Blackboard: http://drake.blackboard.com
Old exams: http://wmboal.com/regress
Email: william.boal@drake.edu

BOAL'S STAT 170

SLIDESHOW HANDOUTS

FALL 2024

PART 1

Introduction and Review

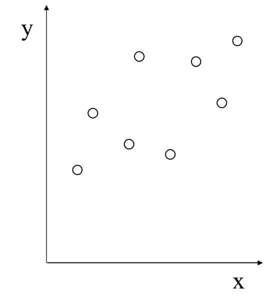
WHAT IS REGRESSION ANALYSIS?

WHAT IS REGRESSION ANALYSIS?

- What is regression analysis?
- What are the two main purposes of regression analysis?

What is regression analysis?

- A type of data analysis.
- Using one or more variables (x) to explain another variable (y).
- Fitting equations or models to noisy data—data that do not lie exactly on the curve.



Ultimate purposes of regression

1. Develop a mathematical model that uses available data (x) to <u>predict</u> y as closely as possible. Close correlation is good enough!

2. Measure the <u>causal effect</u> of x on y.

Examples of prediction

- Using individuals' health status (x) to predict their health insurance claims (y).
- Using individuals' age, sex, driving record, etc. (x) to predict whether they will have an accident next year (y).
- Using characteristics of banks (x) to predict whether they will fail (y).

Successful prediction

- Model predicted values should be close to actual values.
- At a minimum, model should "explain" well the y data used to develop the model.
- In addition, hopefully, the model should predict well outside the sample, too.

Examples of causal inference

- Measuring the effect of a job training program (x) on a typical worker's earnings (y).
- Measuring the effect of hiring more police officers (x) on the crime rate (y).
- Measuring the effect of user fees at national parks (x) on the number of visitors (y).

WHAT IS REGRESSION ANALYSIS?

Successful causal inference

- Mere correlation ≠ causality.
- For causal inference, must measure the effect of x on y, *ceteris paribus.*
- That requires measuring what happens to y when x changes, while holding constant other factors that might influence y.

The challenge of holding other factors constant

- Sometimes data are available on experiments where other factors are held constant through randomization.
  - Example: RAND Health Insurance Experiment (1974-1977).
- But usually we only have nonexperimental (or observational) data.

Perils of causal inference with nonexperimental data: example 1

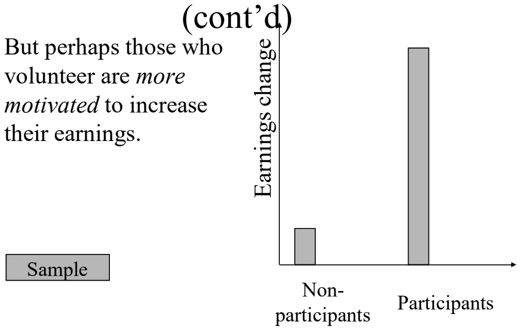
- Suppose we want to measure the effect of a job training program on earnings of participants.
- So we collect data on the change in earnings of people who volunteered for the program and the change in earnings of people who did not.

Perils of causal inference with nonexperimental data: example 1 (cont'd)

- Our goal is to measure the *ceteris paribus* effect—that is, holding other factors constant.
- We want to know how much higher a person's earnings are *as a result of the program,* holding constant differences in ability, attitude, motivation, etc. that also affect earnings.

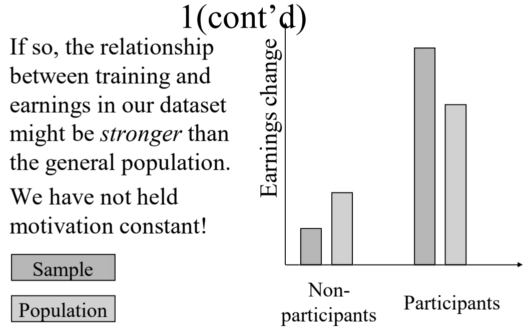
Perils of causal inference with nonexperimental data: example 1 (cont'd)

- But perhaps those who volunteer are *more motivated* to increase their earnings.



Perils of causal inference with nonexperimental data: example 1(cont'd)

- If so, the relationship between training and earnings in our dataset might be *stronger* than the general population.
- We have not held motivation constant!



WHAT IS REGRESSION ANALYSIS?

Perils of causal inference with nonexperimental data: example 2

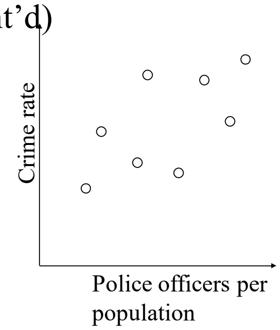
- Suppose we want to measure the effect of hiring more police officers on crime.
- So we collect data from different cities on crime rates and on the number of police officers per 1000 population.

Perils of causal inference with nonexperimental data: example 2 (cont'd)

- Our goal is to measure the *ceteris paribus* effect—that is, holding everything else constant.
- We want to know how much lower a city's crime rate would be *as a result of hiring more police,* holding constant differences poverty rates, average age of the population, etc. that also affect the crime rate.

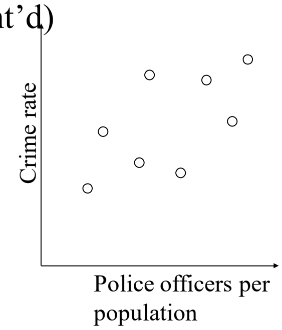
Perils of causal inference with nonexperimental data: example 2 (cont'd)

- But perhaps cities with high crime rates (for whatever reason) *respond* by increasing the number of police officers.
- Then the data might be scattered as shown.



Perils of causal inference with nonexperimental data: example 2 (cont'd)

- Then the relationship between police officers and crime in our data might be the *opposite* of the true *ceteris paribus* relationship!
- We have not held other factors constant.



Conclusions

- The main purposes of regression analysis are
  1. __________________________.
  2. __________________________.
- Successful predictions are close to the actual values of y.
- Successful causal inference holds constant other factors that affect y.

DATA SETS

DATA SETS

•What four forms do data sets usually take?

Structure of economic datasets

- *Datum* = a single number, like 47.5. Plural of datum is_________.
- *Dataset* = array of data to be analyzed.
- Datasets are often arranged so that the rows are _________________ and the columns are _________________.
- Datasets differ in how observations are related to each other.

Types of datasets

1. Cross-sections.
2. Time series.
3. Pooled cross-sections.
4. Panels.

1. Cross-sectional datasets

- All observations collected at roughly the same point in time.
- Observations can be people, firms, industries, cities, countries, etc.

Obs. #	Name	Age	Education	Income
1	B. Smith	34	12 years	$38,845
2	C. Valdez	47	16 years	$65,150
3	J. Huang	24	18 years	$45,275

Commonly used cross-section datasets

- Household surveys such as U.S. Decennial Census, American Community Survey, Current Population Survey, Consumer Expenditure Survey, Survey of Consumer Finances.*
- Point-in-time data sets on firms, products, U.S. states, etc.

* First three are available free at ipums.org .

Cross-sectional datasets are easiest to analyze

- Often we can plausibly assume that observations are a _____________ *sample* from some larger population.
- Each new observation is a fresh draw from the population, unrelated to other observations.
- Observations are thus _________________.

DATA SETS

2. Time-series datasets

- Same individual (person, firm, country) is observed repeatedly over time.
- Frequency might be weekly, monthly, quarterly, or annual.

Obs. #	Year	Unempl. rate	Inflation (CPI)	GR RGDP per capita
1	2000	4.0	3.4	2.5
2	2001	4.7	2.8	-0.6
3	2002	5.8	1.6	1.1

Commonly-used time-series datasets

- Macroeconomic time series (GDP, unemployment, inflation, interest rates, etc.)*
- Currency exchange rates.*
- Stock, bond, and commodity prices.

* Available free at fred.stlouisfed.org.

Patterns in time-series

- Time-series data often show ____________ patterns (unless the data are annual).
  - Electricity use peaks in July or August every year most places.
  - Unemployment peaks in June most years.
- Time series data often show long-run ____________ (usually upward).
  - GDP, employment, and the price level all trend upward.

Time series datasets are harder to analyze

- Time-series observations are *not* usually independent over time.
- Example: If GDP is above trend in one quarter, there is a good chance GDP will be __________ trend in the next quarter, too.
- Each new observation is *not* a fresh draw from the population. Time-series data sets cannot be considered __________ samples.

3. Pooled cross-section datasets

- Several cross-section datasets are combined (or pooled).
- Example: Surveys from several different years, covering different individuals, might be combined into one dataset.
- Observations in the same year might be related to each other, but not to observations in another year.

What a pooled dataset looks like

Obs. #	Year	Name	Age	Income
1	2000	B. Smith	34	$38,845
2	2000	C. Valdez	47	$65,150
3	2000	J. Huang	24	$45,275
4	2002	P. Abdul	65	$55,250
5	2002	A. O'Toole	19	$22,750
6	2002	H. Schmidt	29	$44,500

DATA SETS

Why pooled datasets can be useful

- Can include _________ observations than a single cross-section. The more observations, the more precise are statistical estimates.
- Can estimate relationships _________ each cross section, and compare to see whether relationship has changed over time.

4. Panel (or longitudinal) datasets

- Same cross-section is followed over time.
- Same individuals appear, period after period.
- Example: A few government surveys collect information from the same people month after month.

What a panel dataset looks like

Obs. #	Year	Name	Age	Income
1	2000	B. Smith	34	$38,845
2	2001	B. Smith	35	$40,150
3	2002	B. Smith	36	$42,750
4	2000	C. Valdez	47	$65,150
5	2001	C. Valdez	48	$68,275
6	2002	C. Valdez	49	$70,155

Commonly-used panel data sets

- U.S. Census data (states, cities, counties).
- Financial data sets like Compustat (firms).
- World Economic Outlook, maintained by International Monetary Fund (countries).
- National Longitudinal Surveys and Panel Study of Income Dynamics (households).

Why panel datasets can be useful

- Sometimes can get better *ceteris paribus* measures.
- Extraneous differences between individuals (if constant over time) can be removed by focusing on *changes* over time in the same individual.

Conclusions

- A ____________________ dataset observes many individuals (persons, firms, states, countries, etc.) at one point in time.
- A ____________________ dataset observes one individual repeatedly at many points in time.
- A ____________________ dataset combines several cross-sections.
- A ____________________dataset observes the same set of individuals at different points in time.

THE SUMMATION OPERATOR

THE SUMMATION OPERATOR

- What does the symbol Σ mean?
- How can it be manipulated?

Meaning of summation symbol (Σ)

- *All expressions to the right of Σ should be added, over the specified range of the index.*
- Formally,

$$\sum_{i=m}^{n} x_i \equiv x_m + x_{m+1} + x_{m+2} + \ldots + x_{n-1} + x_n$$

- Example: If $x_1=3$, $x_2=5$, $x_3=6$, and $x_4=8$, then $\sum_{i=1}^{4} x_i =$ ____, and $\sum_{i=2}^{3} x_i =$ ____.

Manipulating the summation symbol (Σ)

- The summation symbol is just shorthand for addition.
- All the properties of addition apply to Σ, including the
  - Commutative law (rearranging order of sum)
  - Distributive law (taking out common factors)

Taking out a common factor

- A common factor (identical for every term in the summation) can be taken outside the summation symbol.

$$\sum_{i=1}^{n} a x_i =$$

- Reason: "distributive law."
- Special case--sum of constants: $\sum_{i=1}^{n} a =$

Rearranging order of sums

- Addition gives the same answer, no matter what order terms are summed in.

$$\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i =$$

- "Commutative law."

Manipulating double sums

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j = \sum_{i=1}^{n} x_i \left(y_1 + \ldots + y_m \right)$$

$$= x_1 y_1 + x_1 y_2 + \ldots + x_1 y_m$$
$$+ x_2 y_1 + x_2 y_2 + \ldots + x_2 y_m$$
$$+ \ldots$$
$$+ x_n y_1 + x_n y_2 + \ldots + x_n y_m$$
$$=$$

THE SUMMATION OPERATOR

NOT true, in general

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$$

- Example: Suppose $x_1=3$, $x_2=2$, $x_3=4$.

$$\sum_{i=1}^{n} x_i^2 = 9 + 4 + 16 =$$

$$but \left(\sum_{i=1}^{n} x_i\right)^2 = (3 + 2 + 4)^2 =$$

NOT true, in general

$$\sum_{i=1}^{n} \frac{x_i}{y_i} \neq \left(\sum_{i}^{n} x_i\right) \Big/ \left(\sum_{i}^{n} y_i\right)$$

- Example: Suppose $x_1=3$, $x_2=2$, $x_3=4$, and $y_1=1$, $y_2=2$, $y_2=2$.

$$\sum_{i=1}^{n} \frac{x_i}{y_i} = \frac{3}{1} + \frac{2}{2} + \frac{4}{2} =$$

$$but \left(\sum_{i}^{n} x_i\right) \Big/ \left(\sum_{i}^{n} y_i\right) = \frac{3+2+4}{1+2+2} =$$

Conclusions

- The summation symbol (Σ) is convenient shorthand for ________________.
- It has all the usual properties of addition, including the _________________ law (rearranging the order) and the __________ law (taking out common factors).

DERIVATIVES OF SUMS

DERIVATIVES OF SUMS

•How can we find the derivative of a function with a Σ symbol?

Rule for derivatives of sums

- From calculus we know that the derivative of a sum of functions is the _______ of the derivatives:

$$\frac{d}{d\alpha}\left(f_1(\alpha)+f_2(\alpha)\right)=\frac{df_1}{d\alpha}+\frac{df_2}{d\alpha}$$

- Example:

$$\frac{d}{d\alpha}\left(2\alpha^2+3\ln(\alpha)\right)=$$

Derivatives of functions with Σ

- Same rule applies to summation symbol.

$$\frac{d}{d\alpha}\sum_{i=1}^{n} f_i(\alpha)=\sum_{i=1}^{n}\frac{df_i}{d\alpha}$$

- We simply take the derivative term-by-term.

Example 1

- Example 1:

$$\frac{d}{d\alpha}\sum_{i=1}^{n}\left(\alpha x_i\right)=$$

How do we know whether to differentiate with respect to α or x?

- If the derivative is to be taken with respect to x, the operator is written as ... $\dfrac{d}{dx}$
- If the derivative is to be taken with respect to α, the operator is written as ... $\dfrac{d}{d\alpha}$

Example 2

- Example 2:

$$\frac{d}{d\alpha}\sum_{i=1}^{n}\left(\alpha x_i+\alpha^2 x_i^2\right)=$$

DERIVATIVES OF SUMS

Example 3

- Example 3:
$$\frac{d}{d\alpha}\sum_{i=1}^{n}\left(x_i-\alpha\right)^2 =$$

More examples

- Example 4:
$$\frac{d}{d\beta}\sum_{i=1}^{n}\left(x_i\beta+y_i\,\beta^2\right)=\sum_{i=1}^{n}$$
- Example 5:
$$\frac{d}{d\gamma}\sum_{i=1}^{n}\left(\gamma+x_i\right)^{-1}=-\sum_{i=1}^{n}$$

Conclusions

- No special formulas are required for taking the derivatives of functions containing the summation symbol (Σ).
- The derivative of a sum is just the ______________________ of all the terms.

AVERAGES AND WEIGHTS

AVERAGES AND WEIGHTS

- How can averages be defined using the symbol Σ ?
- What properties do averages have?

Simple average (or "sample mean")

- Definition: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Example: Suppose $x_1=5$, $x_2=6$, and $x_3=7$.

$$\bar{x} = \tfrac{1}{3}\sum_{i=1}^{3} x_i = \left(\tfrac{1}{3}\right)(5+6+7) =$$

Deviations from sample mean

- Definition: $x_i - \bar{x}$
- Key property of deviations from sample mean: $$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) =$$
- By contrast, $$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \geq$$

Deviations from sample mean

- Definition: $x_i - \bar{x}$
- Key property of deviations from sample mean: $$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$
- By contrast, $$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \geq 0$$

An algebraic identity

(1) $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 =$

Another algebraic identity

(2) $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) =$

AVERAGES AND WEIGHTS

What is a weighted sum?

- Each term (say x_i) is multiplied by some weighting factor (say w_i) before adding:

$$\sum_{i=1}^{n} w_i x_i$$

- Example: Suppose $x_1=5$, $x_2=6$, and $x_3=7$, and weights are $w_i=1/i$. Then

$$\sum_{i=1}^{3} w_i x_i = \left(\tfrac{1}{1}\right)5 + \left(\tfrac{1}{2}\right)6 + \left(\tfrac{1}{3}\right)7 =$$

What is a weighted average?

- A weighted sum whose (nonnegative) weights alone sum to one: $\sum_{i=1}^{3} w_i = 1$
- Example: Again suppose $x_1=5$, $x_2=6$, and $x_3=7$. Now suppose weights are $w_1=1/4$, $w_2=1/4$, and $w_3=1/2$. Then

$$\sum_{i=1}^{3} w_i x_i = \left(\tfrac{1}{4}\right)5 + \left(\tfrac{1}{4}\right)6 + \left(\tfrac{1}{2}\right)7 =$$

Conclusions

- A simple average (or sample mean) can be defined as (1/n) times the sum.
- The sum of deviations from the sample mean is necessarily ______________.
- _________________ multiply each term by some weight, before summing.
- ____________________ have nonnegative weights that sum to one.

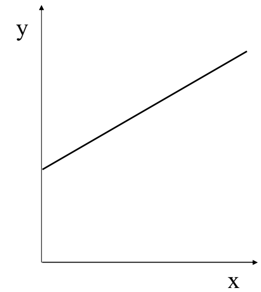
DEFINITION OF LEAST-SQUARES

DEFINITION OF LEAST-SQUARES

- What is the "least-squares principle" for fitting a line to data?
- What are the formulas for the least-squares estimators of the intercept and slope?

Linear relationships

- Suppose x and y have a linear relationship: $y = \beta_1 + \beta_2 x$.
- β_1 = ____________
- β_2 = ____________

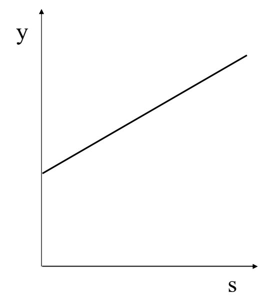


Meaning of slope

- If x increases by a small amount, then y changes by β_2 times that amount.
- Example: Suppose $\beta_2 = 2$ and x increases by 0.4. Then y increases by (approximately) _______.

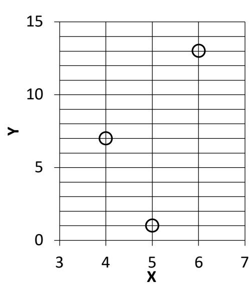
Measuring relationships

- Suppose we have data on x_i and y_i, for $i = 1$ through n.
- We believe that x and y have a roughly linear relationship: $y = \beta_1 + \beta_2 x$.
- How can we estimate β_1 and β_2?



Simple example with 3 observations: data and scatterplot

Obs.	X	Y
1	4	7
2	5	1
3	6	13



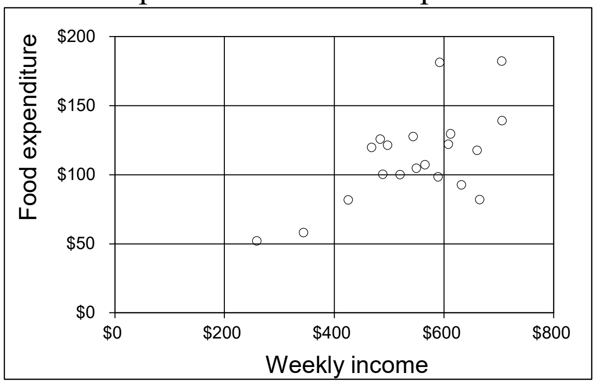
Another example: household data on income and food expenditure

Household no.	Weekly income	Food expenditure
1	258.3	52.25
2	343.1	58.32
3	425	81.79
4	467.5	119.9
5	482.9	125.8
6	487.7	100.46
7	496.5	121.51
8	519.4	100.08
9	543.3	127.75
10	548.7	104.94

Household no.	Weekly income	Food expenditure
11	564.6	107.48
12	588.3	98.48
13	591.3	181.21
14	607.3	122.23
15	611.2	129.57
16	631	92.84
17	659.6	117.92
18	664	82.13
19	704.2	182.28
20	704.8	139.13

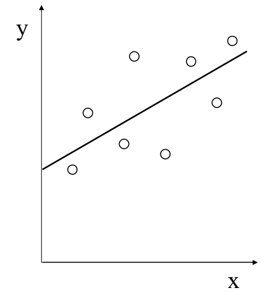
DEFINITION OF LEAST-SQUARES

Scatterplot of household expenditure



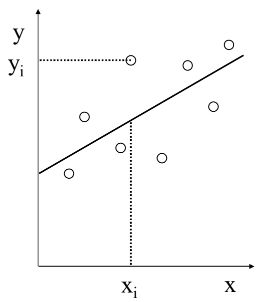
Fitting a line to data

- Economic data rarely lie exactly on a straight line.
- So we must find a line that fits the data "best."
- How to choose the "best"-fitting line?



Deviations from the line

- The "best" line would come close to the actual data points.
- Suppose we measure deviations from the line vertically.



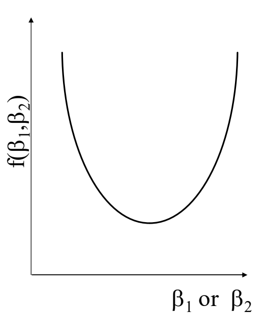
The least-squares principle

- Choose the line that minimizes the sum of the squared vertical deviations.
- In other words, find values of β_1 and β_2 that minimize the following objective function:

$$f(\beta_1,\beta_2)=\sum_{i=1}^{n}\left(y_i-[\beta_1+\beta_2 x_i]\right)^2$$

Minimizing the function

- The least-squares objective function is quadratic in β_1 and β_2 .
- Its minimum occurs where its slope = ________________
- Use this fact to solve for β_1 and β_2 .



First, find formulas for the slopes (or derivatives) of the objective function

Derivative of $f(\beta_1,\beta_2)$ with respect to β_1:

$$\frac{\partial f(\beta_1,\beta_2)}{\partial \beta_1}=\sum_{i=1}^{n}-2\left(y_i-[\beta_1+\beta_2 x_i]\right)$$

Derivative of $f(\beta_1,\beta_2)$ with respect to β_2:

$$\frac{\partial f(\beta_1,\beta_2)}{\partial \beta_2}=\sum_{i=1}^{n}-2\left(y_i-[\beta_1+\beta_2 X_i]\right)x_i$$

DEFINITION OF LEAST-SQUARES

Second, set the derivatives equal to zero

- These equations are called the "first-order necessary conditions" (FONCs), or sometimes "normal equations."

$$0 = \sum_{i=1}^{n} -2\left(y_i - [\beta_1 + \beta_2 x_i]\right)$$

$$0 = \sum_{i=1}^{n} -2\left(y_i - [\beta_1 + \beta_2 X_i]\right) x_i$$

The least-squares estimators

- These equations can be solved to give the least-squares estimators.
- Slope: $\widehat{\beta_2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- Intercept: $\widehat{\beta_1} = \bar{y} - \widehat{\beta_2}\,\bar{x}$

where $\bar{x}$ = sample mean of x_i and $\bar{y}$ = sample mean of y_i.

Least-squares estimates for the simple example

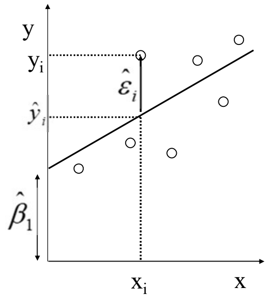
- Here, $\bar{x}$ =_____ and $\bar{y}$ =_____ .
- Slope: $\widehat{\beta_2} = \frac{\sum_{i=1}^{3}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{3}(x_i-\bar{x})^2}$ =_____.
- Intercept: $\widehat{\beta_1} = \bar{y} - \widehat{\beta_2}\,\bar{x}$ =____ .

Least-squares estimates for the household expenditure example

- Here, $\bar{x}$ =_________ and $\bar{y}$ =_________ .
- Slope: $\widehat{\beta_2} = \frac{\sum_{i=1}^{20}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{20}(x_i-\bar{x})^2}$ =_________.
- Intercept: $\widehat{\beta_1} = \bar{y} - \widehat{\beta_2}\,\bar{x}$ =_________.

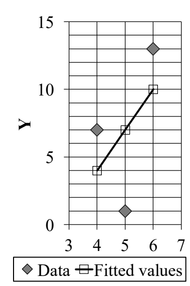
Definition of least-squares fitted values and residuals

- LS "fitted value" or "predicted value" = $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$
- LS "residual" = $\hat{\varepsilon}_i = y_i - \hat{y}_i$



Fitted values for the simple example

Obs.	x_i	y_i	$\hat{y}_i$	$\hat{\varepsilon}_i$
1	6	13		
2	4	7		
3	5	1		



$\hat{\beta}_2 = 3$ $\hat{\beta}_1 = -8$

DEFINITION OF LEAST-SQUARES

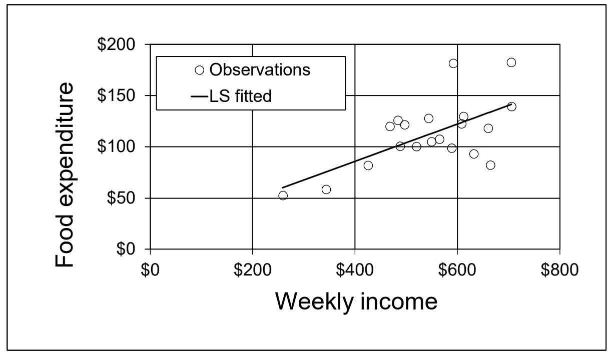
Fitted values for the household expenditure example

- Y-intercept (β_1): 12.94
- Slope (β_2) : 0.18

- Fitted line:

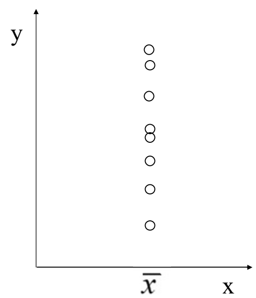
y = __________ + ________ x

Fitted values for the household expenditure example: plot



Remark: LS estimators cannot always be calculated

- Note that $\hat{\beta}_2$ cannot be calculated if x never varies: $x_i = \overline{x}$.
- Reason: $\hat{\beta}_2 = 0/0$
- But this is not a defect of LS. If x never varies, how can we expect to measure its impact on y?



Remark: LS fitted line *must* pass through sample means

- The LS fitted line is given by:

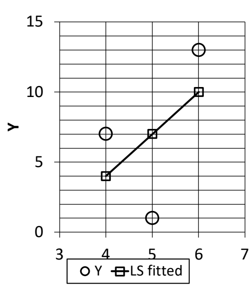
$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$$

- The sample means necessarily lie on the LS fitted line: $$\overline{y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{x}$$

(Follows from formula for $\hat{\beta}_1$.)

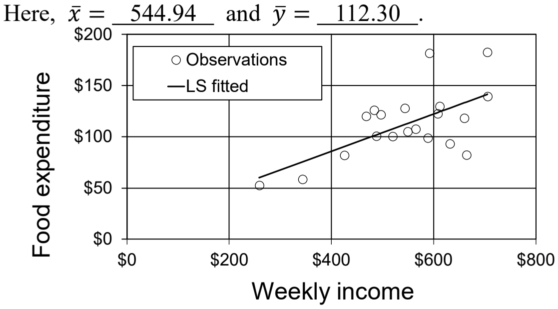
LS fitted line *must* pass through sample means: simple example

- Here, $\bar{x} =$ __5__ and $\bar{y} =$ __7__ .



LS fitted line *must* pass through sample means: household expenditure example

Here, $\bar{x} =$ __544.94__ and $\bar{y} =$ __112.30__ .



DEFINITION OF LEAST-SQUARES

Conclusions

- One way to fit a line to data is to choose the line that minimizes the sum of the ____________________________________.
- This is called the "____________________ principle."
- Using calculus, explicit formulas can be derived for the least-squares estimators of the slope and intercept.

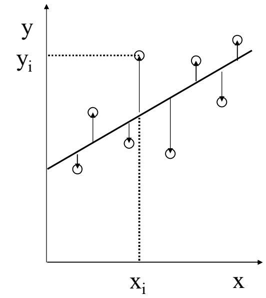
ALTERNATIVES TO LEAST-SQUARES

ALTERNATIVES TO LEAST-SQUARES

•What other principles can be used to estimate the intercept and slope?

Least-squares is only one approach

- Least-squares minimizes the sum of the *squared* deviations, measured *vertically*: $(y_i - [\beta_1 + \beta_2 x_i])^2$.
- How else could we define the "best" fit?

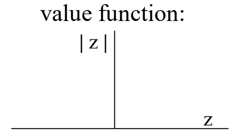


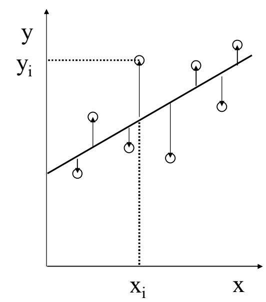
Alternative objective functions

- Many other criteria for "best fit" are possible. We consider two:
  (1) Sum of absolute value of deviations.
  (2) Sum of squared deviations measured horizontally.

(1) Absolute deviations

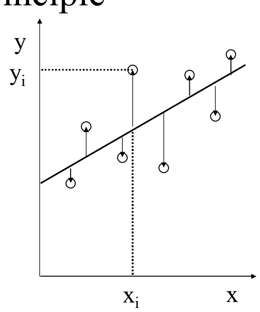
- Absolute value of deviations are given by: $| y_i - [\beta_1 + \beta_2 x_i] |$.
- Recall the absolute value function:





The least-absolute-deviations (LAD) principle

- Minimize the sum of the absolute value of deviations, measured vertically: $| y_i - [\beta_1 + \beta_2 x_i] |$.

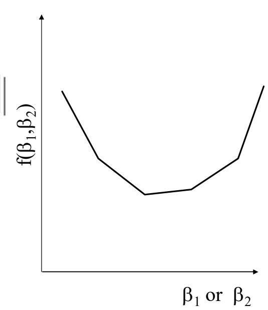


Minimizing the function

- Objective function is:

$$f(\beta_1,\beta_2)=\sum_{i=1}^{n}\left|y_i-\left[\beta_1+\beta_2 x_i\right]\right|$$

- The absolute value function has 1 kink.
- Sum of n absolute values has ___ kinks.



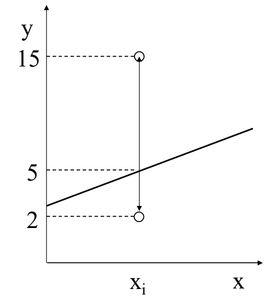
ALTERNATIVES TO LEAST-SQUARES

Finding LAD estimates of β_1 and β_2

- Cannot use calculus because cannot take derivative of kinky functions.
- No formulas exist for β_1 and β_2 .
- Must use trial-and-error (preferably aided by computer) to find LAD estimates.

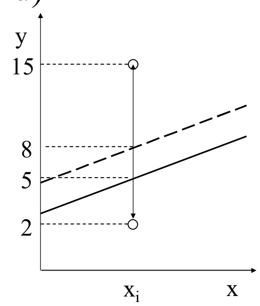
Why LAD is different from LS

- Example: LAD gives less weight to outliers (large deviations).
- Here, the sum of squared resids (SSR) = __________.
- The sum of absolute deviations (SAD) = __________.



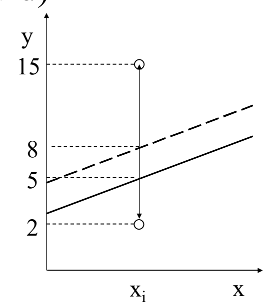
Why LAD is different from LS (cont'd)

- Suppose we move the fitted line up three units, closer to the outlier at the top.
- The SSR falls to ___.
- The SAD __________ ___________________.



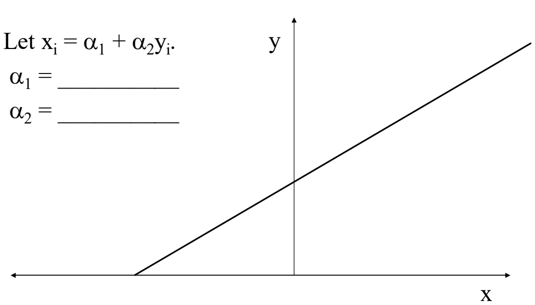
Why LAD is different from LS (cont'd)

- Conclude: Compared to LS, LAD is less responsive to observations far from the regression line.



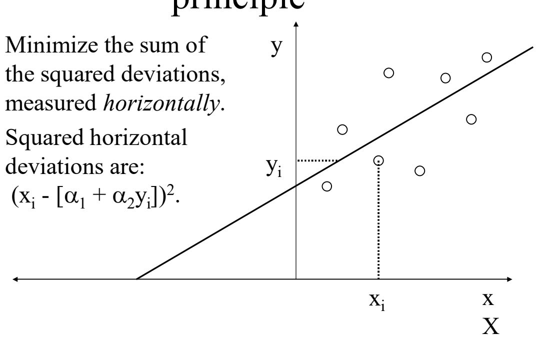
(2) Horizontal deviations

- Let $x_i = \alpha_1 + \alpha_2 y_i$.
- α_1 = ___________
- α_2 = ___________



The reverse-least-squares (RLS) principle

- Minimize the sum of the squared deviations, measured *horizontally*.
- Squared horizontal deviations are: $(x_i - [\alpha_1 + \alpha_2 y_i])^2$.



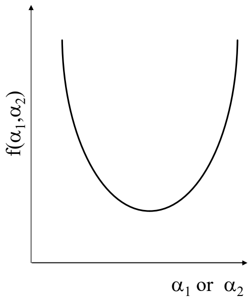
ALTERNATIVES TO LEAST-SQUARES

Minimizing the function

- RLS objective function is quadratic in α_1 and α_2:

$$f(\alpha_1,\alpha_2)=\sum_{i=1}^{n}\left(x_i-[\alpha_1+\alpha_2 y_i]\right)^2$$

- Its minimum occurs where its slope = ________________



Solving for RLS estimates of α_1 and α_2

- Set zero equal to derivative of $f(\alpha_1,\alpha_2)$ with respect to α_1:

$$0=\sum_{i=1}^{n}-2\left(x_i-[\alpha_1+\alpha_2 y_i]\right)$$

- Set zero equal to derivative of $f(\alpha_1,\alpha_2)$ with respect to α_2:

$$0=\sum_{i=1}^{n}-2\left(x_i-[\alpha_1+\alpha_2 y_i]\right)y_i$$

The RLS estimators

- These equations can be solved to give the reverse least-squares estimators:

$$\hat{\alpha}_1=\bar{x}-\hat{\alpha}_2\bar{y}$$

$$\hat{\alpha}_2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Remarks

- Note that $\hat{\alpha}_2$ cannot be calculated if y never varies.
- The sample means necessarily lie on the RLS fitted line:

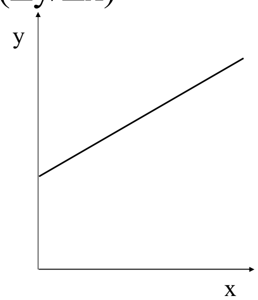
$$\bar{x}=\hat{\alpha}_1+\hat{\alpha}_2\bar{y}$$

RLS estimators for y-intercept and slope ($\Delta y/\Delta x$)

- Solve $x_i=\alpha_1+\alpha_2 y_i$ for y_i to get:

$$y_i=\left(\frac{-\alpha_1}{\alpha_2}\right)+\left(\frac{1}{\alpha_2}\right)x_i$$

- RLS Y-intercept estimator = $\left(\frac{-\hat{\alpha}_1}{\hat{\alpha}_2}\right)$
- RLS slope estimator = $\left(\frac{1}{\hat{\alpha}_2}\right)$

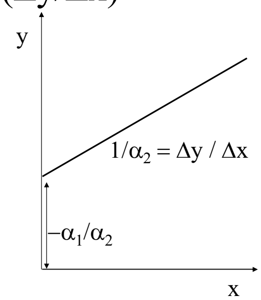


RLS estimators for y-intercept and slope ($\Delta y/\Delta x$)

- Solve $x_i=\alpha_1+\alpha_2 y_i$ for y_i to get:

$$y_i=\left(\frac{-\alpha_1}{\alpha_2}\right)+\left(\frac{1}{\alpha_2}\right)x_i$$

- RLS Y-intercept estimator = $\left(\frac{-\hat{\alpha}_1}{\hat{\alpha}_2}\right)$
- RLS slope estimator = $\left(\frac{1}{\hat{\alpha}_2}\right)$

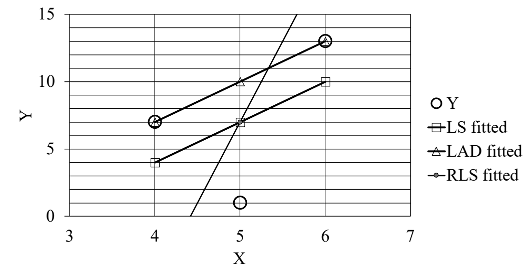


ALTERNATIVES TO LEAST-SQUARES

Alternative estimates for the simple example

	Y-intercept (β_1)	Slope (β_2)
Ordinary LS	-8	3
LAD	-5	3
Reverse LS	-53	12

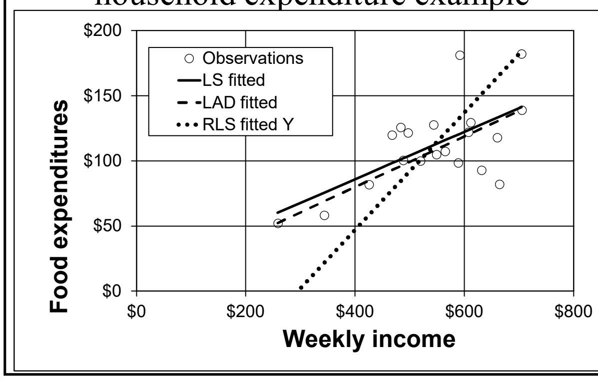
Alternative fitted values for the simple example



Alternative estimates for the household expenditure example

	Y-intercept (β_1)	Slope (β_2)
Ordinary LS	12.94	0.18
LAD	1.99	0.19
Reverse LS	-132.88	0.45

Alternative fitted values for the household expenditure example



Other possible objective functions

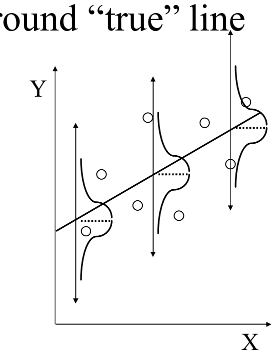
Which principle to use?

- Different objective functions for "best fit" lead to different estimates—sometimes *very* different estimates (see RLS).
- Which objective function is best?
- Answer depends on <u>why we think the data do not lie exactly on the line</u>.

ALTERNATIVES TO LEAST-SQUARES

Data scattered around "true" line

- Perhaps the data are randomly scattered around a "true" line.
- To model this random scattering, and decide which principle is best, we must first review theory of …
  ____
  ____



Conclusions

- Reasonable alternative principles can be used to define the "best fit" and find estimators for the intercept and slope.
- But they sometimes give very ___________ answers.
- To decide which principle is best, we must model the scattering of data off the "true" line using ___________________________.

RANDOM VARIABLES

RANDOM VARIABLES

- What is probability?
- What is a random variable?
- What is the difference between discrete and continuous random variables?

Probability: definition

- *Relative frequency with which an event occurs in repeated trials.*
- Examples:
  - Flip of fair coin: probability of "heads" = _____
  - Toss of fair die: probability of "1" = _____

Properties of probability

- Probabilities must lie between _______ and _______.
- Probabilities of all possible but mutually exclusive outcomes must sum to ______.

Random variable: definition

- *A variable whose value is determined by a random process.*
- Each value that a random variable can take is associated with some probability.
- The sum of all those probabilities = ______.
- Random variables can be discrete or continuous.

Discrete random variable: definition

- *A random variable that can take only values that are separated from each other, such as integers.*
- These values can be listed.
- Example: a random variable that can take only the values 0, 1, and 2.
- The total probability of all possible values = _______.

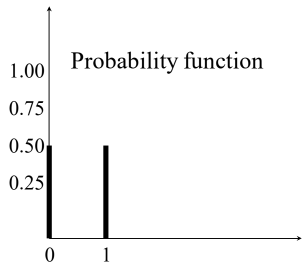
Discrete random variable: examples

- Binary random variables: Whether a person is employed, owns home, belongs to a labor union, owns a smartphone.
- Other discrete random variables: Number of children in household, of cars owned, of visits paid to doctor in a year.

RANDOM VARIABLES

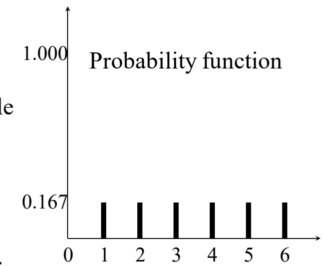
Discrete random variable: example 1

- Coin toss: Let X=1 if "heads" and X=0 if "tails."
- If a fair coin, then Prob{X=1} = _____, and Prob{X=0} = ___.



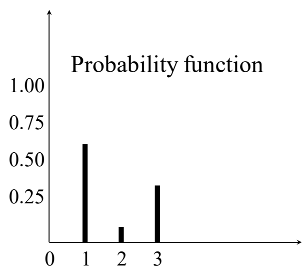
Discrete random variable: example 2

- Roll of die: Let X = number of dots showing.
- X can take six possible values.
- If a fair die, then Prob{X=1} = = Prob{X=2} = … = Prob{X=6} = ____.



Discrete random variable: example 3

- Suppose a person is selected at random from the population.
- Let X = 1 if employed, 2 if unemployed, 3 if out of the labor force (not looking for work).

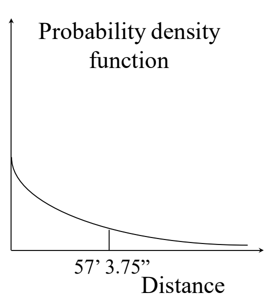


Continuous random variable: definition

- *A random variable that takes a continuous range of values on real number line.*
- The probability of any particular value is essentially _______, but the probability of a range can be positive.
- The total probability over the whole real line must = ______.

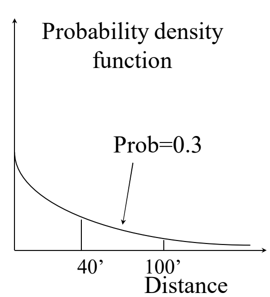
Continuous random variable: example 1

- Let X = distance that a baseball hit from home plate travels before touching the ground.
- Probability of any particular distance (say, 57 ft. 3 1/4 in.) is essentially zero.



Continuous random variable: example 1 (cont'd)

- But probability of a range (say 40-100 ft) is positive.
- Probability of range = area under *probability density function*.



RANDOM VARIABLES

Continuous random variable: applications

- Many variables take so many values that they are most conveniently modeled as continuous.
- Stock prices.
- Health care spending.
- Insurance claims.

Continuous random variable: economic applications

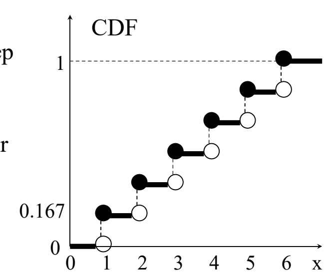
- Microeconomic examples: quantities and prices of electricity, food, steel, automobiles.
- Macroeconomic examples: GDP, money supply, price level, employment, currency exchange rates.

Cumulative distribution function (CDF): definition

- *Function showing the probability that a random variable takes value less than or equal to the argument.*
- F(x) = Prob{X≤x}.
- F(-infinity) = ________.
- F(infinity) = _________.
- F(x) cannot slope __________.

CDF for discrete random variable

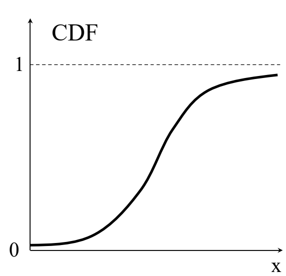
- For discrete random variable, CDF is a step function.
- Example: roll of die, six possible values for X.
- F(x) = Prob{X≤x} jumps up at each of these six values.



CDF for continuous random variable: example

- For continuous random variable, CDF is continuous, upward-sloping.
- F(x) = Prob{X≤x} =

$$\int_{-\infty}^{x} f(z)dz$$



Conclusions

- *Probability* is the relative ___________ with which an event occurs in repeated trials.
- A _____________ random variable takes a countable number of possible values.
- A _____________ random variable takes a continuous range of possible values.
- The _______________ distribution function F(x) shows Prob{X≤x}.

JOINT DISTRIBUTIONS

JOINT DISTRIBUTIONS

- How can we describe random variables that are related to each other?
- What are joint distributions, marginal distributions, and conditional distributions?

Joint distribution: definition

- Any two random variables have a joint distribution.
- *A joint distribution shows the probabilities of any particular combination of values the random variables may take.*

Joint distribution of discrete random variables

- *Probability associated with some particular combination of outcomes of two random variables.*
- $f_{x,y}(x,y)$ = Prob{X=x and Y=y}.

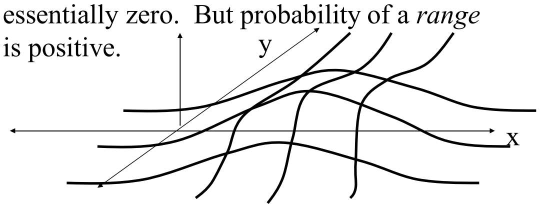
Joint distribution of discrete random variables: example

- For two discrete random variables, joint distribution can be displayed as table.
- Example: $f_{x,y}(1,2)$ = Prob{X=1 and Y=2} = 0.1 .

	Y=1	Y=2	Y=3
X=1	0.2	0.1	0.1
X=2	0.3	0.2	0.1

Joint distribution of continuous random variables

- Can be defined by joint density function $f_{x,y}(x,y)$.
- Probability of any particular combination is essentially zero. But probability of a *range* is positive.



Marginal distribution

- In the context of joint distributions, the distribution of each individual random variable is called its "marginal distribution."

JOINT DISTRIBUTIONS

Marginal probabilities for discrete distributions

- Let $p_{ij} = f_{x,y}(x_i,y_i)$ = Prob{$X=x_i$ and $Y = y_j$}.
- Then the marginal probabilities are given by

$$f_x(x_i) = \sum_j p_{ij}$$

$$f_y(y_j) = \sum_i p_{ij}$$

Marginal probabilities of discrete random variables: example

- For joint discrete distribution, marginal probabilities are column sums and row sums.

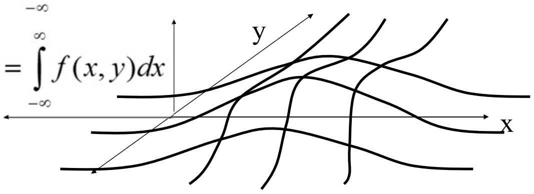
	Y=1	Y=2	Y=3	Marginal X
X=1	0.2	0.1	0.1	
X=2	0.3	0.2	0.1	
Marginal Y				

Marginal density functions for continuous distributions

- Let $f_{x,y}(x,y)$ = joint density function. Then the marginal density functions are given by

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

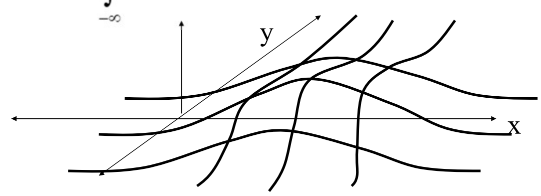
$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$



Graphic interpretation of marginal density functions

- Marginal density is area under a "slice" of the joint density.

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy$$



Independence

- Two random variables are independent if and only if the value taken by one has *no effect* on the distribution of the other.
- Formally, X and Y are independent if and only if $f_{x,y}(x,y) = f_x(x)\, f_y(y)$.

Independence of discrete random variables: example

- Here, X and Y are not independent because 0.1 = Prob{X=1 and Y=2} $\neq 0.4 \times 0.3$.

	Y=1	Y=2	Y=3	Marginal X
X=1	0.2	0.1	0.1	0.4
X=2	0.3	0.2	0.1	0.6
Marginal Y	0.5	0.3	0.2	

JOINT DISTRIBUTIONS

Independence of discrete random variables: more examples

- What is the probability of rolling two dice and getting "boxcars" (two sixes)?
  - If the dice are independent, $f_{x,y}(6,6) = f_x(6)\, f_y(6) = (1/6)(1/6) =$ _____.
- What is the probability of rolling a three and a four?
  - If the dice are independent, $f_{x,y}(3,4) + f_{4,3}(4,3) = (1/6)(1/6) + (1/6)(1/6) =$ _____.

Conditional distribution

- *The conditional distribution of y given x is the distribution of y assuming that x takes some particular value.*
- Write "$f_{y|x}(y|x)$." Read "|" as "given" or "conditional on".
- Calculate as $f_{y|x}(y|x) = f(x,y) / f_x(x)$.
- Probabilities of conditional distributions (like ordinary distributions) must sum to one.

Conditional distribution of discrete random variable: example

- If X=1, restrict our attention to first row, whose total probability is 0.4.
- $f_{y|x}(2|1) =$ ___________________.

	Y=1	Y=2	Y=3	Marginal X
X=1	0.2	0.1	0.1	0.4
X=2	0.3	0.2	0.1	0.6
Marginal Y	0.5	0.3	0.2	

Conditional distributions of independent random variables

- Recall: Two random variables are independent if and only if the value taken by one has *no effect* on the distribution of the other.
- This means that the conditional distribution is always the same, and thus equal to the marginal distribution: $f_{y|x}(y|x) = f_y(y)$.

Conditional distributions of independent random variables: alternative example

- $f_{y|x}(2|1) =$ __________________________.

	Y=1	Y=2	Y=3	Marginal X
X=1	0.05	0.1	0.1	0.25
X=2	0.15	0.3	0.3	0.75
Marginal Y	0.2	0.4	0.4	

Conclusions

- A ___________ distribution shows probabilities of any particular combination of values of random variables.
- A ___________ distribution is just the distribution of one random variable in a joint distribution.
- A ____________ distribution shows probabilities of one variable given that another random variable takes a particular value.
- For ____________ random variables, the marginal distribution equals the conditional distribution.

EXPECTED VALUE OR MEAN

EXPECTED VALUE OR MEAN

- What is the "mean" of a random variable?
- How can we evaluate the mean of a linear function of a random variable?

Central tendency of a random variable

- Often we want to characterize a random variable by the value it tends to take on average--that is, its *central tendency*.
- One measure of central tendency is the *expected value* or *mean.*

Expected value: definition

- The sum of all possible values of a random variable, after first multiplying them by their probabilities.
- Notation: E(X).
- Expected value of random variable sometimes called *mean* or *population mean*.

Expected value for a discrete random variable

- Suppose random variable X can take n possible values: $x_1, x_2, \dots, x_n$.
- Each value x_i has associated probability p_i.
- Then E(X) is given by:

$$\sum_{i=1}^{n} x_i p_i$$

Expected value for a discrete random variable: example 1

- Suppose a fair coin is flipped and a game contestant is awarded \$10 if "heads" shows, and \$50 if "tails" shows.
- Let X = amount awarded.
- Then x_1 = \$10, $p_1 = 1/2$, x_2 = \$50, $p_2 = 1/2$.
- $E(X) = x_1 p_1 + x_2 p_2 =$ _______.

Expected value for a discrete random variable: example 2

- Suppose a fair die is thrown and the game player gets to advance her or his token the number of spaces shown on the face.
- Let X = amount shown on face.
- Then $x_1=1$, $x_2=2$, $x_3=3$, $x_4=4$, $x_5=5$, $x_6=6$.
- $\text{Prob}\{x_1=1\} = \dots = \text{Prob}\{x_6=6\} = 1/6$.
- $E(X) = x_1p_1 + x_2p_2 + \dots + x_6p_6 =$ _______.

EXPECTED VALUE OR MEAN

Expected value for a discrete random variable: example 3

- Suppose X takes three possible values.
  - Prob{X=1} = 0.5
  - Prob{X=3} = 0.25
  - Prob{X=11} = 0.25
- Then E(X) = 1*0.5 + 3*0.25 + 11*0.25 = _______.

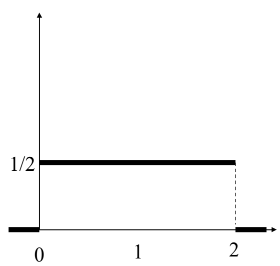
Expected value for a continuous random variable

- Suppose random variable X can take a continuous range of possible values.
- The probability of any subrange is given by the area under the density function f(x).
- Then E(X) is given by: $$\int_{-\infty}^{\infty} x\, f(x)\, dx$$

Expected value for a continuous random variable: example 1

- Suppose a continuous random variable has density function f(x) = 1/2 from x=0 to 2, and zero otherwise.
- *Then E(X) =*

$$\int_0^2 x\left(\tfrac{1}{2}\right)dx = \tfrac{1}{4}x^2\Big|_{x=0}^{x=2} =$$



Sample mean versus population mean

- *Sample mean* = average of outcomes in a sample of n observations chosen from a larger population or distribution.
- Denote sample mean as $\overline{x}$.
- Sample mean need ________ equal population mean because sample is just a subset of population.

Expectations of functions of random variables

- Let g(x) be some function. Its expectation is defined as follows.
- If X is a discrete random variable:

$$E(g(X)) = \sum_{i=1}^{n} g(x_i)\, p_i$$

- If X is a continuous random variable:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

Expectation of linear functions of random variables

- For most functions g(x), the expectation E(g(X)) is a mess. But it is easy to show that for linear functions, E(g(X)) is simple.
- Let X and Y denote random variables, and let a and b denote constants.
- E(aX + b) = a E(X) + b. ("linear operator")
- E(X + Y) = E(X) + E(Y).

EXPECTED VALUE OR MEAN

Expectation of linear functions of random variables: examples

- Suppose E(X) = 5 and E(Y) = 2.
  - If Z = 3X + 7 then E(Z) = ______.
  - If W = X + Y then E(W) = ______.
- Suppose we have n random variables $X_1, \ldots, X_n$ and each of them has the same mean $E(X_i) = 11$.
  - Then $E\left(\sum X_i\right) =$ ______.

Expectation of nonlinear functions of random variables

- There are no such simple rules for nonlinear functions.
- $E(X^2) \neq (E(X))^2$.
- $E(X^3) \neq (E(X))^3$.
- $E(\ln(X)) \neq \ln(E(X))$.

Mean of product of random variables

- In general, E(XY) ≠ (EX)(EY).
- However, in the very special case where X and Y are *independent* random variables, then E(XY) = (EX)(EY).

Remarks

- Synonym for mean = “first moment.”
- Mean need not be finite. Example: Cauchy distribution (= “t” distribution with 1 degree of freedom) has no finite mean.
- However, mean will be finite for all distributions we will use in this course.

Conclusions

- *Expected value (or population mean)* = average value of a random variable.
- Expected value is computed by multiplying each possible value by its ____________ , and then summing the results.
- For a linear function of a random variable, the mean of the linear function is the linear function of the ____________.

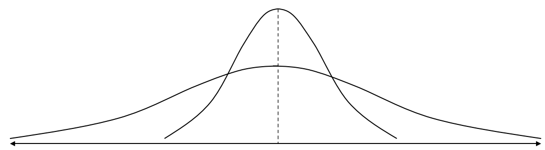
VARIANCE AND STANDARD DEVIATION

VARIANCE AND STANDARD DEVIATION

- What is the “variance” of a random variable?
- What is the “standard deviation”?
- How can we evaluate the variance of a linear function of a random variable?

Dispersion of a random variable

- Often we want to measure a random variable’s *dispersion* or spread around its central tendency.
- One measure of dispersion is the *variance.*



Variance: definition

- Consider the difference between a random variable X and its mean E(X): X-E(X).
- *The sum of all possible squared differences, after first multiplying them by their probabilities.*
- Variance of X = E $[X - E(X)]^2$.
- Notation: Var(X).

Remarks

- Synonym for variance = “second moment about the mean.”
- Variance need not be finite. Example: “t” distribution with 2 degrees of freedom has no finite variance.
- However, variance will be finite for all distributions we will use.

Variance of a discrete random variable

- Suppose random variable X can take n possible values: $x_1, x_2, \ldots, x_n$.
- Each value x_i has associated probability p_i.
- Then Var(X) is given by:

$$\sum_{i=1}^{n} \left(x_i - E(X)\right)^2 p_i$$

Variance of a discrete random variable: example

- Suppose X takes three possible values.
  - Prob{X=1} = 0.5
  - Prob{X=3} = 0.25
  - Prob{X=11} = 0.25
- It is easy to show that E(X) = ______.
- So Var(X) = $(1-4)^2$ * 0.5 + $(3-4)^2$ * 0.25 + $(11-4)^2$ * 0.25 = _____.

VARIANCE AND STANDARD DEVIATION

Variance of a continuous random variable

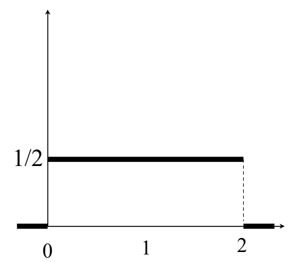
- If X is a continuous random variable, then Var(X) is given by:

$$\int_{-\infty}^{\infty} \left(x - E(x)\right)^2 \; f(x)\, dx$$

Variance of a continuous random variable: example

- Suppose a continuous random variable has density function f(x) = 1/2 from x=0 to 2, and zero otherwise.
- Then Var(X) =

$$\int_0^2 (x-1)^2 \left(\tfrac{1}{2}\right) dx =$$



Key properties of variance

- The following properties are not hard to show.
- Suppose X is a random variable.
  - Then $\mathrm{Var}(X) = E(X^2) - (EX)^2$.
- Suppose a and b are constants and X is a random variable.
  - Then $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.
  - Also Var(a) = Var(b) = 0.

Examples of properties of variance

- Suppose Var(X) = 5.
  - Then Var(3X + 13) = ______.
- Suppose Var(X) = 7.
  - Then Var(2X – 5) = ______.
- Also, Var(7) = _____.

Definition of standard deviation

- Standard deviation is the square root of the variance: $\mathrm{SD}(X) = [\mathrm{Var}(X)]^{1/2}$.
- Key properties of standard deviation follow from properties of variance.
- Suppose a and b are constants and X is a random variable. Then, obviously,
  - SD(a) = _________.
  - SD(aX + b) = _________.

Conclusions

- *Variance* = expected value of the squared deviation of a random variable from its mean.
- *Standard deviation* = _________________ of variance.
- For a linear function of a random variable, the variance of the function equals the coefficient ______________ times the variance of the random variable itself.

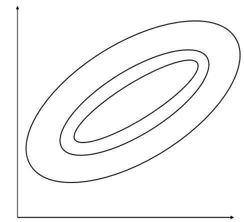
COVARIANCE, CORRELATION, AND CONDITIONAL EXPECTATION

COVARIANCE, CORRELATION, AND CONDITIONAL EXPECTATION

- What are "covariance" and "correlation"?
- What is the mean of a random variable, "conditional" on another random variable?

Measures of association between random variables

- Often we want to measure how closely two random variables move together.
- Two such *measures of association* are
  - covariance
  - correlation



Covariance: definition

- *Expected value of the product of the deviations of two random variables from their respective means.*
- Cov(X,Y) = E [(X-EX)(Y-EY)].
- Measures how the variables move together.

Covariance for discrete random variables

- Suppose X and Y are discrete random variables, taking n and m different values respectively.
- Cov(X,Y) =

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_i - EX\right)\left(y_j - EY\right)p_{ij}$$

Covariance for continuous random variables

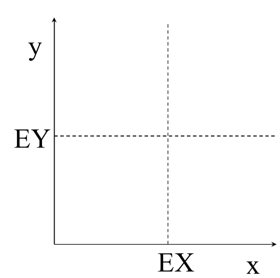
- Suppose X and Y are continuous random variables.
- Cov(X,Y) =

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x - EX\right)\left(y - EY\right) f_{xy}\left(x, y\right) dydx$$

Meaning of positive covariance

If Cov(X,Y)>0, then

- When X is above its mean, Y is usually also above its mean.
- When X is below its mean, Y is usually also below its mean.

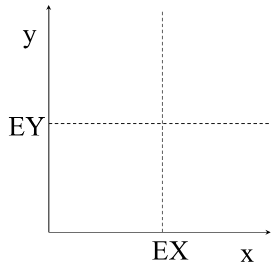


COVARIANCE, CORRELATION, AND CONDITIONAL EXPECTATION

Meaning of negative covariance

If Cov(X,Y)<0, then

- When X is above its mean, Y is usually also below its mean.
- When X is below its mean, Y is usually also above its mean.



Alternative expressions for covariance

- It can be shown that **Cov(X,Y)**
  = E [(X-EX)(Y-EY)]
  = E [(X-EX) Y]
  = E [X (Y-EY)]
  = E(XY) – EX EY.
- Note that if EX=0 or EY=0, then Cov(X,Y) = ____________.

Properties of covariance

- Covariance can be, positive, negative, or zero.
- If X and Y are independent, Cov(X,Y)=0.
  - But converse is not necessarily true.
- Cov (aX+b, cY+d) = ac Cov(X,Y).
- Cov (X,X) = Var(X).
- |Cov (X,Y)| ≤ SD(X) SD(Y) ["Schwarz inequality"].

Variance of sum of random variables

- **Var(X+Y)**
  $= E[(X+Y) - (EX+EY)]^2$
  $= E[(X-EX) + (Y-EY)]^2$
  $= E(X-EX)^2 + E(Y-EY)^2$
  $\quad + 2E[(X-EX)(Y-EY)]$
  = Var(X) + Var(Y) + 2 Cov(X,Y).
- **Var(aX+bY) =**
  a^2 Var(X) + b^2 Var(Y) + 2 ab Cov(X,Y).

Examples

- Suppose Var(X) = 4, Var(Y) = 9, and Cov(X,Y) = -3.
- Then Var(X+Y) = 4 + 9 + 2(-3) = ______.
- Also, Var(3X + 5Y)
  = 9(4) + 25(9) + 2(3)(5)(-3)
  = ________.

Special case: variance of sums of random variables with no covariance

- If Cov(X,Y) = 0,
  - Var(X+Y) = Var(X) + Var(Y).
  - Var(X-Y) = Var(X) + Var(Y).
- If $X_1, X_2, \ldots, X_n$ all have pairwise zero covariance,

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$

COVARIANCE, CORRELATION, AND CONDITIONAL EXPECTATION

Example

- Suppose $Var(X_i) = 7.2$ for $i=1, ..., 20$ and the X_i have pairwise zero covariance.
- Then $$Var\left(\sum_{i=1}^{20} X_i\right) = \sum_{i=1}^{20} Var(X_i)$$

Correlation coefficient

- *Covariance divided by product of standard deviations.*
- Corr(X,Y) = Cov(X,Y) / [SD(X) SD(Y)].
- By Schwarz inequality, $-1 \leq Corr(X,Y) \leq 1$.

Properties of correlation coefficient

- Corr (aX+b, cY+d)
  = Corr(X,Y) if (ac)>0.
  = - Corr(X,Y) if (ac)<0.
  - Thus correlation is unaffected by scaling, only by sign.
- Corr (X,X) = ______.
- Corr (X,-X) = ______.

Y given X

- Often we want to know what value Y will likely take, *given* that X takes some given value.
- Example: What wage (Y) will a person likely earn, *given* that person has 16 years of education (X)?
- Example: What will be tax revenue (Y) *given* that GDP (X) is 3 percent higher than last year?

Conditional expectation

- Covariance and correlation coefficient cannot answer this kind of question.
- We need *conditional mean or expectation*.
- E(Y|X=x) = expected value of Y given that X takes the particular value x.
- Expectation is computed using the _______________ distribution.

Formulas for conditional expectation

- If Y is discrete,
$$E(Y \mid X = x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j \mid x)$$
- If Y is continuous,
$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) dy$$

COVARIANCE, CORRELATION, AND CONDITIONAL EXPECTATION

Calculating E(Y|X=x) if Y is discrete: example

- First compute marginal probabilities for X.
- Then compute conditional probabilities given X=1: f(1|X=1)=______________, f(2|X=1)=_____, f(3|X=1)=_____.

	Y=1	Y=2	Y=3
X=1	0.2	0.1	0.1
X=2	0.3	0.2	0.1

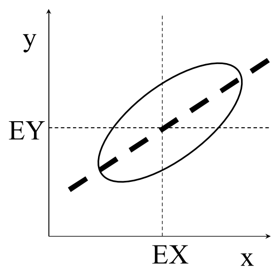
Calculating E(Y|X=x) if Y is discrete: example (cont'd)

- Finally, use the conditional probabilities to compute the conditional mean: E(Y|X=1) = 0.5(1)+0.25(2)+0.25(3)=_____

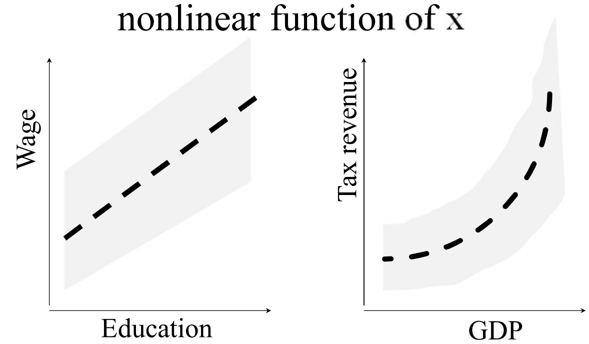
	Y=1	Y=2	Y=3	Marginal X
X=1	0.2	0.1	0.1	0.4
X=2	0.3	0.2	0.1	0.6

Conditional expectation as a function

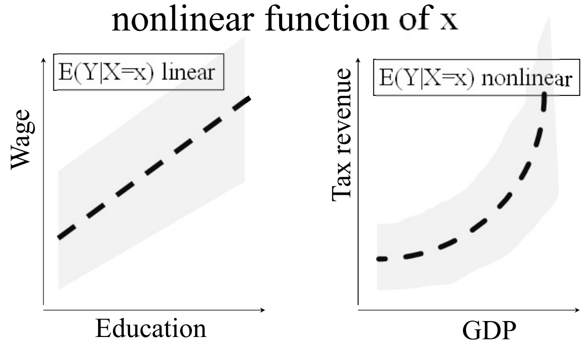
- E(Y|X=x) can be viewed as a function of x.
- But E(Y|X=x) is NOT the inverse of E(X|Y=y).



E(Y|X=x) can be a linear or nonlinear function of x



E(Y|X=x) can be a linear or nonlinear function of x



Conditional mean is the best predictor

- Suppose we want to predict someone's wage given her or his education.
- No prediction is 100% accurate, but suppose we want to choose a prediction formula that minimizes the mean squared prediction error.
- It can be shown that our best choice for a predictor is the ______________________: E(wage|education).

COVARIANCE, CORRELATION, AND CONDITIONAL EXPECTATION

Conditional mean is the best predictor (cont'd)

- Suppose we want to predict tax revenues given GDP.
- Again, suppose we want to minimize the mean squared prediction error.
- Then our best choice for a predictor is the ______________________: E(tax revenue|GDP).

Properties of conditional expectation

- If X and Y are independent, then E(Y|X=x) = E(Y) and thus not a function of x.*
- If E(Y|X=x) = E(Y), then
  - Cov(X,Y) = 0 = Corr(X,Y).
  - Any function of X is uncorrelated with Y.

*But the converse is not true.

Conditional variance

- Var(Y|X=x) = Variance of Y given that X takes the particular value x.
  - Variance is taken around the ______________ mean, using the ______________ distribution.
- If X and Y are independent, Var(Y|X=x) = Var(Y).

Conclusions

- Covariance and the correlation coefficient are measures of association.
- Covariance can take any value, but the correlation coefficient is bounded between ______________.
- Conditional expectation gives the expected value of one random variable __________ a value of another.

THE BERNOULLI DISTRIBUTION

THE BERNOULLI DISTRIBUTION

- An important discrete distribution.

A two-valued random variable

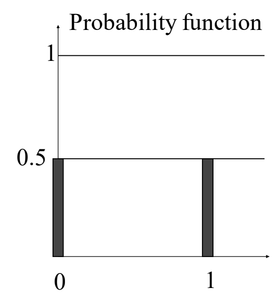
- Suppose a random variable can take only two values, say zero and one.
- Let p = Prob{X=1}.
- Then Prob{X=0} =________.

Mean and variance

- E(X) = p (1) + (1-p) (0) = _____.
- Var(X) = P $(1\text{-}p)^2$ + (1-p) $(0\text{-}p)^2$
  = p $(1\text{-}p)^2$ + (1-p) p^2
  = p (1-p) [(1-p) + p]
  = __________.

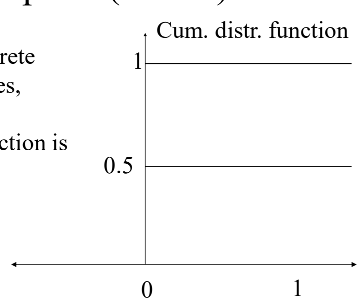
Example 1

- Suppose X is distributed as Bernoulli with parameter p = 0.5.
- Then E(X) = ______.
- And Var(X) = p (1-p) = ______.



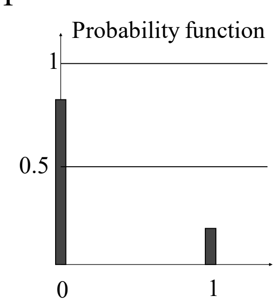
Example 1 (cont'd)

- As with all discrete random variables, the cumulative distribution function is a step-function.



Example 2

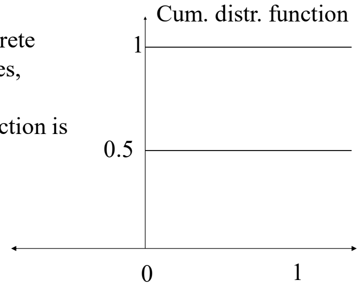
- Suppose X is distributed as Bernoulli with parameter p = 0.2.
- Then E(X) = _____.
- And Var(X) = _____.



THE BERNOULLI DISTRIBUTION

Example 2 (cont'd)

- As with all discrete random variables, the cumulative distribution function is a step-function.



Conclusions

- A Bernoulli random variable takes two values, zero and one.
- If Prob{X=1} = p, the E(X) = _____ and Var(X) = _________.
- The cumulative distribution function has two steps.

THE NORMAL DISTRIBUTION

THE NORMAL DISTRIBUTION

- An important continuous distribution.

Definition of the normal distribution

- Density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

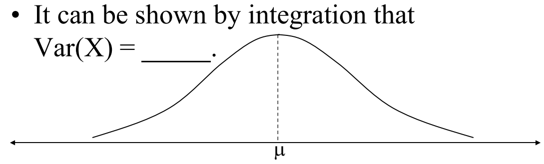
- A bell-shaped curve, symmetric about x = ______.
- Note that as x gets far from μ, the term in parentheses gets more negative, so f(x) approaches ____________.

Normally-distributed random variables

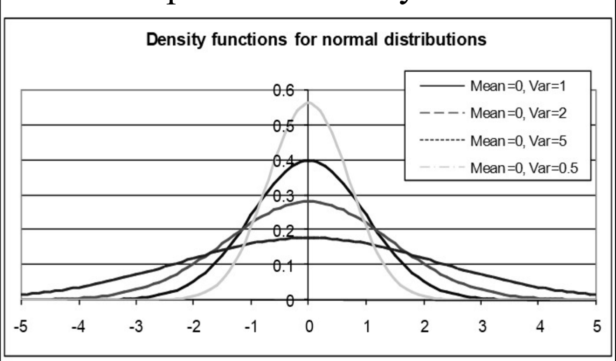
- Are continuous random variables.
- Can take any real-number value—positive, negative or zero.
- Notation: "X ~ N(μ, σ^2)" means "X is normally-distributed with parameters μ and σ^2."

Mean and variance

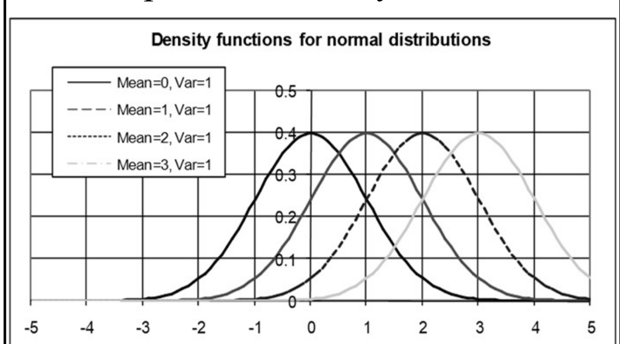
- It can be shown by integration that E(X) = ______.
- Because distribution is symmetric around μ, μ is also the median and the mode.
- It can be shown by integration that Var(X) = _____.



How the σ^2 parameter affects the shape of the density function

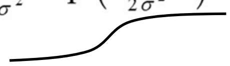


How the μ parameter affects the shape of the density function



THE NORMAL DISTRIBUTION

Cumulative distribution function

- Cumulative distribution function is
$$\Phi\left(z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$
An S-shaped curve.

- Integral has no closed form—no simple formula.
- But function available in Excel and statistical software.

Linear functions of normal random variables are also normal

- If $X \sim N(\mu, \sigma^2)$, then $Z = aX + b$ is also normally distributed.
- Using the formulas for linear functions of any random variable, $E(Z) = a\mu+b$ and $Var(Z) = a^2\sigma^2$.
- So $Z \sim N(a\mu+b, a^2\sigma^2)$.

Joint normal distribution

- If X and Y are jointly normally distributed and their covariance is zero, then X and Y are independent.
  - Recall: for other distributions, zero covariance does not necessarily imply independence. But here it does!
- Any linear combination of jointly normal random variables is also normal.

The standard normal distribution

- *The special normal distribution with $\mu = 0$ and $\sigma^2 = 1$.*
- Thus, standard normal: $Z \sim N(0,1)$.
- Functions for standard normal cumulative distribution are available in Excel and statistical software.

"Standardizing" a normal distribution

- Suppose $X \sim N(\mu, \sigma^2)$.
- Then $Z = (X-\mu)/\sigma \sim N(0,1)$.
- So Prob{ X< a }
= Prob{ $(X-\mu)/\sigma < (a-\mu)/\sigma$ }
= Prob{ $Z < (a-\mu)/\sigma$ }
- In words, if we subtract the mean and divide by the standard deviation, we have a ______________ normal random variable.

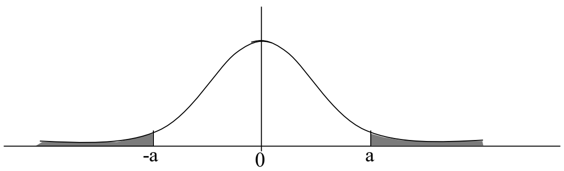
"Standardizing" a normal distribution: example

- Suppose $X \sim N(2, 9)$.
- What is Prob{X< 4} ?
- Prob{X< 4} = Prob{ Z < (4-2)/3 }
= Prob{Z < 0.67}, where $Z \sim N(0,1)$.
- From table of standard normal cumulative distribution in textbook,
Prob{X<4} = Prob{Z < 0.67} = 0.7486 .

THE NORMAL DISTRIBUTION

Standard normal is symmetric around zero

- Prob{Z<-a} = Prob{Z>a}.
- Prob{|Z|>a} = 2 Prob{Z>a}.



Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are independent identically-distributed random variables (not necessarily normal) each with mean $E(X_i)=\mu$ and variance $Var(X_i)=\sigma^2$.
- Then

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}} \overset{A}{\sim} N(0,1)$$

Another way of stating the Central Limit Theorem

- This result can alternatively be expressed as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{A}{\sim} N\left(\mu, \tfrac{\sigma^2}{n}\right)$$

- This asymptotic normal distribution gets more accurate as n increases.

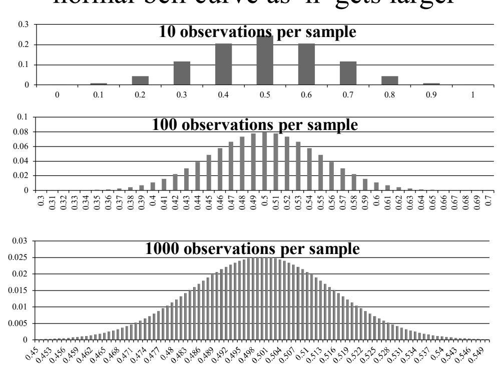
What the Central Limit Theorem means

- Suppose we compute the sample mean $\overline{X}$ from a random sample $X_1, \ldots, X_n$.
- Now $\overline{X}$ is itself a _____________ variable, varying randomly from one sample to the next.
- If the samples are large enough, then $\overline{X}$ will behave as if it were normally distributed, *regardless of distribution of X_i*.

What the Central Limit Theorem means: example

- Example: Suppose a sample of n values X_i are drawn from a Bernoulli distribution with mean p = 0.5.
- Then we compute $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- If we draw many samples, and compute $\overline{X}$ each time, what does the distribution of $\overline{X}$ look like?

Distribution of $\overline{X}$ looks increasingly like a normal bell curve as n gets larger



THE NORMAL DISTRIBUTION

Applying the Central Limit Theorem

- Suppose X_i , i=1, ..., 100 are Bernoulli random variables with Prob{X_i=1} = 0.4.
- Then E(X_i) = 0.4 and Var(X_i) = 0.24.
- By the Central Limit Theorem,

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i \overset{A}{\sim} N\left(0.4, \frac{0.24}{100}\right)$$

Applying the Central Limit Theorem

- Suppose X_i , i=1, ..., 100 are Bernoulli random variables with Prob{X_i=1} = 0.4.
- Then E(X_i) = 0.4 and Var(X_i) = 0.24.
- By the Central Limit Theorem,

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i \overset{A}{\sim} N\left(0.4, \frac{0.24}{100}\right)$$

Bell-shaped curve

$$= \boxed{N}(0.4, 0.0024)$$

Conclusions

- The normal distribution is a _______________ distribution with a bell-shaped density function.
- The ____________ normal distribution has mean = 0 and variance = 1.
- Linear functions of joint normal random variables are also normally-distributed.
- The sample mean of n independent identically-distributed random variables is _______________ normally-distributed.

DISTRIBUTIONS RELATED TO THE NORMAL DISTRIBUTION

DISTRIBUTIONS RELATED TO THE NORMAL DISTRIBUTION

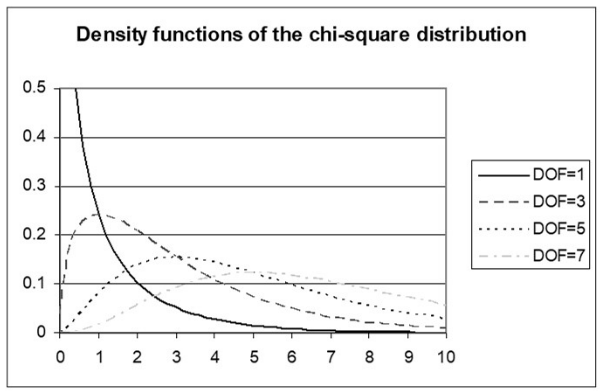
- What is the chi-square distribution?
- What is the t distribution?
- What is the F distribution?

The chi-square distribution

- Suppose $Z_1, Z_2, \ldots, Z_n$ are independent N(0,1) random variables. Then
$$Y = \sum_{i=1}^{n} Z_i^2$$
is distributed as chi-square with n degrees of freedom (DOF=n).
- Notation: $Y \sim \chi^2(n)$.

Properties of chi-square random variables

- $E(Y) = n$, and $Var(Y) = 2n$.
- $Y \geq 0$, so distribution is skewed to right. However, becomes more symmetric as n gets large.
- If $Y_1 \sim \chi^2(n_1)$ and $Y_2 \sim \chi^2(n_2)$, then $Y = (Y_1+Y_2) \sim \chi^2(n_1+n_2)$.



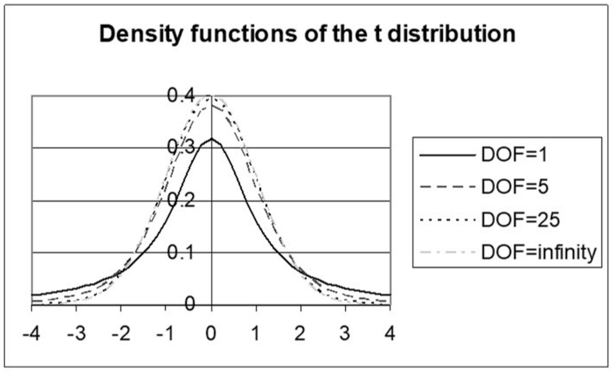
The t distribution

- Suppose $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$ are independent random variables. Then
$$W = \frac{Z}{\sqrt{Y/n}}$$
is distributed as t with n degrees of freedom (DOF=n).
- Notation: $W \sim t(n)$.

Properties of the t distribution

- Density is bell-shaped curve, symmetric around zero (=median and mode).
- Density > 0 for all x (never touches axis).
- $E(W) = 0$, for $n > 1$.
- $Var(W) = n/(n-2) > 1$, for $n > 2$.
- As n approaches infinity, t(n) approaches N(0,1).

DISTRIBUTIONS RELATED TO THE NORMAL DISTRIBUTION

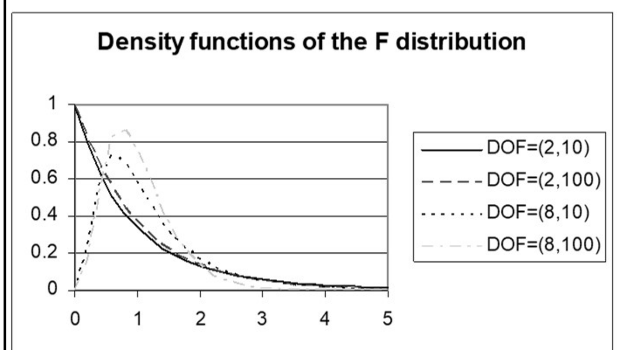


The F distribution

- Suppose $Y_1 \sim \chi^2(n_1)$ and $Y_2 \sim \chi^2(n_2)$ are independent random variables. Then
$$V = \frac{Y_1 / n_1}{Y_2 / n_2}$$
is distributed as F with n_1 degrees of freedom in the numerator and n_2 degrees of freedom in the denominator.
- Notation: $V \sim F(n_1, n_2)$.

Properties of the F distribution

- $V \geq 0$, so distribution is skewed to right. However, becomes more symmetric as n_1 gets large.
- t and F distributions are related:
  If $W \sim t(n)$, then $W^2 \sim F(1,n)$.
- $E(V) = n_2/(n_2-2)$. Thus $E(V)$ approaches 1 as n_2 approaches infinity.



Conclusions

- A ___________ random variable takes only positive values and its mean equals its DOF.
- A __________ random variable is similar to a standard normal random variable, but it has fatter tails.
- An __________ random variable is similar to a chi-square, but its mean approaches ______ as the DOF in the denominator become large.

RANDOM SAMPLES AND ESTIMATORS

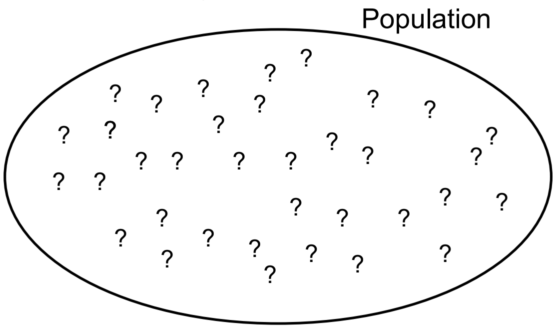
RANDOM SAMPLES AND ESTIMATORS

•What can a small sample tell us about a larger population?

Random samples

- A random sample is a set of observations chosen at random from a fixed larger population.
- Example: A random sample of heights might be taken by choosing individuals at random (e.g., from a phone book or roster) and measuring them.

A sample is an observed subset of a larger population

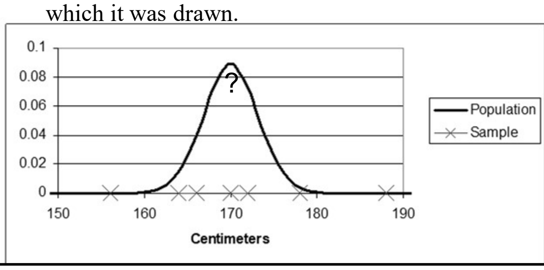


The problem of estimation

- Given a random sample, we wish to guess the population distribution from which it was drawn.
- We cannot observe the entire population, however, due to financial or physical constraints.
- The sample is _______________, whereas the population distribution is fixed but ___________________.

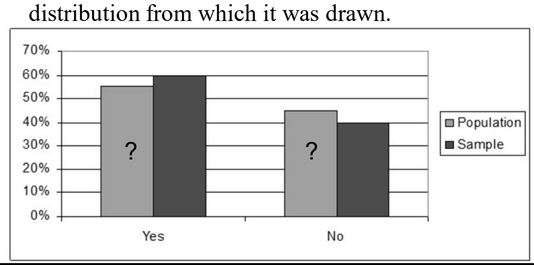
Example for continuous population distribution

- Given a sample of heights of students, we might wish to guess the population distribution from which it was drawn.



Example for discrete population distribution

- Given a sample of opinions from a telephone survey, we might wish to guess the population distribution from which it was drawn.



RANDOM SAMPLES AND ESTIMATORS

The parametric approach to estimation

- We can simplify the problem by assuming the general form of the distribution (e.g., Bernoulli, normal, etc.).
- The problem is thus reduced to finding the true population values of a few unknown parameters (e.g., p for Bernoulli, μ and σ^2 for normal, etc.).

"Estimator" versus "estimate"

- An *estimator* is a formula, that when applied to the data in a sample, gives a value for the unknown parameters.
- Estimators for unknown parameters are typically denoted with "^".
- Since sample data are random (they vary from sample to sample) the estimator is itself a ___________________.
- An *estimate* is a particular value taken by the estimator for a particular set of data.

"Estimator" versus "estimate": example

- Suppose we wish to guess the population distribution of heights of all Drake students using a sample of 10 students.
- Taking a *parametric approach*, we assume the population distribution is normal, and seek to estimate μ and σ^2 .

"Estimator" versus "estimate": example (cont'd)

- As our *estimator* for μ, we might choose the sample mean:

$$\hat{\mu} = \overline{X} = \tfrac{1}{n}\sum X_i$$

- Using measurements x_i from our sample of 10 students, we apply our estimator and compute an *estimate* of

$$\hat{\mu} = \bar{x} = 175 \text{ centimeters.}$$

Many estimators for the same unknown parameter

- Many estimators can be used to estimate the same unknown parameter.
- For example, suppose we have a random sample $X_1, \ldots, X_n$ which we believe is drawn from $N(\mu,\sigma^2)$ where μ and σ^2 are unknown parameters.
- Here are four possible estimators for μ and two possible estimators for σ^2 .

Some possible estimators for μ

$$\hat{\mu}_1 = \overline{X} = \left(\tfrac{1}{n}\right)\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_2 = \text{sample median of } \{X_i\}$$

$$\hat{\mu}_3 = \left(\tfrac{1}{n+1}\right)\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_4 = 47$$

RANDOM SAMPLES AND ESTIMATORS

Some possible estimators for σ²

$$\hat{\sigma}_1^2 = \left(\tfrac{1}{n}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$\hat{\sigma}_2^2 = \left(\tfrac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

No estimator is perfect

- Estimators (except for silly ones) are random variables whose values vary from sample to sample.
- Estimators are __________ guaranteed to equal or even to be close to the fixed but unknown true population values.
- So if we want to estimate some parameter, which estimator should we choose?

What makes a good estimator?

- We want the estimator that is "closest" to the true value of the unknown parameter "most" of the time.
- But what does that mean? Need more precise criteria. Need to compare the *properties* of alternative estimators.

Two kinds of properties of estimators

- *Exact or "small-sample" properties:* Hold exactly for any sample.
  - Most useful criteria, but may be difficult to evaluate.
- *Asymptotic or "large sample" properties:* Approximate tendencies of estimators as sample size increases.
  - Asymptotic properties hold approximately if the sample size is large.

Conclusion

- An ________________ is a formula. Its value varies randomly from sample to sample.
- An ________________ is the particular value taken by that formula in a particular sample.
- ___________-sample properties describe the behavior of an estimator exactly.
- ___________-sample properties describe the approximate tendencies of the estimator as the sample size increases.

EXACT FINITE-SAMPLE PROPERTIES OF ESTIMATORS

EXACT FINITE-SAMPLE PROPERTIES OF ESTIMATORS

- What are "bias," "variance," and "mean squared error"?
- What makes an estimator "linear" or "best-unbiased"?

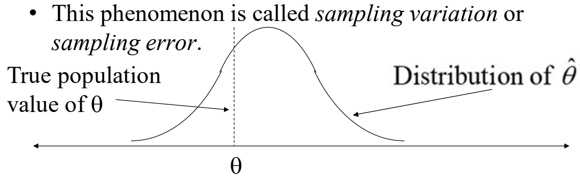
Evaluating exact finite-sample properties of estimators

- In this slideshow, we define some properties of estimators that describe their behavior exactly, even in small samples.
- We then evaluate those properties for the estimators of the mean and variance of a normal distribution, defined in a previous slideshow:

$$\left\{\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4\right\} \text{ and } \left\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\right\}$$

Estimators are themselves random variables

- Estimators are computed from observations, sampled at random.
- From one sample to another, the same estimator yields different values—it varies randomly.
- This phenomenon is called *sampling variation* or *sampling error*.



Means and variances of estimators

- Estimators have their own means and variances, which may be different from those of the population:

$$E\left(\hat{\theta}\right) \text{ and } Var\left(\hat{\theta}\right)$$

Definition of linear estimator

- A *linear estimator* is a linear function of the observations $X_1, X_2, \ldots, X_n$.
- It has the general form $a_1X_1+a_2X_2+\ldots+a_nX_n$ or

$$\hat{\theta} = \sum_{i=1}^{n} a_i X_i$$

where $a_1, a_2, \ldots, a_n$ are constant numbers.

Why linearity a useful property

- Linearity is not itself a desirable property.
- But linear estimators have relatively simple formulas.
- It is much easier to find formulas for the mean and variance of a linear estimator than of a nonlinear estimator.

EXACT FINITE-SAMPLE PROPERTIES OF ESTIMATORS

Are these estimators linear?

$$\hat{\mu}_1 = \overline{X} = \left(\tfrac{1}{n}\right)\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_2 = \text{sample median of } \{X_i\}$$

$$\hat{\mu}_3 = \left(\tfrac{1}{n+1}\right)\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_4 = 47$$

Are these estimators linear?

$$\hat{\sigma}_1^2 = \left(\tfrac{1}{n}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$\hat{\sigma}_2^2 = \left(\tfrac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

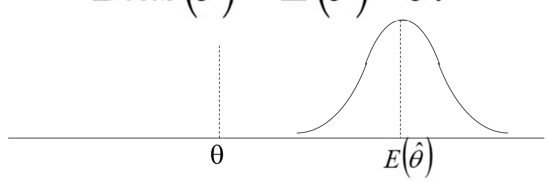
Mean and variance of linear functions of random variables

- Recall that the mean of a linear function equals the linear function of the means: $E(a_1X_1+a_2X_2+\ldots+a_nX_n) =$
- Also, if the random variables have zero covariances, the variance of the sum equals the sum of the variances: $Var(a_1X_1+a_2X_2+\ldots+a_nX_n) =$

Definition of bias

- Suppose $\hat{\theta}$ is an estimator for the true unknown population parameter θ. Then:

$$Bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta.$$

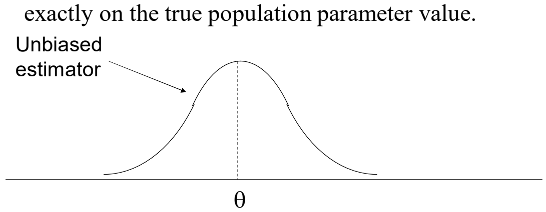


Why bias is an undesirable property

- Bias measures the difference between the mean of the estimator and the true population parameter we are trying to estimate.
- All else equal, estimators with _________________ bias are desirable.

Biased versus unbiased estimators

- If $E\left(\hat{\theta}\right) = \theta$, that is, if the estimator's bias is zero, then the estimator is called *unbiased.*
- An unbiased estimator's distribution is centered exactly on the true population parameter value.



EXACT FINITE-SAMPLE PROPERTIES OF ESTIMATORS

Specific examples: finding formulas for bias in estimators for μ

$$E(\hat{\mu}_1) = E\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i\right) = \tfrac{1}{n} E\left(\sum_{i=1}^{n} X_i\right)$$

$$= \tfrac{1}{n}\left(\sum_{i=1}^{n} EX_i\right) = \tfrac{1}{n}\left(\sum_{i=1}^{n} \mu\right) = \mu$$

- So $\hat{\mu}_1$ is unbiased.
- It can also be shown that $\hat{\mu}_2$, the sample median, is unbiased when sampling from a normal distribution.

Specific examples: finding formulas for bias in estimators for μ (cont'd)

$$E(\hat{\mu}_3) = \tfrac{1}{n+1}\left(\sum_{i=1}^{n} \mu\right) = \frac{n\mu}{n+1},$$

$$\text{so } Bias(\hat{\mu}_3) = \frac{n\mu}{n+1} - \mu = \frac{-\mu}{n+1}.$$

- Thus, $\hat{\mu}_3$ is __________.
- Also, the silly estimator $\hat{\mu}_4$ is biased, since $Bias(\hat{\mu}_4) = E(\hat{\mu}_4) - \mu = 47 - \mu \neq 0.$

Specific examples: finding formulas for bias in estimators for σ^2

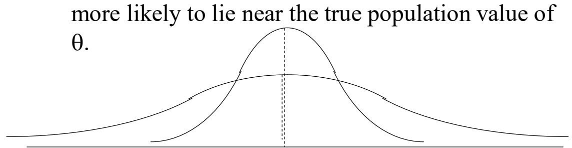
- It is not hard to show that
  $$E(\hat{\sigma}_1^2) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2,$$
  so $\hat{\sigma}_1^2$ is biased: $Bias(\hat{\sigma}_1^2) = -\sigma^2 / n$.
- However, $E(\hat{\sigma}_2^2) = \left(\frac{n-1}{n-1}\right)\sigma^2 = \sigma^2,$
  so $\hat{\sigma}_2^2$ is unbiased.

Definition of variance of estimators

- Most estimators have variance:
  $$E\left(\hat{\theta} - E\hat{\theta}\right)^2$$
- If an estimator is unbiased, then its variance is given by $$E\left(\hat{\theta} - \theta\right)^2$$

Why low variance is a desirable property

- Suppose two estimators have the same mean.
- If one estimator has lower variance than another, then its distribution is bunched more tightly.
- Assuming the first estimator's bias is low, it is more likely to lie near the true population value of θ.



θ

Definition of minimum variance

- An estimator $\hat{\theta}$ is *minimum variance* (MV) among a given set of alternative estimators if it has the smallest variance, whatever the true population value of θ.
- It is easy to find estimators with very low variance, but often they are not good estimators for other reasons.

EXACT FINITE-SAMPLE PROPERTIES OF ESTIMATORS

Specific examples: finding formulas for variance of estimators for μ

- We can use the rules for variances of linear functions, and the fact that observations in a random sample have zero covariance, to get:

$$Var(\hat{\mu}_1) = \left(\tfrac{1}{n}\right)^2 \sum_{i=1}^{n} Var(X_i) = \left(\tfrac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

$$Var(\hat{\mu}_3) = \left(\tfrac{1}{n+1}\right)^2 \sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{(n+1)^2}.$$

$$Var(\hat{\mu}_4) = 0.$$

Specific examples: finding formulas for variance of estimators for μ (cont'd)

- Var($\hat{\mu}_2$) is quite complicated, but when n is large, it equals approximately

$$Var(\hat{\mu}_2) \approx \left(\frac{\pi}{2}\right)\left(\frac{\sigma^2}{n}\right).$$

- It follows that

$$Var(\hat{\mu}_4) < Var(\hat{\mu}_3) < Var(\hat{\mu}_1) < Var(\hat{\mu}_2).$$

- So the silly estimator $\hat{\mu}_4$ is MV in this set of four estimators, followed by $\hat{\mu}_3$.

Variance of estimators for σ^2

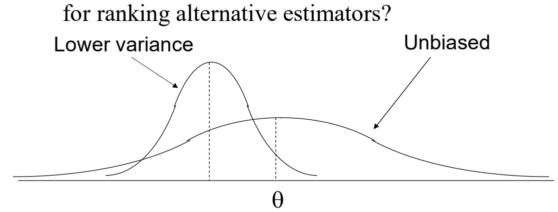
- It can be shown that

$$Var(\hat{\sigma}_1^2) < Var(\hat{\sigma}_2^2)$$

so $\hat{\sigma}_1^2$ is MV in this set of two estimators.

Trading off properties

- Low bias and low variance are both desirable properties, usually.
- How can we combine them into a single criterion for ranking alternative estimators?



Ruling out silly estimators

- From another perspective, how can we rank alternative estimators in a reasonable way, yet rule out silly estimators like $\hat{\mu}_4$?
- Two ways: *best unbiased*, and *mean square error.*
  1. *Best unbiased* simply ignores biased estimators.
  2. *Mean square error* combines bias and variance into a single formula.

1. Definition of *best unbiased*

- An estimator $\hat{\theta}$ is the *best unbiased estimator* (BUE) if it is MV among all possible unbiased estimators of θ .
- Common synonyms for *best unbiased*:
  - *Efficient.*
  - *Uniformly minimum-variance unbiased estimator* (UMVUE).

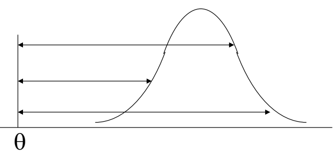
EXACT FINITE-SAMPLE PROPERTIES OF ESTIMATORS

Examples of best unbiased estimators

- It can be shown that $\hat{\mu}_1$ is the BUE for μ when sampling from a normal distribution.
- It can be shown that $\hat{\sigma}_2^2$ is the BUE for σ^2 when sampling from a normal distribution.

2. Definition of *mean squared error*

- Definition: $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$.
- In words: For each possible value of estimator, find the distance to true parameter value, square this distance, multiply by associated probability, and sum.
- Estimators with _____ MSE are desirable.



Useful formulas for MSE

- If $\hat{\theta}$ is unbiased, then
$$MSE(\hat{\theta}) = Var(\hat{\theta}).$$
- More generally, it is not hard to show that
$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2.$$

Specific examples: finding formulas for MSE of estimators for μ

- Since $\hat{\mu}_1$ and $\hat{\mu}_2$ are both unbiased, their MSEs are equal to their variances. So
$$MSE(\hat{\mu}_1) < MSE(\hat{\mu}_2).$$
- By contrast, $\hat{\mu}_3$ is biased, so its MSE is given by
$$MSE(\hat{\mu}_3) = \frac{n\sigma^2}{(n+1)^2} + \left(\frac{-\mu}{n+1}\right)^2 = \frac{n\sigma^2 + \mu^2}{(n+1)^2}$$
- This expression will be much larger than $MSE(\hat{\mu}_1) = \frac{\sigma^2}{n}$ if μ is large, relative to n and σ^2.

Specific examples: finding formulas for MSE of estimators for μ (cont'd)

- Since the silly estimator $\hat{\mu}_4$ has zero variance, its MSE is given by the square of its bias: $(47-\mu)^2$.
- Unlike the MSEs for the other three estimators, this MSE does *not* decrease as the sample size (n) increases.
- So the MSEs of the other three must eventually beat this one (unless by some miracle the true value of μ is exactly 47 !).

MSE of estimators for σ^2

- It can be shown that
$$MSE(\hat{\sigma}_1^2) < MSE(\hat{\sigma}_2^2).$$

even though $\hat{\sigma}_1^2$ is biased.

EXACT FINITE-SAMPLE PROPERTIES OF ESTIMATORS

Conclusion

- Estimators are random variables, and have their own means and variances.
- An _____________ estimator has mean equal to the true population parameter.
- A ________________________________ (or efficient) estimator has the lowest variance among unbiased estimators.
- The ______________________ of an estimator is the sum of its variance and the square of its bias.
- A good estimator has low bias, low variance, and especially low MSE.

ASYMPTOTIC PROPERTIES OF ESTIMATORS

ASYMPTOTIC PROPERTIES OF ESTIMATORS

•What are "asymptotic bias" and "consistency"?

Evaluating asymptotic properties of estimators

- In this slideshow, we define two properties of estimators that describe their behavior as the sample size (n) grows without bound.
- We then evaluate those properties for specific examples of estimators defined in a previous slideshow:

$$\{\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4\} \text{ and } \{\hat{\sigma}_1^2, \hat{\sigma}_2^2\}$$

Why asymptotic properties?

- We never have an infinite sample (n= ∞) so why care about asymptotic properties?
- *Indicators of reasonableness:* A good estimator should get closer to the true value as the sample size increases.
- *Handy approximations for computation:* Asymptotic distributions are often simpler to work with than exact distributions of estimators.

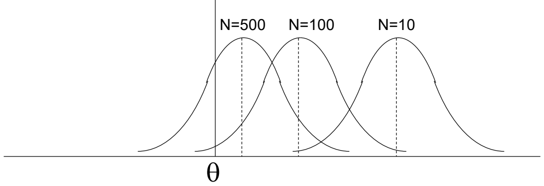
Definition of asymptotic bias

- *Asymptotic bias* of an estimator $\hat{\theta}$ is limit of bias as sample size increases without bound.
- Formally,

$$\lim_{n \to \infty} Bias(\hat{\theta}) = \lim_{n \to \infty} E(\hat{\theta}) - \theta$$

Asymptotically unbiased estimators

- A good estimator, if biased, should have a bias that disappears as the sample size increases.
- Thus estimators with __________ asymptotic bias are desirable.



"Unbiased" implies "asymptotically unbiased"

- Estimators which are unbiased in finite sample must obviously be asymptotically unbiased, too.
- Thus the sample mean $\hat{\mu}_1$, and the sample median $\hat{\mu}_2$ are asymptotically unbiased because they are unbiased in finite sample.

ASYMPTOTIC PROPERTIES OF ESTIMATORS

Checking asymptotic bias of estimators for μ

- What about the sample mean with n replaced by n+1 ($\hat{\mu}_3$) and the silly estimator ($\hat{\mu}_4$) ?
- Recall from previous presentation,
  $$Bias(\hat{\mu}_3) = \frac{-\mu}{n+1}.$$
- Also recall from previous presentation,
  $$Bias(\hat{\mu}_4) = 47 - \mu.$$

Checking asymptotic bias of estimators for σ^2

- Consider $\hat{\sigma}_1^2$, the variance estimator dividing by n. In previous slideshow, we claimed that
  $$Bias(\hat{\sigma}_1^2) = -\sigma^2 / n.$$
- Also claimed that $\hat{\sigma}_2^2$, the variance estimator dividing by (n-1), was unbiased, so it must be asymptotically unbiased.

Summary of results for example estimators

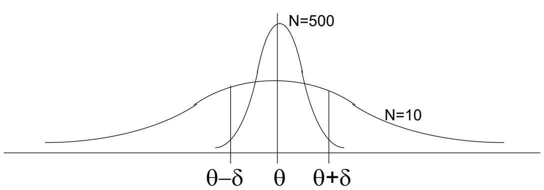
- $\hat{\mu}_1, \hat{\mu}_2$ and $\hat{\mu}_3$ are all asymptotically unbiased, so they are all good estimators by that criterion.
- The silly estimator $\hat{\mu}_4$ is __________ asymptotically unbiased, so it is not a good estimator.
- $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are both asymptotically unbiased, so they are good estimators.

Definition of consistency

- An estimator is *consistent* if the probability that the estimator is more than any given distance from the true value converges to zero as the sample grows without bound.
- Formally, $\hat{\theta}$ is consistent if for any positive number δ,
  $$\lim_{n \to \infty} \text{Prob}\left\{ \left| \hat{\theta} - \theta \right| > \delta \right\} = 0$$

Consistent estimators

- Most estimators occasionally yield values far from the true value, due to sampling variation.
- A consistent estimator does this less and less frequently, as the sample size increases.



Alternate terminology for consistency

(1) The estimator *converges in probability* to the true population parameter. Notation:

$$\hat{\theta} \xrightarrow{p} \theta$$

(2) The true population parameter is the *probability limit* of the estimator. Notation:

$$\text{plim}(\hat{\theta}) = \theta$$

ASYMPTOTIC PROPERTIES OF ESTIMATORS

Consistent estimators are desirable

- Of course, our samples usually do not grow in size spontaneously.
- Nevertheless, any estimator that would not get closer to the true value, as the sample size increased, is surely suspect.

How to check consistency

- It can be shown that if the *MSE of an estimator converges to zero* as the sample size (n) increases without bound, then the estimator is consistent.
- We now apply this handy result to the familiar estimators

$$\{\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4\} \text{ and } \{\hat{\sigma}_1^2, \hat{\sigma}_2^2\}$$

Checking consistency of estimators for μ using MSE

$$MSE(\hat{\mu}_1) = \frac{\sigma^2}{n}.$$

$$MSE(\hat{\mu}_2) \approx \left(\frac{\pi}{2}\right)\left(\frac{\sigma^2}{n}\right).$$

$$MSE(\hat{\mu}_3) = \frac{n\sigma^2}{(n+1)^2} + \frac{\mu^2}{(n+1)^2}.$$

$$MSE(\hat{\mu}_4) = (47-\mu)^2.$$

Checking consistency of estimators for σ^2

- It can be shown that the MSEs for both $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ converge to zero as n increases without bound, so both estimators are consistent.

What kinds of estimators are consistent?

- The mean of a random sample drawn from *any* distribution is a consistent estimator for the unknown true population mean, according to a theorem called the Law of Large Numbers.

  (See a mathematical statistics textbook for formal proof).

Conclusions

- An estimator is *asymptotically unbiased* if its bias converges to ______________ as n grows without bound.
- An estimator is *consistent* if the probability that it is more than any given distance from the true value converges to ______________ as the sample grows without bound.
- Good estimators should be asymptotically unbiased and consistent.

ASYMPTOTIC NORMALITY

ASYMPTOTIC NORMALITY

- What makes an estimator "asymptotically normal"?
- Why is that a useful property?

Definition of asymptotic normality

- An *asymptotically normal* estimator has a distribution which approaches the normal distribution as the sample size grows without bound.
- Notation: $\hat{\theta} \overset{A}{\sim} N\left(\theta, Var\left(\hat{\theta}\right)\right)$

  or $\dfrac{\hat{\theta} - \theta}{SD(\hat{\theta})} \overset{A}{\sim} N(0,1)$

Why asymptotic normality is useful

- Often the exact distribution of an estimator is hopelessly complicated.
- But its asymptotic distribution may provide a good approximation for large samples.
- If the asymptotic distribution is normal, we can use a normal table in a textbook, or a normal function in Excel or other software, to evaluate it.

What kinds of estimators are asymptotically normal?

- The mean of a random sample drawn from *any* distribution is asymptotically normal, according to the Central Limit Theorem.
- Almost all estimators encountered in practice can be shown to be asymptotically normal.

Example: Bernoulli distribution

- Suppose
  - $X_i = 1$ with probability = p
  - $X_i = 0$ with probabilility = ________.
- $E(X_i)$ = ______.
- $Var(X_i)$ = __________.

Example (cont'd)

- Suppose sample of size n is observed.
- Sum of the X_i can take _________ different values: 0, 1, 2, 3, ... n.
- Thus sample mean $\overline{X}$ also takes __________ different values: 0, 1/n, 2/n, 3/n, …, 1.
- Thus the exact distribution of $\overline{X}$ quickly becomes hopelessly complicated!

ASYMPTOTIC NORMALITY

Example (cont'd)

- But mean and variance of $\overline{X}$ are simple.
  - $E(\overline{X}) = E(X_i) =$ _________.
  - $Var(\overline{X}) = Var(X_i) / n =$ _____________.
- So we can apply the Central Limit Theorem to find the asymptotic distribution.

Example (cont'd)

- Applying the Central Limit Theorem, the asymptotic distribution of $\overline{X}$ must be

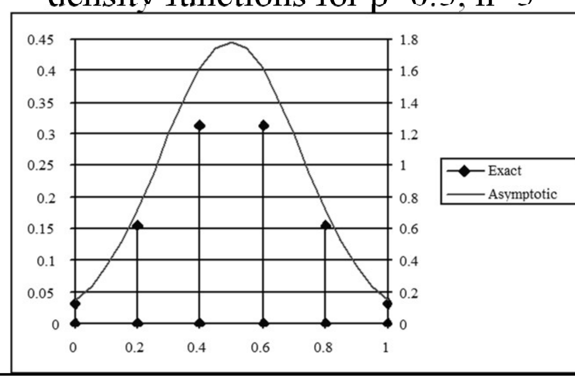
$$\overline{X} \overset{A}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

$$\text{or} \quad \frac{\overline{X}-p}{\sqrt{p(1-p)/n}} \overset{A}{\sim} N(0,1)$$

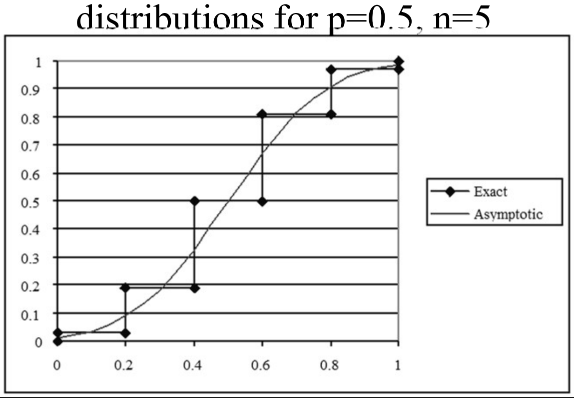
How close an approximation is the asymptotic distribution?

- How close an approximation does the Central Limit Theorem provide?
- In other words, how close is the exact distribution of $\overline{X}$ to its asymptotic distribution?
- The following charts compare the exact and asymptotic distribution functions of $\overline{X}$ for p=0.5, and n = 5, 10, and 25.

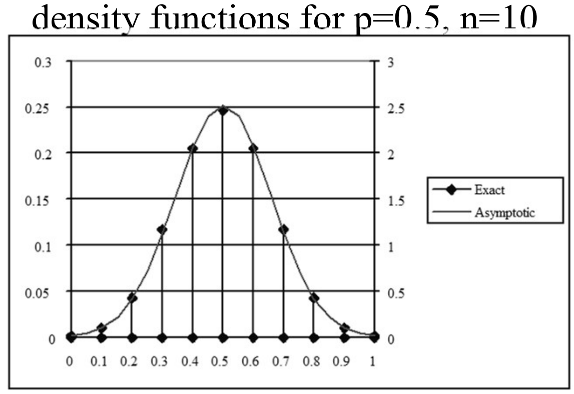
Exact probability and asymptotic density functions for p=0.5, n=5



Exact and asymptotic cumulative distributions for p=0.5, n=5

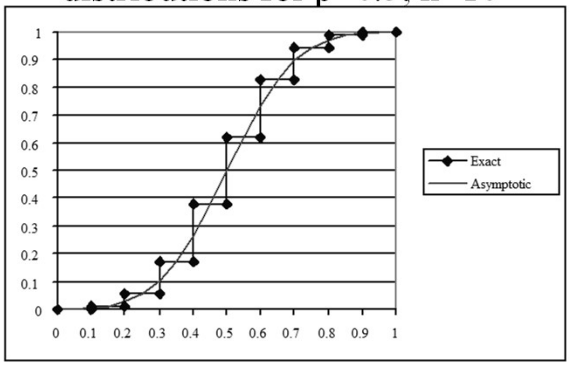


Exact probability and asymptotic density functions for p=0.5, n=10

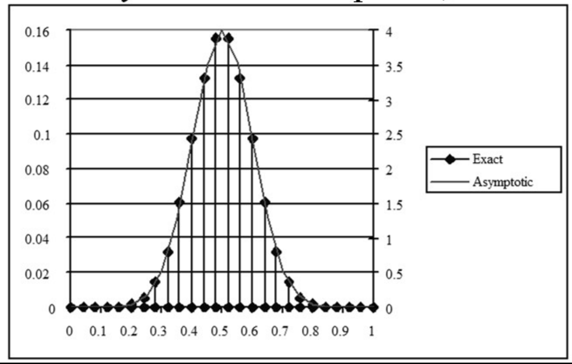


ASYMPTOTIC NORMALITY

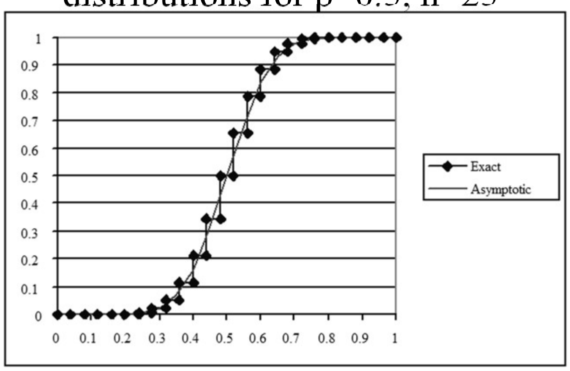
Exact and asymptotic cumulative distributions for p=0.5, n=10



Exact probability and asymptotic density functions for p=0.5, n=25



Exact and asymptotic cumulative distributions for p=0.5, n=25



Asymptotic distribution often provides a very good approximation

- Clearly, the asymptotic distribution appears to be a good approximation to the exact distribution in this example for n=25.
- Moreover, the approximation gets better as the sample size (n) increases.
- But in addition to comparing graphs, we can compare calculations of probabilities.

Using the asymptotic distribution to calculate probabilities: example

- Suppose one had a sample of 25 observations from Bernoulli distribution with population mean p = 0.5 .
- What is the probability that the sample mean would be 0.65 or greater?
- We now use the asymptotic normal distribution to compute the answer.

Example (cont'd)

$$\text{Prob}\left\{\overline{X} > 0.65\right\}$$

$$= \text{Prob}\left\{\frac{\overline{X} - 0.5}{\sqrt{0.5(1-0.5)/25}} > \frac{0.65 - 0.5}{\sqrt{0.5(1-0.5)/25}}\right\}$$

$$= \text{Prob}\left\{Z > \frac{0.65 - 0.5}{\sqrt{0.5(1-0.5)/25}}\right\}$$

$$= \text{Prob}\left\{Z > 1.5\right\} = \_\_\_\_\_\_\_\_\_\text{, according to table.}$$

- By comparison, exact probability = 0.0539 .

ASYMPTOTIC NORMALITY

Example (cont'd)

$$\text{Prob}\left\{\overline{X} > 0.65\right\}$$

$$= \text{Prob}\left\{\frac{\overline{X} - 0.5}{\sqrt{0.5(1-0.5)/25}} > \frac{0.65 - 0.5}{\sqrt{0.5(1-0.5)/25}}\right\}$$

$$= \text{Prob}\left\{Z > \frac{0.65 - 0.5}{\sqrt{0.5(1-0.5)/25}}\right\}$$

$$= \text{Prob}\left\{Z > 1.5\right\} = \underline{0.0668}, \text{ according to table.}$$

- By comparison, exact probability = 0.0539 .

What if the true population standard deviation is unknown?

- Asymptotic normality still holds when an estimate of the standard deviation is used instead of its true population value.

Conclusions

- An *asymptotically normal* estimator has a distribution which approaches the ____________ distribution as the sample size grows.
- This property gives a convenient approximation to the exact distribution of the estimator when the sample size is ____________.

RELIABLE PRINCIPLES FOR FINDING GOOD ESTIMATORS

RELIABLE PRINCIPLES FOR FINDING GOOD ESTIMATORS

- What is the "method-of-moments" principle?
- What is the "maximum-likelihood" principle?

Principles for finding estimators

- Suppose we want to estimate unknown parameters of a distribution that we believe is generating our data.
- How should we begin?
- Two general principles work very well in most cases.

(1) "Method-of-moments" principle

(2) "Maximum-likelihood" principle

(1) "Method of moments" principle

- First, find formulas for the true population moments of the distribution in terms of the unknown parameters.
- Second, set the sample moments equal to the formulas for the true population moments.
- Finally, solve for estimators of the parameters.

Application to Bernoulli sample: true population moment

- Consider coin toss with possibly unfair coin.
- Let X = 1 if heads, = 0 if tails.
- Probability of heads = p, unknown parameter. Probability of tails = 1 - p .
- The first moment is the mean:
  - E(X) = (p) 1 + (1-p) 0 = ______.

Application to Bernoulli sample: sample moment

- Suppose we have observations on n coin tosses: $X_1, \ldots X_n$.
- To estimate p using method-of-moments principle, set E(X) = p = $\overline{X}$
- Solving for p gives (immediately) a formula for the method-of-moments estimator for p:

$$\hat{p}_{MOM} = \overline{X}$$

Numerical example of MOM estimator

- Suppose out of 20 coin tosses, 12 tosses were heads.
- Then method-of-moments estimate is:

$$\hat{p}_{MOM} = \overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i =$$

RELIABLE PRINCIPLES FOR FINDING GOOD ESTIMATORS

What's so great about MOM estimators?

- Method-of-moments estimators are almost always
  - Consistent.*
  - Asymptotically normal.**
- Also, MOM estimation does not require us to make assumptions about the whole distribution of X_i , only the moments.

* Because of the Law of Large Numbers theorem.
** Because of the Central Limit Theorem.

(2) "Maximum likelihood" principle

- First, find the likelihood function of the sample.
- Second, solve for the value(s) of the parameter(s) that maximize this function.
- But what is the "likelihood function"?

The joint density function of a random sample

- Suppose we are willing to assume that our sample comes from a particular distribution.
- If we have a random sample, the observations are independent.

The joint density function of a random sample (cont'd)

- Then the joint density (or joint probability function) of the sample is the product of the individual density functions (or probability functions) of the observations:

  $f(x_1, x_2,\ldots, x_n) = f(x_1)\, f(x_2) \ldots f(x_n)$.

From joint density function to likelihood function

- The joint density function depends on the values of the observations (x_i) and the parameters of the distribution.
- If the values of observations are *known* and the values of the parameters are *unknown*, this function is called the "_______________ function."

Application to Bernoulli sample: finding the likelihood function

- Recall the Bernoulli distribution:
  - X = 1 with probability p.
  - X = 0 with probability (1-p).
- One way to represent this as a probability function is:
  $\text{Prob}\{X=x\} = f(x) = p^x (1-p)^{1-x}$
  for x = 0, 1.

RELIABLE PRINCIPLES FOR FINDING GOOD ESTIMATORS

Application to Bernoulli sample: finding the likelihood function (cont'd)

- The likelihood function is the joint density of the sample, with p viewed as unknown:

$$f(x_1,\dots,x_n;p)$$
$$=\left(p^{x_1}(1-p)^{(1-x_1)}\right)\left(p^{x_2}(1-p)^{(1-x_2)}\right)\dots\left(p^{x_n}(1-p)^{(1-x_n)}\right)$$
$$=p^{\sum x_i}(1-p)^{\left(n-\sum x_i\right)}$$

Application to Bernoulli sample: maximizing the likelihood function

- Second, maximize the likelihood function with respect to unknown p, by setting

$$0=\frac{d}{dp}\left[p^{\sum x_i}(1-p)^{\left(n-\sum x_i\right)}\right] .$$

- The derivative is a little messy, but with some algebraic manipulation reduces to

$$0 = p^{\left(\sum x_i-1\right)}(1-p)^{\left(n-\sum x_i-1\right)}\left(\sum x_i-np\right).$$

Application to Bernoulli sample: maximizing the likelihood function (cont'd)

- The derivative equals zero if and only if

$$0=\left(\sum x_i-np\right).$$

- Setting this derivative equal to zero and solving for p gives

$$\hat{p}_{ML} =$$

Comparing ML estimator with MOM estimator

- In this application, the maximum-likelihood estimator $\hat{p}_{ML}$ is identical to the method-of-moments estimator $\hat{p}_{MOM}$.
- This is often, but not always, the case.
- When it happens, the estimator is sure to be a good one!

Numerical example of ML estimator

- Suppose out of n=20 coin tosses, 12 tosses were heads.
- Then the maximum likelihood estimate is the same as the method-of-moments estimate :

$$\hat{p}_{ML}=\bar{x}=\frac{1}{20}\sum_{i=1}^{20}x_i=$$

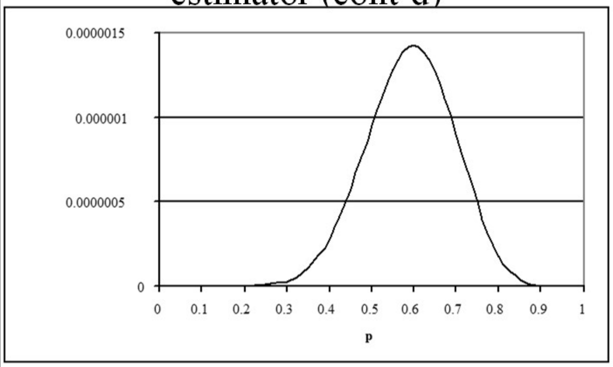
Numerical example of ML estimator (cont'd)

- But wait! Let's check graphically whether 0.6 really maximizes the likelihood function in this case.
- In this case, the likelihood function is:

$$f(x_1,\dots,x_n;p)=p^{\sum x_i}(1-p)^{n-\sum x_i}$$
$$=$$

RELIABLE PRINCIPLES FOR FINDING GOOD ESTIMATORS

Numerical example of ML estimator (cont'd)



What's so great about ML estimators?

- Maximum likelihood estimators are almost always
  - Consistent.
  - Asymptotically normal.
- Also, if they are unbiased, ML estimators are always *minimum-variance unbiased* (or "*best unbiased*" or "*efficient*").

Conclusions

- The ________________________ principle proposes equating sample moments with true (or "population") moments.
- The ________________________ principle proposes substituting data into the joint density function and finding the values of the unknown parameters that maximize this "likelihood function."

STANDARD ERRORS

STANDARD ERRORS

•What is the standard error of an estimator or estimate?
•How can it be computed for an estimator like the sample mean?

Limitations of point estimates

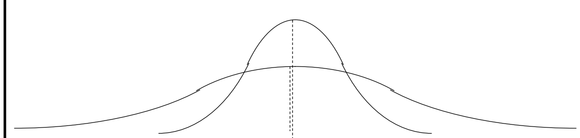
- A particular estimate, by itself, is not convincing unless we have some measure of its precision.
- Although population parameters are ___________, estimates ___________ from sample to sample because of sampling error.
- For example, if we are sampling hourly wages of workers in Des Moines, one sample might yield a sample mean of $12.35, another a sample mean of $14.07.

Variance and standard deviation of estimators

- Estimators are themselves random variables because they are computed from random samples.
- A natural measure of precision is thus the *variance* of the estimator.
- Another is the square root of the variance: the *standard deviation* of the estimator.

Low variance or standard deviation is a desirable property

- If one estimator has lower variance than another, then its distribution is bunched more tightly.
- An estimator with lower variance (or standard deviation) is more precise.



Definition of standard error

- Usually the variance and standard deviation of an estimator depend on the unknown population parameters we are trying to estimate.
- So the true variance and standard deviation of the estimator are ____________________.
- But they can be estimated.
- The estimated standard deviation is called the _____________________ (SE).

An important distinction

- Here, we do *not* want an estimate of the *population standard deviation*—that is, the standard deviation of a single observation.
- Instead we want an estimate of the *standard deviation of our estimator*, a formula based on all observations in our sample.
- But the former will usually help us find the latter.

STANDARD ERRORS

Application: the sample mean from a normal distribution

- Suppose we have a random sample of n observations from a normal population with unknown true population mean μ and unknown true population variance σ^2.
- We want to use the sample mean to estimate μ:

$$\overline{X} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i$$

- Suppose we also want to calculate the standard error of $\overline{X}$ to measure its precision.

Variance of the sample mean

- Using the theoretical properties of variance, we know that:

$$Var(\overline{X}) = Var\left(\left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i\right) =$$

- Moreover, if the observations are uncorrelated, the variance of the sum is the sum of the variances:

$$Var(\overline{X}) = \left(\frac{1}{n}\right)^2\left(\sum_{i=1}^{n} Var(X_i)\right) = \left(\frac{1}{n}\right)^2 n\sigma^2 =$$

Standard deviation of the sample mean

- Because the standard deviation is just the square root of the variance,

$$SD(\overline{X}) = \sqrt{Var(\overline{X})} =$$

- But this depends on σ^2, which is unknown.

Using an estimate of the true population variance

- We compute SE($\overline{X}$) by replacing the unknown true population σ^2 with its estimate in the formula for SD($\overline{X}$).
- A good choice for an estimate of the unknown σ^2 of a normal population is the unbiased estimate

$$\hat{\sigma}^2 = \left(\tfrac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Standard error for the sample mean from a normal distribution

- So the standard error of $\overline{X}$ from a normal population is given by:

$$SE(\overline{X}) =$$

where $\hat{\sigma}^2 = \left(\tfrac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

Example: Standard error for the sample mean from a normal distribution

- Suppose we have a sample of heights of n=50 persons, measured in centimeters.
- We have computed

$$\sum_{i=1}^{50}\left(x_i - \overline{x}\right)^2 = 1568.$$

STANDARD ERRORS

Example: Standard error for the sample mean from a normal distribution (cont'd)

- The unbiased estimator of the sample variance is

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{50}\left(x_i - \bar{x}\right)^2 = \frac{1}{49}\left(1568\right) =$$

- So the standard error of the sample mean is

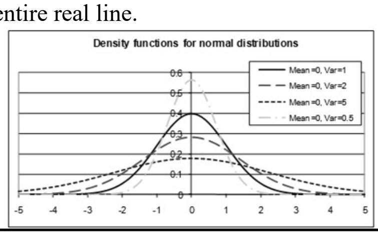
$$SE = \sqrt{\tfrac{1}{50}\hat{\sigma}^2} =$$

Standard error for the sample mean from a non-normal distribution

- Often we work with data whose population distribution is not normal.
- Examples:
  - Income (non-negative, right-skewed).
  - Family size (discrete).
  - Opinion polls (yes-no).

The normal distribution is special and does not fit all data

- Recall that the normal distribution is continuous and symmetric, with bell-shaped density function.
- Normal random variables can take any value on the entire real line.



Example of non-normal population distribution: family income

- Suppose we are investigating family income, where the parameter of interest is average household income.
- But household income cannot be normally-distributed because
  - income *cannot be negative*.
  - the population distribution of income is not symmetric. It is _______________ to the right because a few households have very high income.

Another example of non-normal population distribution: family size

- Suppose we are investigating family size, where the parameter of interest is the mean number of children in a family.
- But family size cannot be normally-distributed because
  - the number of children in a family must be a nonnegative *whole number*: 0, 1, 2, 3, etc.
  - this is obviously a ________________ random variable, bounded at zero.

Another example of non-normal population distribution: yes-no opinions

- Suppose we are investigating public opinion, where the parameter of interest is the fraction of the population that approves of the president.
- But opinion data cannot be normally-distributed because
  - opinion is either yes (X=1) or no (X=0).
  - this is a Bernoulli random variable.

STANDARD ERRORS

Exact finite-sample standard errors for non-normal populations

- In principle, standard errors can be constructed for any distribution.
- However, exact finite-sample formulas for standard errors can be very complicated if the population distribution is not normal.
- Moreover, if we are not sure of the underlying distribution, an exact finite-sample formula cannot be found.

Asymptotic ("large sample") standard errors

- An easier approach is possible if the sample is large.
- We use the same basic formula,

$$\text{asymptotic } SE(\overline{X}) = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

- But then we can use *any consistent estimator* for the population variance.

Consistent estimators for the population variance

- Consistent estimators for any population variance include

$$\hat{\sigma}^2 = \left(\tfrac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$\tilde{\sigma}^2 = \left(\tfrac{1}{n}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Consistent estimators for the population variance for Bernoulli random variables

- For yes-no data (like opinion polls) then the following is often used:

$$\tilde{\sigma}^2 = \hat{p}\left(1-\hat{p}\right)$$

where $\hat{p}$ = m/n, that is, the number of yeses divided by the total number of responses.

Example: Standard error for the sample mean from a Bernoulli distribution

- Suppose we have a sample of opinions (yes/no) of n=50 persons.
- The number of people saying "yes" is m=32, so the sample mean $\hat{p}$ = m/n =
- So $\tilde{\sigma}^2 = \hat{p}\left(1-\hat{p}\right) =$
- So the asymptotic standard error of the sample mean is $SE = \sqrt{\tfrac{1}{50}\tilde{\sigma}^2} =$

Interpreting the standard error

- The larger the standard error, the _______ precise is the estimator (in the preceding example, the sample mean).
- Honesty in research requires that whenever an estimate is reported, its standard error should be reported, too.
- Of course, the value of the standard error is itself computed from the data, so it will vary from sample to sample, just like the estimator.

STANDARD ERRORS

Conclusions

- All estimators are subject to ________________.
- Any estimate should therefore be accompanied by a measure of its ______________.
- A natural measure of precision is the standard deviation of the estimator, but usually this is ______________.
- The standard error is an ____________ of the true standard deviation of an estimator.

CONFIDENCE INTERVALS

CONFIDENCE INTERVALS

•What are confidence intervals?
•What is the formula for the CI for the mean of a normal distribution?
•What is the formula for the CI for the mean of an arbitrary distribution?

Limitations of point estimators

- We cannot estimate the true population parameter exactly if we have only a sample.
- Any estimator is subject to sampling error, varying randomly from sample to sample.
- For example, if we are sampling household incomes in Iowa, one sample might yield a sample mean of $39,150, another a sample mean of $46,400.

Measuring the precision of an estimator

- So an estimate, by itself, tells us little about the true population parameter unless we known how precise that estimate is.
- One measure of precision is the *standard error* of the estimator, discussed earlier.
- A more elaborate measure of precision is the *confidence interval* (CI), discussed here.

Bounding the true value

- We cannot know the true population parameter's value exactly.
- But it would be helpful if we could at least *bound* the true population parameter.
- For example, if we could at least say "the true mean income of Iowa households is between $41,500 and $42,250."

Bounds are necessarily random

- Can we bound the true population parameter *for sure*?
- ______! Any bounds we construct must be calculated from the sample, so they must be subject to random sampling variation, too.
- So the best we can do is construct bounds that *probably* contain the true population parameter—that is, *confidence intervals*.

Formal definition of confidence interval (CI)

- Let γ denote some level of confidence, like 80%, 90%, 95%, or 99%.
- Then a γ *confidence interval* is a pair of estimators—call them $\hat{\theta}_1$ and $\hat{\theta}_2$ —that bound the unknown true population parameter with probability γ :

$$\text{Prob}\left\{\hat{\theta}_1 \le \theta \le \hat{\theta}_2\right\} = \gamma$$

CONFIDENCE INTERVALS

Confidence interval for the mean of a normal distribution

- Suppose we have a random sample of 20 observations from a normal distribution with unknown population mean μ and unknown population variance σ^2.
- We are using the sample mean $\overline{X}$ to estimate the true population mean μ but we would like also to construct a 95% CI for μ.

Confidence interval for the mean of a normal distribution (cont'd)

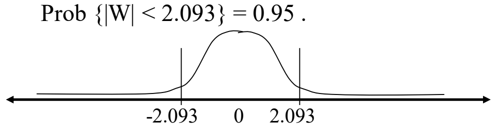
- It can be shown (see a mathematical statistics book) that the following expression, a random variable, follows a t distribution with 19 degrees of freedom:

$$W = \frac{\left(\overline{X} - \mu\right)}{SE\left(\overline{X}\right)} = \frac{\left(\overline{X} - \mu\right)}{\sqrt{\hat{\sigma}^2 / 20}}$$

where $$\hat{\sigma}^2 = \left(\tfrac{1}{19}\right)\sum_{i=1}^{20}\left(X_i - \overline{X}\right)^2$$

Using the t table

- Open your textbook to the table of the t distribution in the back. Focus on the 2-tailed probabilities.
- Depending on the format of the table, it shows that, if W ~ t(19), then either
  Prob {|W| > 2.093} = 0.05 or
  Prob {|W| < 2.093} = 0.95 .



From table values to confidence interval

- Here, W = $\left(\overline{X} - \mu\right) / SE\left(\overline{X}\right)$
- We now use the values from the table, and the formula for W, to derive formally the formula for the confidence interval.

Formal derivation of the formula for the 95% confidence interval (CI)

$$0.95 = \text{Prob}\left\{ \left| \frac{\left(\overline{X} - \mu\right)}{SE\left(\overline{X}\right)} \right| < 2.093 \right\}$$

$$= \text{Prob}\left\{ 2.093 > \frac{\left(\overline{X} - \mu\right)}{SE\left(\overline{X}\right)} > -2.093 \right\}$$

$$= \text{Prob}\left\{ 2.093\, SE\left(\overline{X}\right) > \left(\overline{X} - \mu\right) > -2.093\, SE\left(\overline{X}\right) \right\}$$

$$= \text{Prob}\left\{ -\overline{X} + 2.093\, SE\left(\overline{X}\right) > -\mu > -\overline{X} - 2.093\, SE\left(\overline{X}\right) \right\}$$

$$= \text{Prob}\left\{ \overline{X} - 2.093\, SE\left(\overline{X}\right) < \mu < \overline{X} + 2.093\, SE\left(\overline{X}\right) \right\}$$

Formula for the 95% CI with 20 observations

- Bottom line: for a random sample of 20 observations from a normal distribution, the 95% CI for the mean is

$$\left(\overline{X} - 2.093\ SE\left(\overline{X}\right),\quad \overline{X} + 2.093\ SE\left(\overline{X}\right)\right)$$
$$= \overline{X} \pm 2.093\ SE\left(\overline{X}\right)$$

where $$SE\left(\overline{X}\right) = \sqrt{\hat{\sigma}^2 / 20}$$

CONFIDENCE INTERVALS

Other levels of confidence

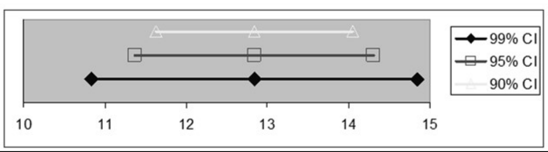
- The same t table shows that, if $W \sim t(19)$,
  Prob $\{|W| < 1.729\} = 0.90$ and
  Prob $\{|W| < 2.861\} = 0.99$.
- So for a 90% CI, replace "2.093" with "_________" in the formula.
- For an 99% CI, replace "2.093" with "_________."

Numerical example 1

- Suppose we have 20 observations from a normal distribution.
- Suppose the sample mean is $\overline{X} =$ 12.84 and the standard error is SE($\overline{X}$) = 0.7.
- Then the 95% CI is 12.84 $\pm$ 1.4651
- The 90% CI is 12.84 $\pm$ 1.2103
- The 99% CI is 12.84 $\pm$ 2.0027

Numerical example 1 (cont'd)

- Note that the higher the confidence, the ___________ the interval.
- Makes intuitive sense: to be more confident you have "captured" the unknown parameter, you must cast a wider net.



Other sample sizes

- If the sample mean is computed from a random sample of n observations from a normal distribution, then the following expression, a random variable, follows a t distribution with (n-1) degrees of freedom:

$$\left(\overline{X}-\mu\right)/\,SE\left(\overline{X}\right) \;=\; \left(\overline{X}-\mu\right)/\sqrt{\hat{\sigma}^2/n}$$

where $\hat{\sigma}^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$

General formula for the CI of the mean from sample of n observations

- So with a random sample of n observations from a normal distribution, the 95% CI for the mean is

$$\left(\overline{X}-c\,SE\left(\overline{X}\right),\,\overline{X}+c\,SE\left(\overline{X}\right)\right)$$
$$=\overline{X}\pm c\,SE\left(\overline{X}\right)$$

General formula for the CI of the mean from sample of n observations (cont'd)

- Here, $SE\left(\overline{X}\right)=\sqrt{\hat{\sigma}^2/n}$

$$\hat{\sigma}^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

and the constant c is taken from the t table with (n-1) degrees of freedom and the desired level of confidence.

CONFIDENCE INTERVALS

Numerical example 2

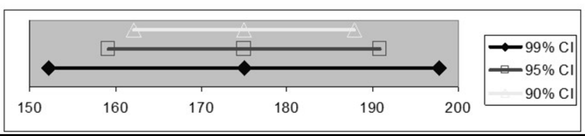
- Suppose we have 10 observations on heights of Drake students.
- Suppose the sample mean is 175 centimeters, and the standard error of the sample mean is 7 centimeters.

Numerical example 2 (cont'd)

- For n=10, the degrees of freedom = ______.
- Consulting the t table, we see that
  - for 90% confidence, c = _______.
  - for 95% confidence, c = _______.
  - for 99% confidence, c = _______.

Numerical example 2 (cont'd)

- 90% CI = 175 ± 1.833 (7) = 175 ± 12.831 = (__________,__________) centimeters.
- 95% CI = 175 ± 2.262 (7) = 175 ± 15.834 = (__________,__________) centimeters.
- 99% CI = 175 ± 3.250 (7) = 175 ± 22.75 = (__________,__________) centimeters.



Other parameters

- It is also mathematically possible to construct confidence intervals for σ^2.
- These are usually of less interest so not discussed here.

Confidence intervals for non-normal population distributions

- In many settings, the population distribution is definitely not normal.
  - Family size (discrete, not continuous).
  - Opinion polls (yes-no).
  - Income (non-negative, skewed to the right)
- Yet we may still want to compute a confidence interval for the mean (or some other parameter). How to do this?

Confidence interval for the mean of an arbitrary distribution

- Suppose we have a random sample of n observations from an arbitrary distribution with unknown population mean μ and unknown population variance σ^2.
- Suppose we wish to construct a 95% confidence interval for the population mean.
- Let $\overline{X}$ denote the sample mean.

CONFIDENCE INTERVALS

Use the *asymptotic* distribution of the estimator

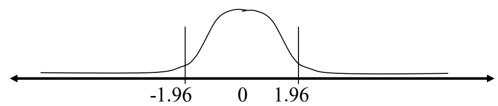
- From the central limit theorem we know that if n is large, $\frac{\overline{X}-\mu}{SE(\overline{X})} \overset{A}{\sim} N(0,1)$
- Here, $SE(\overline{X})$ denotes the asymptotic standard error. Often the following is used:

  asymptotic $SE(\overline{X}) = \sqrt{\tilde{\sigma}^2 / n}$

  where $\tilde{\sigma}^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

Using the standard normal table

- From a table of the standard normal distribution (or the bottom (∞) row of a *t* table) we see that if Z ~ N(0,1), Prob {|Z| < 1.96} = 0.95 .



Deriving the formula for the asymptotic CI

$$
\begin{aligned}
0.95 &= \text{Prob}\left\{\left|\frac{(\overline{X}-\mu)}{SE(\overline{X})}\right| < 1.96\right\} \\
&= \text{Prob}\left\{1.96 > \frac{(\overline{X}-\mu)}{SE(\overline{X})} > -1.96\right\} \\
&= \text{Prob}\left\{1.96\,SE(\overline{X}) > (\overline{X}-\mu) > -1.96\,SE(\overline{X})\right\} \\
&= \text{Prob}\left\{-\overline{X}+1.96\,SE(\overline{X}) > -\mu > -\overline{X}-1.96\,SE(\overline{X})\right\} \\
&= \text{Prob}\left\{\overline{X}-1.96\,SE(\overline{X}) < \mu < \overline{X}+1.96\,SE(\overline{X})\right\}
\end{aligned}
$$

Formula for the 95% asymptotic confidence interval

- So the 95% asymptotic CI for the mean is

$$
\begin{aligned}
&\left(\overline{X}-1.96\,SE(\overline{X}),\ \overline{X}+1.96\,SE(\overline{X})\right) \\
&= \overline{X} \pm 1.96\,SE(\overline{X})
\end{aligned}
$$

Other levels of confidence

- The same standard normal table shows that, if Z ~ N(0,1),
  Prob {|Z| < 1.645} = 0.90 and
  Prob {|Z| < 2.576} = 0.99 .
- So for a 90% CI, replace "1.96" with "________" in the formula.
- For an 99% CI, replace "1.96" with "________."

Numerical example 3

- Suppose we have polled 500 people for their opinion on some topic.
- We obtain 280 "yes" answers (x=1) and 220 "no" answers (x=0).
- Our estimate of the unknown fraction "yes" in the larger population is the fraction in our sample: $\hat{p} = \overline{X} = \frac{280}{500} = 0.56.$

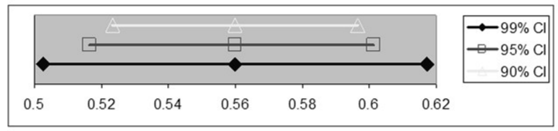
CONFIDENCE INTERVALS

Numerical example 3 (cont'd)

- How precise is this estimate?
- Variance of a Bernoulli random variable = p(1-p), so a consistent estimate of the unknown population variance =
  $\tilde{\sigma}^2 = 0.56\,(1-0.56) = 0.2464$
- Asymptotic standard error =
  $\sqrt{\tilde{\sigma}^2 / n} = \sqrt{0.2464 / 500} = 0.0222$

Numerical example 3 (cont'd)

- 90% CI = 0.56 ± 1.645 (0.0222) = 0.56 ± 0.0365 = (__________,___________).
- 95% CI = 0.56 ± 1.96 (0.0222) = 0.56 ± 0.0435 = (__________,___________).
- 99% CI = 0.56 ± 2.576 (0.0222) = 0.56 ± 0.0572 = (__________,___________).



The meaning of "confidence"

- The level of confidence is the probability that the formula encloses the true parameter, when the same formula is applied in repeated samples.
- In repeated samples, a 95% CI formula encloses the true parameter value __________ of the time.
- Of course, particular *values* of a CI computed from a particular sample are *numbers*, not random variables.

What is the source of randomness?

- Which is random—the true population parameter or the confidence interval?
- In classical statistics, the true population parameter is assumed __________, though unknown. It is NOT random.
- By contrast, the CI is a formula using the data in the sample, so its value varies ____________ from sample to sample.

Conclusions

- A γ *confidence interval* is a pair of estimators that probably bound the unknown true population parameter, with probability γ.
- Sampling from a normal distribution, the CI for the mean is $\bar{X} \pm c\,SE(\bar{X})$ where c is from the t table with (______) DOF.
- Sampling from an arbitrary distribution, the same formula is used, but with the asymptotic standard error, and c taken from _______________ table.

BASIC CONCEPTS OF HYPOTHESIS TESTS

BASIC CONCEPTS OF HYPOTHESIS TESTS

- What is a statistical hypothesis test?
- What do "power" and "size" mean in statistics?

Key values of parameters

- In many settings, a *key value* of an unknown parameter is economically important.
- Example from microeconomics:
  - Suppose we are estimating the price elasticity of demand for cigarettes. Call it θ.
  - The value $\theta=0$ is key because it implies that cigarette buyers do not respond to price.

Key values of parameters (cont'd)

- Another example from microeconomics:
  - Suppose we are estimating the returns to scale parameter (the sum of the exponents in a Cobb-Douglas production function) for an industry. Call it θ.
  - The value $\theta=1$ is key because it implies constant returns to scale. Large firms have no advantage over small firms so the industry has no tendency to consolidate into monopoly.

Key values of parameters (cont'd)

- Example from macroeconomics:
  - Suppose we are estimating the effect of inflation on unemployment (the reciprocal of the slope of the Phillips curve). Call it θ.
  - The value $\theta=0$ is key because it implies a "vertical Phillips curve": no tradeoff between inflation and unemployment.

Key values of parameters (cont'd)

- Another example from macroeconomics
  - Suppose we are estimating the slope of the consumption function, the marginal propensity to consume (MPC).
  - The value $\theta=1$ is key because if the MPC = 1 then the multiplier is not a meaningful concept.

Testing a key value

- If a key value is economically important, we may want to know whether or not the data agree with that key value.
- But estimates almost never equal the true value exactly, due to sampling error, so we cannot base our decision on whether our estimate equals the key value exactly.
- Rather, we must base our decision whether our estimate is "close" to the key value.

BASIC CONCEPTS OF HYPOTHESIS TESTS

Definition of hypothesis test

- *A decision rule, based on the data, that permits one to choose between two hypotheses about an unknown parameter θ.*
- The *null hypothesis* (H_0) supposes that the true unknown parameter θ equals some key value (say, zero or one).
- The *alternative hypothesis* (H_1) supposes that the true unknown parameter θ lies in some range of alternative plausible values.

Two-sided alternatives

- Sometimes the range of alternative plausible values is greater or less than the key value.
- Example: The returns-to-scale parameter in and industry might plausibly be greater or less than one.
  - If $\theta > 1$, there are increasing returns to scale.
  - If $\theta < 1$, there are decreasing returns to scale.

One-sided alternatives

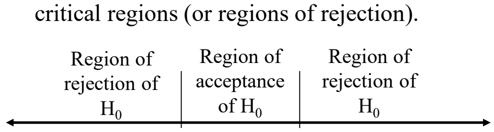
- Sometimes the range of alternative plausible values lies only on one side of the key value. Examples:
  - The elasticity of demand cannot be _________, even for cigarettes.
  - The MPC cannot be greater than __________.
  - The effect of inflation on unemployment cannot be ______________.

Components of a hypothesis test

- *Test statistic:* a formula to be computed from data. Related to, but not the same as, the parameter of interest.
- *Critical region:* a range of possible values of the test statistic which indicate the null hypothesis (H_0) should be rejected.
- Boundaries of the critical region are called *critical points*.

Rejecting the null hypothesis

- If the test statistic falls in the critical region, we *reject the null hypothesis* (H_0).
- Thus the critical region is sometimes called the *region of rejection*.
- If the alternative is two-sided, there may be two critical regions (or regions of rejection).



Accepting the null hypothesis?

- If the test statistic does not fall in the critical region, some people say the test statistic falls in the *region of acceptance* and that we *accept the null hypothesis.*
- However, this terminology is perhaps misleading.
  - Often the test statistic falls outside the critical region just because we have too few observations.
- Better terminology might be to say we *cannot reject the null hypothesis.*

BASIC CONCEPTS OF HYPOTHESIS TESTS

Errors in hypothesis tests

- Test statistics are computed from data in a random sample.
- Hence test statistics are random variables.
- Sometimes they accidentally land in the critical region, even if H_0 is true.
- Sometimes they accidentally fall outside the critical region, even if H_0 is false.

Possible outcomes

		Correct hypothesis in reality: H_0	Correct hypothesis in reality: H_1
Decision indicated by test	H_0	**No error**	**Type II error: fails to reject null when it is false.**
	H_1	**Type I error: Rejects null when it is true.**	**No error**

Probabilities of errors

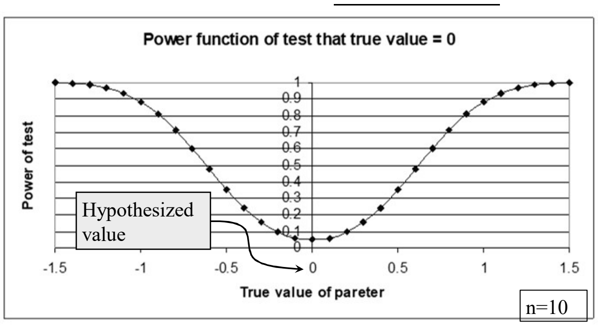
- *Size (or significance) of a test* = probability of a Type I error, of rejecting H_0 when it is really true.
- *Power of a test* = probability of rejecting H_0 when it is false.
- A good test has low ___________________ and high _____________.

Power function

- Usually the power of a test depends on the particular value taken by the parameter, among possible alternative values.
- *Power function* = probability of rejecting H_0, as a function of the true value of the parameter.
- By definition, at the hypothesized value of the parameter, power function = size of test.

Example of power function for a two-sided test

The size of this test is ____________.



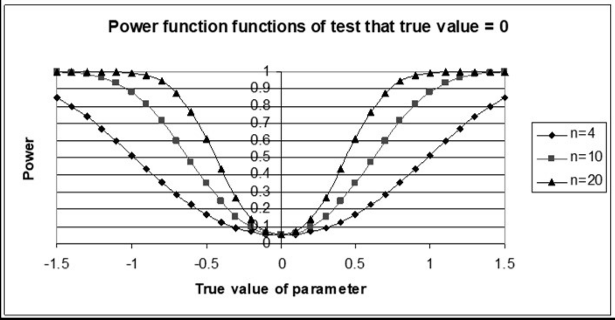
Shape of power function

- The power function usually rises as true value differs from hypothesized value.
- Examples on previous slide:
  - If the true value of the parameter is 1.2, the test rejects the null hypothesis over 95% of the time.
  - But if the true value of the parameter is -0.2, the test rejects the null hypothesis only 10% of the time.

BASIC CONCEPTS OF HYPOTHESIS TESTS

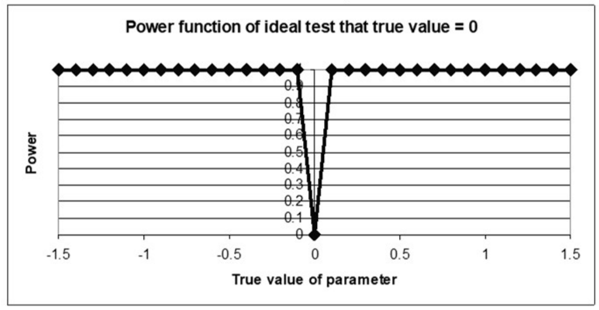
Shape of power function (cont'd)

We can raise the power function (except at the hypothesized value) if we increase the number of observations.



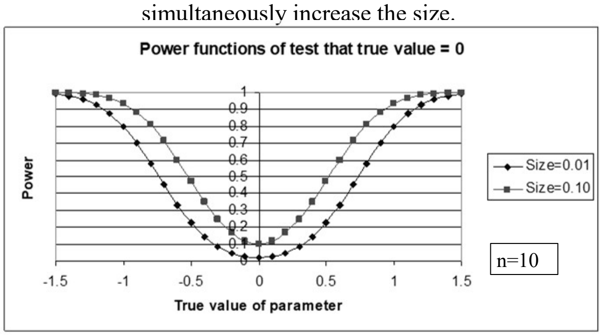
Power function of the ideal test

If we had an infinite number of observations, we could construct the ideal test: never rejects H_0 when it is true, always rejects H_0 when it is false.



A tradeoff between size and power

With a *fixed* number of observations, we can only increase the power of the test if we also simultaneously increase the size.



Most powerful tests

- A good test should have low size and high power function away from the hypothesized value.
- But for a given number of observations, there is a tradeoff between reducing size and increasing power.
- A test that maximizes power for a specified size (say, 5%) is called the *most powerful test* of that size.
- Tests presented in textbooks have usually been proven mathematically to be most powerful tests.

Conclusions

- A *hypothesis test* is a decision rule between a ___________ hypothesis (H_0) that a parameter equals a particular key value and an ___________ hypothesis (H_1) that it equals some other value.
- The probability that the test mistakenly rejects H_0 is called the __________ of the test.
- The probability that the test correctly rejects H_0 is called the __________ of the test.
- A good test has low ________ and high _______.

TESTING THE MEAN OF A DISTRIBUTION

TESTING THE MEAN OF A DISTRIBUTION

- How can we test a hypothesis about the mean of a normal distribution?
- How can we test a hypothesis about the mean of an arbitrary distribution?

Testing the mean of a normal distribution

- Suppose we have a random sample of 20 observations from a population that we are (for some reason) sure is normally-distributed.
- However, we are unsure about the true population mean μ.
- We wish to test the hypothesis that μ equals some key value—say, 5.

The hypotheses

- Our null hypothesis is thus:
  H_0: $\mu = 5$.
- Suppose that the range of alternative plausible values of μ includes values both greater than or less than 5.
- The alternative hypothesis is thus:
  H_1: $\mu \neq 5$.

Distribution of the test statistic

- The assumption that our sample is taken from a normally-distributed population implies the following useful fact.
- It can be shown (see a mathematical statistics book) that the formula below, a random variable, follows a *t* distribution with 19 degrees of freedom:

$$\frac{(\bar{X}-\mu)}{SE(\bar{X})} \quad \text{where} \quad SE(\bar{X}) = \sqrt{\hat{\sigma}^2/20}, \quad \hat{\sigma}^2 = \left(\tfrac{1}{19}\right)\sum_{i=1}^{20}\left(X_i - \bar{X}\right)^2$$

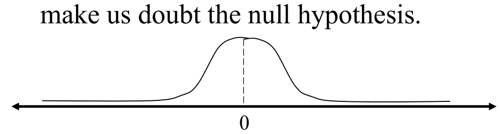
Distribution of the test statistic

- A test statistic must be computable from data. Since we do not know μ, we cannot use the formula on the previous slide.
- But now replace μ by its hypothesized value, 5.
- If the null hypothesis is true, the following *test statistic*, computable from data, follows a *t* distribution with 19 degrees of freedom:

$$W = \frac{(\bar{X}-5)}{SE(\bar{X})}$$

Behavior of the test statistic under the null hypothesis H_0: $\mu = 5$

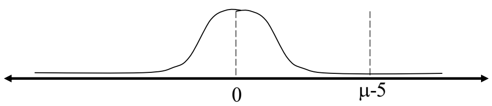
- If the true population mean equals 5, then W will be scattered in a bell-shaped curve centered at zero.
- We should be surprised to find W very far from zero. Finding W far from zero should make us doubt the null hypothesis.



TESTING THE MEAN OF A DISTRIBUTION

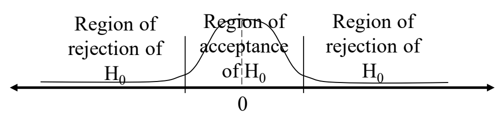
Behavior of the test statistic under the alternative hypothesis H_1: $\mu \neq 5$

- If the true population mean does *not* equal 5, then W will be scattered in a bell-shaped curve centered at the true population mean minus 5.
- If the true population mean is much larger or much smaller than 5, we should expect W to be much larger or much smaller than zero.



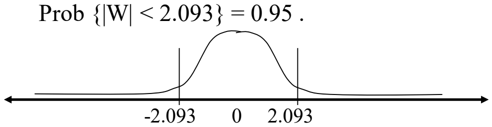
Critical region (or region of rejection)

- So the region of acceptance should be ________ to zero, and the region of rejection should be __________ from zero on _____________ sides.
- How can we find the borders of the critical region—the critical _____________?



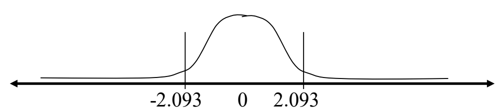
Using the *t* table

- Open your textbook to the table of the *t* distribution in the back. Focus on the 2-tailed probabilities.
- Depending on the format of the table, it shows that, if W ~ t(19), then either
  Prob {|W| > 2.093} = 0.05 or
  Prob {|W| < 2.093} = 0.95 .



Critical points

- It is common to choose a size of 0.05 (5%).
- Recall that "size" means the probability of a Type I Error, of rejecting the H_0 when it is really true.
- According to the table, if we reject H_0 only when |W|>2.093, then the size of the test is 0.05.
- So the critical points are ±2.093 .

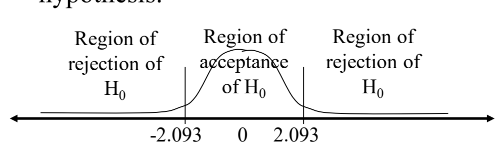


Numerical example

- Suppose for our sample of 20 observations, we find that $\bar{X} = 4.1$ and $SE(\bar{X}) = 0.5$
- Then W = (4.1-5)/0.5 = -1.8.

Numerical example (cont'd)

- Since -1.8 falls in the region of acceptance for a test of size 0.05 (5%), we _________ the null hypothesis.
- Or more properly, we *fail to reject* the null hypothesis.



TESTING THE MEAN OF A DISTRIBUTION

Other sizes (or "levels of significance")

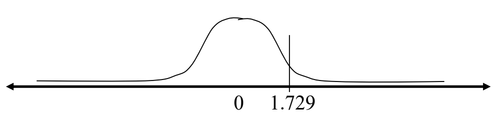
- The same t table shows that, if W ~ t(19), Prob {|W| > 1.729} = 0.10 and Prob {|W| > 2.861} = 0.01 .
- So for a test of size 0.10 (10%), the critical points are ±__________ instead of ±2.093.
- For a test of size 0.01 (1%), the critical points are ±__________ instead of ±2.093.

Test results at other sizes (or "levels of significance)

- Our test statistic W = -1.8 falls in the region of _____________ for a test of size 0.10.
- It falls in the region of ______________ for a test of size 0.01.
- The larger the size, the smaller the region of acceptance, and the __________ likely the test will reject the null hypothesis.

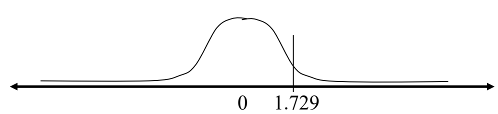
One-sided alternatives

- Suppose the range of alternative plausible values of μ includes only *positive* values.
- Focus on one-tailed probabilities.
- The same t table shows that Prob {W > 1.729} = 0.05 .



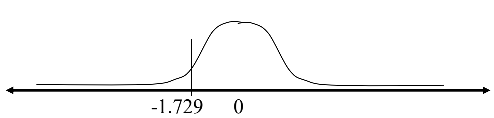
One-sided alternatives (cont'd)

- Thus we should use the one-tailed test.
- For size = 0.05, there is a single critical point 1.729.



One-sided alternatives (cont'd)

- Suppose only the range of alternative plausible values of μ includes only *negative* values.
- Then we should again use the one-tailed test, with the single critical point -1.729.



General form of t-test

- Suppose we have n observations from a normal population with unknown mean and variance.
- We wish to test the null hypothesis that the true population mean equals b, that is, H_0: $\mu = b$.
- We use the test statistic

$$\frac{\left(\bar{X}-b\right)}{SE\left(\bar{X}\right)} \quad \text{where} \quad SE\left(\bar{X}\right)=\sqrt{\hat{\sigma}^2/n}, \quad \hat{\sigma}^2=\left(\tfrac{1}{n-1}\right)\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$$

TESTING THE MEAN
OF A DISTRIBUTION

General form of t-test (cont'd)

- Under the null hypothesis H_0: $\mu = b$, this test statistic is distributed as t with (n-1) degrees of freedom.
- For a two-tailed test, there are two ± critical points, found on the row of the table with (n-1) degrees of freedom, and the column with two tails at the desired size significance level.

General form of t-test (cont'd)

- For a one-tailed test, there is just one critical point, found on the row with (n-1) degrees of freedom, and the column with one tail at the desired size significance level.
- Use a positive critical point if the alternative hypothesis is H_1: $\mu > b$.
- Use a negative critical point if the alternative hypothesis is H_1: $\mu < b$.

Testing the mean of non-normal population distributions

- In many settings, the population distribution is definitely not normal.
  - Family size (discrete, not continuous).
  - Opinion polls (yes-no).
  - Income (non-negative, skewed to the right)
- Yet we may still want to test a hypothesis about mean.

Asymptotic t-test

- Suppose we have n observations from an arbitrary distribution with unknown mean and variance. Assume n is a large number.
- We wish to test the null hypothesis that the true population mean equals b, that is, H_0: $\mu = b$.
- We use the test statistic

$$\frac{\left(\bar{X} - b\right)}{SE\left(\bar{X}\right)} \quad \text{where} \quad SE\left(\bar{X}\right) = \sqrt{\tilde{\sigma}^2 / n}$$

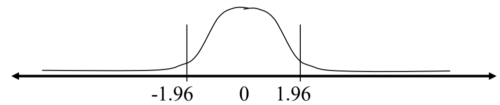
Asymptotic t-test (cont'd)

- Note that $SE\left(\bar{X}\right)$ is the asymptotic standard error.
- Here, $\tilde{\sigma}^2$ is any consistent estimator of the unknown true population variance σ^2. Often the following is used.

$$\tilde{\sigma}^2 = \left(\tfrac{1}{n}\right)\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Using the standard normal table

- From the central limit theorem we know that if n is large, $\dfrac{\bar{X} - \mu}{SE\left(\bar{X}\right)} \overset{A}{\sim} N(0,1)$
- From a table of the standard normal distribution, or the bottom (∞) row of a t table, we see that if $Z \sim N(0,1)$, Prob $\{|Z| < 1.96\} = 0.95$.



TESTING THE MEAN OF A DISTRIBUTION

Critical points for asymptotic t-test of size 0.05

- Thus for a two-tailed test, we use critical points ±1.96.
- For a one-tailed test, we use a single critical point, depending on the alternative hypothesis.
  - Use +1.645 if H_1: $\mu > b$.
  - Use -1.645 if H_1: $\mu < b$.

Critical points for asymptotic t-test of other sizes (or significance levels)

- Critical points for other sizes can be found from the same table.
- Size = 0.10
  - Two-tailed critical points: ±1.645 .
  - One-tailed critical point: 1.282 or -1.282.
- Size = 0.01
  - Two-tailed critical points: ±2.576 .
  - One-tailed critical point: 2.326 or -2.326.

Conclusions

- To test whether the true population mean equals a particular value *b*, compute the following test statistic: $\dfrac{(\bar{X}-b)}{SE(\bar{X})}$
- If the population is assumed normal, choose critical point(s) from the __________ table.
- If the sample is large, regardless of the population distribution, choose critical point(s) from the ________________________ table.

P-VALUES

P-VALUES

- How are P-values computed?
- What do they tell us?

General idea behind testing hypotheses

- Because data are random, any value of the test statistic is *possible*, whether the null hypothesis is true or false.
- But we reject the null hypothesis only if the value of our test statistic would be *very* ______________ if the null hypothesis were true.
- How unusual?

Size (or significance level) of test

- Suppose you choose a size of 0.05.
- That means you have decided to reject the null hypothesis if the test statistic is so large, it would take this value less than ____ percent of the time if the null hypothesis were ____________.

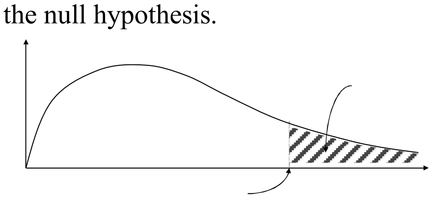
Two ways to decide whether to reject the null hypothesis

(1) Find the critical point (the boundary of the critical region) given your chosen size. Then compare the test statistic with the critical point.

(2) Directly compute *how unusual* the value of the test statistic is, given the null hypothesis. That is, compute the *p-value*. Then compare it with your chosen size.

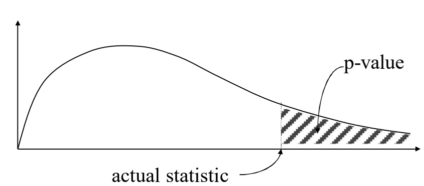
Definition

- The *p-value* of a test statistic = probability of obtaining a test statistic equal to or greater than the one actually obtained, under the null hypothesis.



What the p-value indicates

- The *p-value* of a test statistic tells us *how unusual* the value of the test statistic is, assuming the null hypothesis is true.



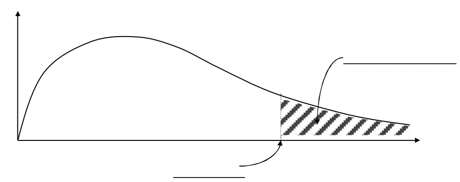
P-VALUES

Example: chi-square test

- Suppose a test statistic is distributed as chi-square with DOF=5 under the null hypothesis and the value of the statistic actually obtained turns out to be 8.7.
- The p-value (computed using the *chidist* function in Excel) is p = 0.1216.

Example: chi-square test (cont'd)

- Thus, the probability of obtaining a test statistic equal to or greater than 8.7, under the null hypothesis, is 0.1216.

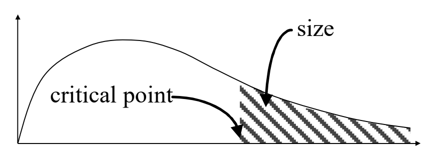


Example: chi-square test (cont'd)

- What does this p-value indicate?
- Even if the null hypothesis were true, a value of the test statistic greater than or equal to 8.7 would still occur _________ % of the time—not so unusual!
- If you had chosen a size of 5% or even 10%, you could __________ reject the null hypothesis.

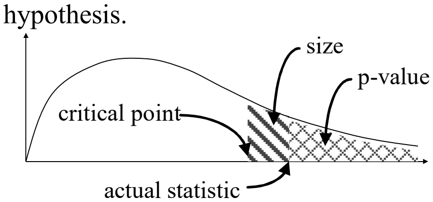
Size versus p-value

- Recall that the size (or significance level) of the test is the area to the *right* of the critical point.



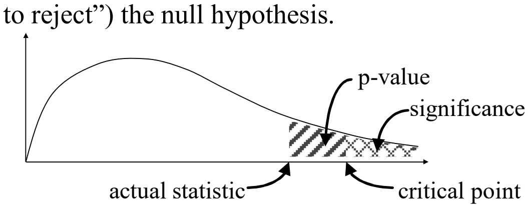
What if p-value < size?

- The p-value < size if and only if actual statistic > critical point.
- So if p-value < size, ___________ the null hypothesis.



What if p-value > size?

- The p-value > size if and only if actual statistic < critical point.
- So if p-value > size, ___________ (or "fail to reject") the null hypothesis.



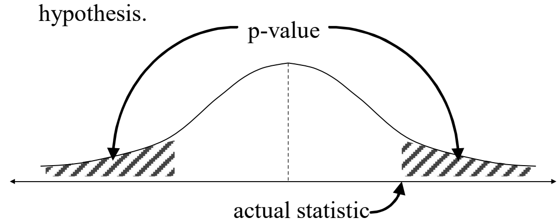
P-VALUES

Example: chi-square test (cont'd)

- Since the p-value = 0.1216, then the actual value of the test statistic must be
  - to the left of the critical point at 10% significance. So we __________ reject the null hypothesis at 10%.
  - to the right of the critical point at 20% significance. So we __________ reject the null hypothesis at 20%.

P-value for two-tailed tests

- *p-value* of a test statistic = probability of obtaining a test statistic as high or higher *in absolute value* than the one actually obtained, under the null hypothesis.

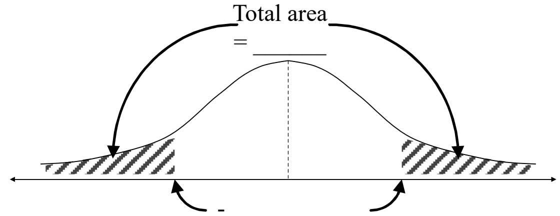


Example: normal test

- Suppose a test statistic is distributed as standard normal under the null hypothesis, and the value of the statistic turns out to be 2.5.
- The two-sided p-value, computed using *(1-NORMSDIST(2.5))*2* in Excel, is 0.0124.

Example: normal test (cont'd)

- Thus, the probability of obtaining a test statistic greater than 2.5 or less than -2.5, under the null hypothesis, is 0.0124.



Example: normal test (cont'd)

- What does this p-value indicate?
- If the null hypothesis were true, a value of the test statistic with an absolute value of 2.5 or more would only occur _________ % of the time—fairly unusual!
- If you had chosen a size of 5% or even 2%, you would have to _____________ the null hypothesis.

Example: normal test (cont'd)

- Given that the p-value = 0.0124, we should
  - ________________ the null hypothesis at 5% significance.
  - ________________ the null hypothesis at 2% significance.
  - ________________ the null hypothesis at 1% significance.

P-VALUES

How and why report p-values?

- Calculating p-values is easy when a spreadsheet or statistical software is used to compute the test statistic.
  - Just use built-in function for the cumulative probability distribution function.
- Reporting p-values saves the reader from having to refer to a table of critical points.

Conclusions

- The *p-value* of a test statistic = probability of obtaining a test statistic ____________ than or equal to the one actually obtained, under the null hypothesis.
- If p-value < size, ____________ the null hypothesis.
- If p-value > size, ____________ (or "fail to reject") the null hypothesis.

PART 2

Two-Variable Regression

ALGEBRAIC PROPERTIES OF LEAST-SQUARES

ALGEBRAIC PROPERTIES OF LEAST-SQUARES

- What properties of LS estimates must hold regardless of data assumptions?

The least-squares principle

- Choose the line that minimizes the sum of the squared vertical deviations.
- Find values of β_1 and β_2 that minimize the following objective function:

$$f(\beta_1, \beta_2) = \sum_{i=1}^{n} \left(y_i - [\beta_1 + \beta_2 x_i] \right)^2$$

First-order necessary conditions (FONCs) for LS estimates of β_1 and β_2

(1) Set zero equal to derivative of $f(\beta_1,\beta_2)$ with respect to β_1:

$$0 = \sum_{i=1}^{n} -2 \left(y_i - [\beta_1 + \beta_2 x_i] \right)$$

(2) Set zero equal to derivative of $f(\beta_1,\beta_2)$ with respect to β_2:

$$0 = \sum_{i=1}^{n} -2 \left(y_i - [\beta_1 + \beta_2 X_i] \right) x_i$$

The least-squares estimators

- The FONCs can be solved to give the least-squares estimators:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

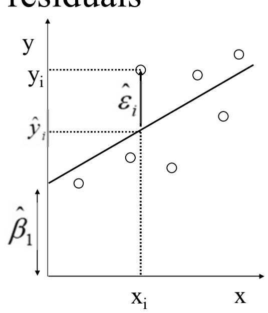
Definition of least-squares fitted values and residuals

- LS "fitted value" or "predicted value" =

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

- LS "residual" =

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$



Rewriting the first-order necessary conditions

- We can use the fitted values to simplify the FONCs defining the LS estimators for β_1 and β_2 :

(1) $$0 = \sum_{i=1}^{n} \left(y_i - [\hat{\beta}_1 + \hat{\beta}_2 x_i] \right) = \sum_{i=1}^{n} y_i - \hat{y}_i$$

(2) $$0 = \sum_{i=1}^{n} \left(y_i - [\hat{\beta}_1 + \hat{\beta}_2 x_i] \right) x_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_i$$

ALGEBRAIC PROPERTIES OF LEAST-SQUARES

Algebraic properties

- Using the FONCs and the definitions of fitted values and residuals, we can derive algebraic properties of LS.
- These properties hold *automatically*
  - no matter what data are used.
  - no matter whether our model is right or wrong.

Why algebraic properties are useful

- The fact that these algebraic properties hold tells us *nothing* about whether the LS estimates are accurate or useful.
- It just tells us our computer is not broken.
- But these algebraic properties can help us
  - check our calculations. (If the properties do not hold, we made an arithmetic mistake!)
  - make further calculations. (Such as the r^2 value—see below.)

Algebraic property 1

- The sum of the LS fitted values must equal the sum of the actual values.

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$$

- Proof: Follows from FONC (1).

Algebraic property 2

- The sum of the LS residuals must equal zero.

$$0 = \sum_{i=1}^{n} \hat{\varepsilon}_i$$

- Proof: Follows from FONC (1).

Algebraic property 3

- The sum of the products of the LS residuals and the X's must equal zero.

$$0 = \sum_{i=1}^{n} \hat{\varepsilon}_i x_i$$

- Proof: Follows from FONC (2).

Algebraic property 4

- The sum of the products of the LS residuals and the LS fitted values must equal zero.

$$0 = \sum_{i=1}^{n} \hat{\varepsilon}_i \hat{y}_i$$

- Proof: Follows from both FONCs (see next slide).

ALGEBRAIC PROPERTIES OF LEAST-SQUARES

Proof of property 4

$$\sum_{i=1}^{n} \hat{\varepsilon}_i \hat{y}_i = \sum_{i=1}^{n} \hat{\varepsilon}_i \left(\hat{\beta}_1 + \hat{\beta}_2 x_i\right)$$
$$= \sum_{i=1}^{n} \hat{\varepsilon}_i \hat{\beta}_1 + \sum_{i=1}^{n} \hat{\varepsilon}_i \hat{\beta}_2 x_i$$
$$= \hat{\beta}_1 \sum_{i=1}^{n} \hat{\varepsilon}_i + \hat{\beta}_2 \sum_{i=1}^{n} \hat{\varepsilon}_i x_i = 0$$

Algebraic property 5

- The sum of the squared deviations of Y around its mean must equal the sum of the squared LS residuals PLUS the sum of the squared deviations of the fitted values around the mean of Y.

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + \sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2$$

Proof of property 5

$$\sum\left(y_i - \overline{y}\right)^2 = \sum\left(\hat{\varepsilon}_i + \hat{y}_i - \overline{y}\right)^2$$
$$= \sum\left(\hat{\varepsilon}_i + \left[\hat{y}_i - \overline{y}\right]\right)^2$$
$$= \sum\left(\hat{\varepsilon}_i^2 + \left[\hat{y}_i - \overline{y}\right]^2 + 2\hat{\varepsilon}_i\left[\hat{y}_i - \overline{y}\right]\right)$$
$$= \sum\hat{\varepsilon}_i^2 + \sum\left[\hat{y}_i - \overline{y}\right]^2 + 2\sum\hat{\varepsilon}_i\left[\hat{y}_i - \overline{y}\right]$$

Proof of property 5 (cont'd)

- But the third term is zero because

$$\sum\hat{\varepsilon}_i\left[\hat{y}_i - \overline{y}\right] = \sum\hat{\varepsilon}_i\hat{y}_i - \sum\hat{\varepsilon}_i\overline{y}$$
$$= \left(\sum\hat{\varepsilon}_i\hat{y}_i\right) - \overline{y}\left(\sum\hat{\varepsilon}_i\right)$$

- So we are left with

$$\sum\left(y_i - \overline{y}\right)^2 = \sum\hat{\varepsilon}_i^2 + \sum\left[\hat{y}_i - \overline{y}\right]^2$$

Property 5 restated as "sums-of-squares decomposition"

"Residual sum of squares" *aka* "error sum of squares" $\sum\hat{\varepsilon}_i^2$

\+ "explained sum of squares" *aka* "regression sum of squares" $+\sum\left[\hat{y}_i - \overline{y}\right]^2$

= "total sum of squares."

Measuring goodness-of-fit

- A natural measure is the *fraction of the total sum of squares that is explained by the xs.*
- This is the R^2 value:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

ALGEBRAIC PROPERTIES OF LEAST-SQUARES

Another definition of R^2

- Using property 5, we can find an alternative definition of R^2 as

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Interpreting R^2

- Note that R^2 must lie between __________ and ___________ (by property 5).
- R^2 equals one if and only if the residuals are all zero—that is, the fit is ____________.
- It can be shown that R^2 is the square of the sample correlation between x and y (or between $\hat{y}$ and y).

Conclusions

- The sum of the LS fitted values must equal the sum of the ___________ values of y.
- The LS residuals, the products of the LS residuals with the x's, and the products of the residuals with the fitted values, must each sum to ___________.
- The total sum of squares equals the residual sum of squares plus the _____________ sum of squares. This motivates the R^2 measure.

FUNDAMENTAL ASSUMPTIONS

FUNDAMENTAL ASSUMPTIONS

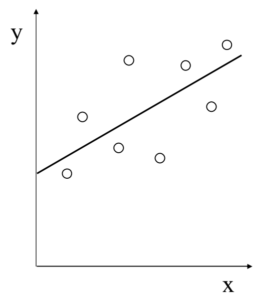
- What basic statistical assumptions do we need to justify least-squares?

Assumptions dictate method

- There are many methods of fitting a line to a set of data points. Examples:
  - Least-squares
  - Least absolute deviation
  - Reverse least-squares
- To decide which method or principle is best, we consider why the data are scattered around the "true" line.

Data scattered around "true" line

- We assume the "true" relationship between x and y is given by: $y = \beta_1 + \beta_2 x$.
- Here β_1, β_2 are unknown.
- But the observations are scattered around that true relationship.

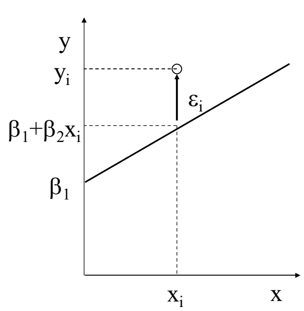


Why are the data scattered?

- Perhaps y is not accurately measured. It might be an estimate or an approximation.
  - Examples:
- Perhaps other variables influence y besides x.
  - Examples:

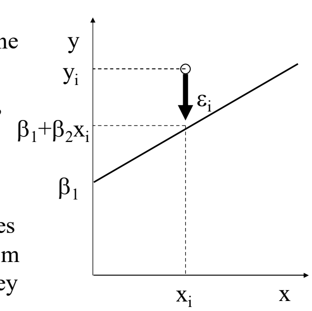
The error term

- In any case, we assume the observations are displaced from the true line by a random "error term": $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$,
- Here ε_i is a random "error term."



The error term is unobserved

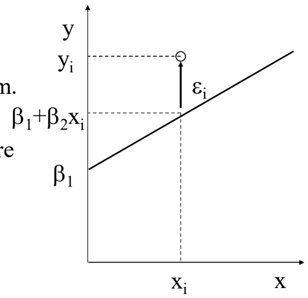
- Unlike x_i and y_i, the ε_i are *not* observed.
- If ε_i *were* observed, we could easily find the "true" line by subtracting: $y_i - \varepsilon_i$!
- The ε_i are sometimes called "latent" random variables because they are unobserved.



FUNDAMENTAL ASSUMPTIONS

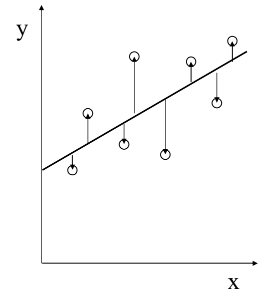
Implications of random error term

- Since $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, y_i is thus also random.
- So the n observations (x_i, y_i) are a random sample.



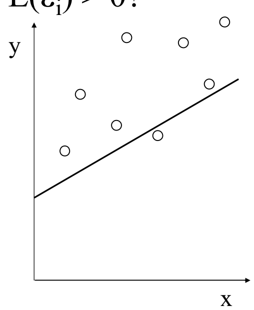
Assumption #1: $E(\varepsilon_i) = 0$

- In other words: $E(y_i|x_i) = \beta_1 + \beta_2 x_i$
- Observations tend to be scattered "evenly" above and below the line.



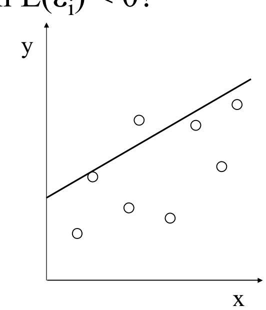
What if assumption #1 were violated, with $E(\varepsilon_i) > 0$?

- Then observations tend to be above the line.
- LS intercept estimate would be biased ___________.



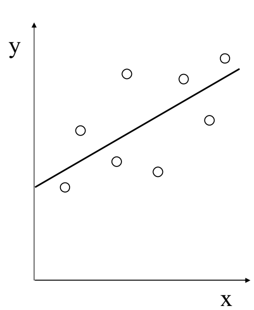
What if assumption #1 were violated, with $E(\varepsilon_i) < 0$?

- Then observations tend to be below the line.
- LS intercept estimate would be biased ___________.
- But usually the slope is of greater interest.



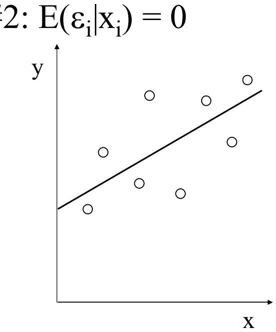
Assumption #2: $E(\varepsilon_i|x_i) = 0$

- This implies $E(\varepsilon_i x_i) = 0$.
- If x_i is random, then also $Cov(x_i,\varepsilon_i) = Corr(x_i,\varepsilon_i) =$ _____.



Meaning of Assumption #2: $E(\varepsilon_i|x_i) = 0$

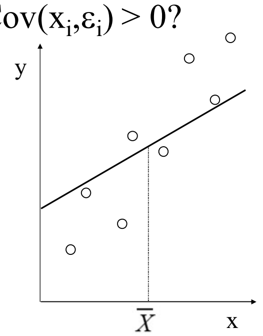
- The value of the error term ε_i is not affected by the value of x_i.
- All other factors that affect y_i are uncorrelated with x_i .



FUNDAMENTAL ASSUMPTIONS

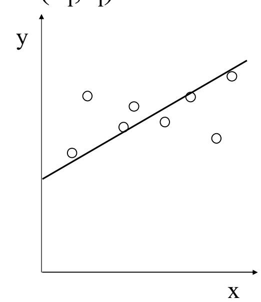
What if assumption #2 were violated, with $Cov(x_i,\varepsilon_i) > 0$?

- x_i and ε_i are positively correlated.
- Observations tend to be above line for large x_i, but below line for small x_i.
- LS slope estimate is biased ________.



What if assumption #2 were violated, with $Cov(x_i,\varepsilon_i) < 0$?

- x_i and ε_i are negatively correlated.
- Observations tend to be below line for large x_i, but above line for small x_i.
- LS slope estimate is biased __________.



The "method-of-moments" principle

- Set the moments of the sample equal to the formulas for the theoretical (or population) moments.
- Solve for estimators of the parameters of interest.
- Examples:

"Method of Moments" estimation: using assumption #1

- By assumption #1, $E(\varepsilon_i) = 0$, so set

$$0 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - [\beta_1 + \beta_2 x_i]\right)$$

"Method of Moments" estimation: using assumption #2

- By assumption #2, $E(\varepsilon_i x_i) = 0$, so set

$$0 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i x_i$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - [\beta_1 + \beta_2 x_i]\right) x_i$$

"Method of Moments" estimators for β_1 and β_2

- Thus together, assumptions #1 and #2 and the "method-of-moments" principle imply the following equations:

$$0 = \sum_{i=1}^{n} \left(y_i - [\beta_1 + \beta_2 x_i]\right)$$
$$0 = \sum_{i=1}^{n} \left(y_i - [\beta_1 + \beta_2 x_i]\right) x_i$$

FUNDAMENTAL ASSUMPTIONS

Least-squares again!

- Exactly same "normal equations" we derived from the least-squares principle!
- Conclude: Under assumptions #1 and #2, least-squares estimators satisfy the "method-of-moments principle."

Conclusions

- Fundamental assumptions are:
  - Assumption #1: $E(\varepsilon_i) = 0$.
  - Assumption #2: $E(\varepsilon_i|x_i) = 0$.
- They imply that the observations are scattered evenly above and below the true regression line for all values of x.
- They also imply that LS estimators satisfy the ________________________ principle.

PROPERTIES UNDER FUNDAMENTAL ASSUMPTIONS

PROPERTIES UNDER FUNDAMENTAL ASSUMPTIONS

•How can we justify LS using only these basic statistical assumptions?

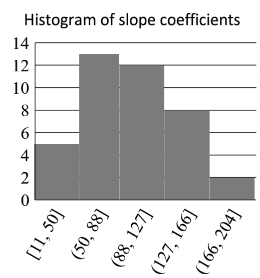
Estimator versus estimate

- Estimator = formula. Takes different values for different samples.
  - *A random variable.*
- Estimate = particular value taken for a particular sample.
  - *An ordinary number.*

Least-squares estimates vary from sample to sample

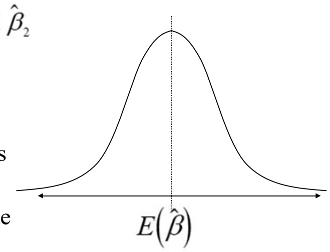
Example: 40 samples of 50 observations each, from Current Population Survey.

- $y = \beta_1 + \beta_2 x$, where y = weekly earnings, x = years of schooling



Least-squares estimators are themselves random variables

- LS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are functions of the random ys (or εs) in the particular sample.
- $\hat{\beta}_1$ and $\hat{\beta}_2$ have means and variances.
- What can we conclude about them?

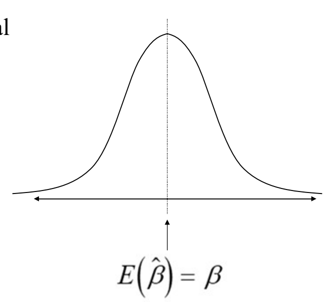


Least-squares estimators are unbiased

- Under the fundamental assumptions #1 and #2, the LS estimators are unbiased.

$$E\left(\hat{\beta}_1\right) = \beta_1$$

$$E\left(\hat{\beta}_2\right) = \beta_2$$



Proof that least-squares slope estimator is unbiased

- Recall slope formula: $\hat{\beta}_2 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- Use following fact to simplify this slope formula:

$$\begin{aligned} y_i - \bar{y} &= (\beta_1 + \beta_2 x_i + \varepsilon_i) - \tfrac{1}{n}\sum(\beta_1 + \beta_2 x_i + \varepsilon_i) \\ &= (\beta_1 + \beta_2 x_i + \varepsilon_i) - (\beta_1 + \beta_2 \bar{x} + \bar{\varepsilon}) \\ &= \beta_2(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \end{aligned}$$

PROPERTIES UNDER FUNDAMENTAL ASSUMPTIONS

Proof that least-squares slope estimator is unbiased (cont'd)

- Substituting:

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})\left(\beta_2[x_i - \bar{x}] + [\varepsilon_i - \bar{\varepsilon}]\right)}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum \beta_2 (x_i - \bar{x})^2 + \sum (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum (x_i - \bar{x})^2}$$

Proof that least-squares slope estimator is unbiased (cont'd)

- Substituting:

$$\hat{\beta}_2 = \beta_2 \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} + \frac{\sum (x_i - \bar{x})\varepsilon_i - \bar{\varepsilon}\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

Proof that least-squares slope estimator is unbiased (cont'd)

- Simplifying:

$$\hat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2}$$

- The mean of this expression is:

$$E\left(\hat{\beta}_2\right) = \beta_2 + \frac{\sum (x_i - \bar{x})E(\varepsilon_i)}{\sum (x_i - \bar{x})^2} =$$

Proof that least-squares intercept estimator is unbiased

- Recall intercept formula: $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$

$$E\left(\hat{\beta}_1\right) = E(\bar{y}) - E\left(\hat{\beta}_2\right)\bar{x}$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n}(\beta_1 + \beta_2 x_i + \varepsilon_i)\right) - \beta_2 \bar{x}$$

$$= \beta_1 + \beta_2 \bar{x} + \frac{1}{n}\sum_{i=1}^{n} E(\varepsilon_i) - \beta_2 \bar{x} \quad = \beta_1$$

Variance of least-squares slope estimator

$$Var\left(\hat{\beta}_2\right) = E\left(\beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2} - E\left(\hat{\beta}_2\right)\right)^2$$

$$= E\left(\frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2}\right)^2 = \frac{E\left(\sum (x_i - \bar{x})\varepsilon_i\right)^2}{\left(\sum (x_i - \bar{x})^2\right)^2}$$

Variance of least-squares slope estimator (cont'd)

$$Var\left(\hat{\beta}_2\right) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 E\left(\varepsilon_i^2\right)}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2} + \frac{\sum_{i=1}^{n}\sum_{j \neq i} (x_i - \bar{x})(x_j - \bar{x}) E\left(\varepsilon_i \varepsilon_j\right)}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2}$$

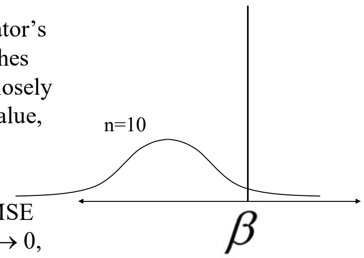
PROPERTIES UNDER FUNDAMENTAL ASSUMPTIONS

Variance of LS estimators under Assumptions #1 and #2

- Formula for variance of $\hat{\beta}_2$ is still fairly complicated .
- Cannot be simplified without further assumptions.
- Similar result can be shown for $\hat{\beta}_1$.

Consistency: definition

- *Consistent* estimator's distribution bunches more and more closely around the true value, as $n \to \infty$.
- An estimator is *consistent* if its MSE $=\text{bias}^2+\text{variance} \to 0$, as $n \to \infty$.



Least-squares estimators are consistent

- Bias of LS estimators = zero, so we need only show that variance approaches zero.
- Sufficient conditions: σ_i^2 and $\text{Cov}(\varepsilon_i,\varepsilon_j)$ are bounded, and the variation of x around its mean does not diminish.
- Notation for *consistency*:

$$\hat{\beta}_1 \xrightarrow{P} \beta_1 \qquad \hat{\beta}_2 \xrightarrow{P} \beta_2$$

Functions of least-squares estimators are also consistent

- An important theorem shows that continuous functions of consistent estimators are themselves always consistent.
- Application: LS can be used to estimate consistently the x-intercept ($-\beta_1/\beta_2$).

$$-\frac{\hat{\beta}_1}{\hat{\beta}_2} \xrightarrow{P} -\frac{\beta_1}{\beta_2}$$

Conclusions

Assuming $E(\varepsilon_i)=0$ and $E(\varepsilon_i|x_i)=0$, LS estimators are

- ____________________ (meaning their expected values equal the true coefficients).
- ____________________ under modest assumptions (meaning their distributions bunch more closely around the true coefficients as the sample size increases).

However, formulas for variance of LS estimators are fairly complicated without further assumptions.

ADDITIONAL USEFUL ASSUMPTIONS

ADDITIONAL USEFUL ASSUMPTIONS

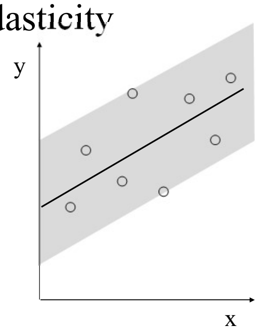
•What additional assumptions do we need to gauge the precision of our LS estimates?

Less fundamental assumptions

- The following assumptions are not as critical as $E(\varepsilon_i) = 0$ and $Cov(x_i,\varepsilon_i) = 0$.
- They may not hold in some datasets.
- But if they do hold, they help drastically simplify the formula for variance of LS estimators.

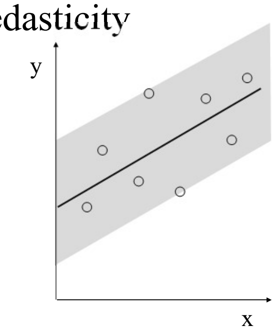
Assumption #3: homoskedasticity

- $Var(\varepsilon_i) = E(\varepsilon_i)^2 = \sigma^2$.
- In other words: $Var(y_i|x_i) = \sigma^2$.
- Variance is _______ for all observations, regardless of x .
- Note: opposite of homoskedasticity is "______skedasticity."



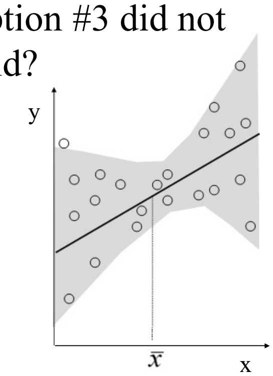
Assumption #3: homoskedasticity

- $Var(\varepsilon_i) = E(\varepsilon_i)^2 = \sigma^2$.
- In other words: $Var(y_i|x_i) = \sigma^2$.
- Variance is __same__ for all observations, regardless of x .
- Note: opposite of homoskedasticity is "__hetero__skedasticity."



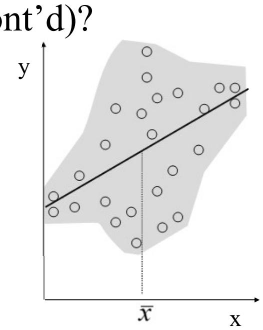
What if assumption #3 did not hold?

- Heteroskedasticity.
- In this graph, $Var(\varepsilon_i)$ _______________ as x gets farther from its mean.



What if assumption #3 did not hold (cont'd)?

- Heteroskedasticity.
- In this graph, $Var(\varepsilon_i)$ _______________ as x gets farther from its mean.



ADDITIONAL USEFUL ASSUMPTIONS

Assumption #4: no autocorrelation

- Cov (ε_i, ε_j) = E (ε_i ε_j) = 0, for $i \neq j$.
- Thus error terms for different observations are uncorrelated.
- Always satisfied if data come from a random sample.
- Usually satisfied for cross-section datasets.

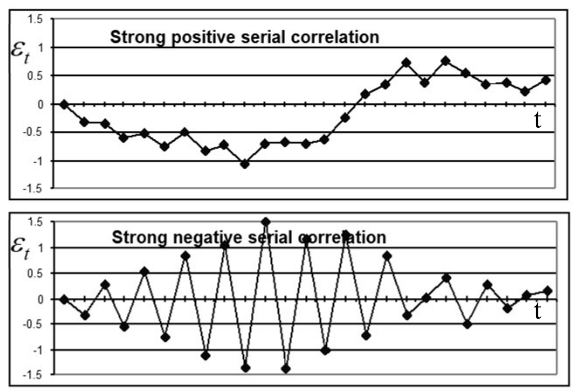
What if assumption #4 did not hold?

- Unobserved factors influencing y in one observation are correlated with those in another observation.
- Example: Cross-section data—neighboring states or cities correlated.
- Example: Time-series data—serial correlation.

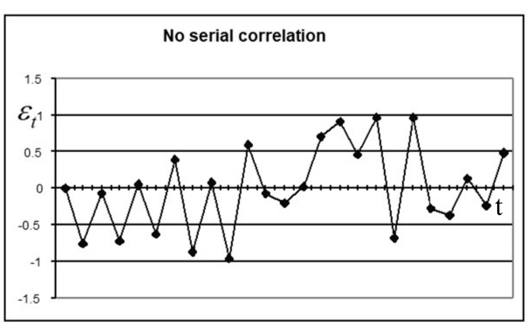
Serial correlation, the most common kind of autocorrelation

- Serial correlation can be
  - positive: Cov (ε_t, ε_{t-1}) > 0.
  - or negative: Cov (ε_t, ε_{t-1}) < 0.
- Positive serial correlation means if ε_t will tend to have the ____________ sign as ε_{t-1}.
- Negative serial correlation means if ε_t will tend to have the ____________ sign as ε_{t-1}.

Examples of serial correlation



Example of no serial correlation



Conclusions

- Additional useful assumptions are:
  - Assumption #3: Var (ε_i) = σ^2 (________________________).
  - Assumption #4: Cov (ε_i,ε_j) = 0 (no ______________________).
- Assumptions #1 through #4 are sometimes called "Gauss-Markov assumptions."

PROPERTIES UNDER ADDITIONAL ASSUMPTIONS

PROPERTIES UNDER ADDITIONAL ASSUMPTIONS

- What do the additional assumptions of homoskedasticity and no autocorrelation buy us?

Additional assumptions yield additional properties

- Under these additional assumptions:
  - Assumption #3: (homoskedasticity).
  - Assumption #4: (no autocorrelation).

  can drastically simplify the formula for variance of LS estimators.
- Can also show additional useful properties of LS:
  - Gauss-Markov theorem
  - asymptotic normality.

Variance of least-squares estimator for slope

$$Var(\hat{\beta}_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(\varepsilon_i)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} + \frac{\sum_{i=1}^{n}\sum_{j\neq i}(x_i - \bar{x})(x_j - \bar{x})Cov(\varepsilon_i, \varepsilon_j)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Implications of no autocorrelation

$$Var(\hat{\beta}_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(\varepsilon_i)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} + \frac{\sum_{i=1}^{n}\sum_{j\neq i}(x_i - \bar{x})(x_j - \bar{x})Cov(\varepsilon_i, \varepsilon_j)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Implications of homoskedasticity

$$Var(\hat{\beta}_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(\varepsilon_i)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Formulas for the true variances and covariance of LS estimators

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

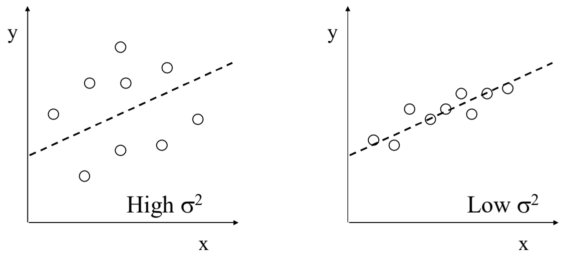
$$Var(\hat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}$$

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\bar{x}\sigma^2}{\sum(x_i - \bar{x})^2}$$

PROPERTIES UNDER ADDITIONAL ASSUMPTIONS

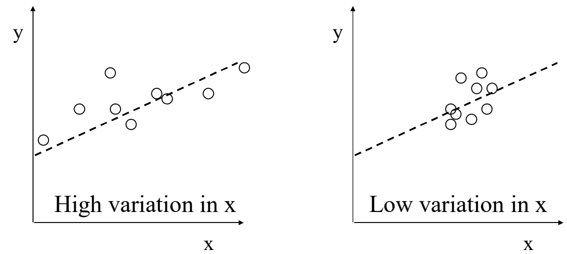
These variance formulas show that the greater the variance of the error term (σ^2) ...

- … the _____________ the variances of the least-squares estimators.



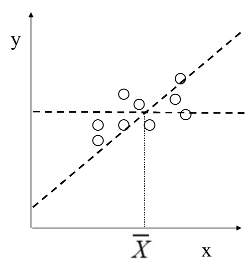
These variance formulas also show that the greater the variation in x around its sample mean ...

- … the _____________ the variance of the least-squares estimators.



These variance formulas also show that if the sample mean of x is positive...

- … the slope and intercept estimators are _______________ correlated with each other.



Estimating the variance of the error term

- The true value of σ^2 is unknown. So these formulas cannot be applied directly.
- But an unbiased estimator of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}^2$$

Estimating the variances and covariance of LS estimators

$$\text{Estimator for } Var\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

$$\text{Estimator for } Var\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2}$$

$$\text{Estimator for } \mathrm{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) = \frac{-\bar{x}\,\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

Standard errors of LS estimators

- Estimates of the standard deviations of the LS estimators are given by:

$$SE\left(\hat{\beta}_2\right) = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}$$

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2}}$$

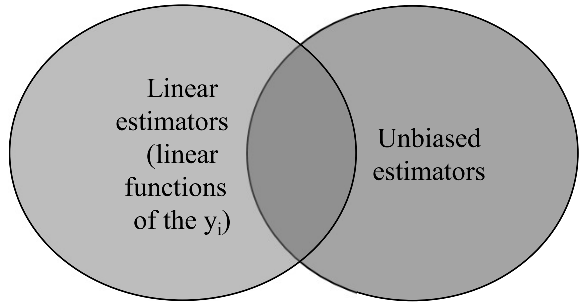
PROPERTIES UNDER ADDITIONAL ASSUMPTIONS

Reporting SEs

- Regression software always computes SEs.
- In research papers reporting results for LS regression, SEs are usually reported underneath coefficient estimates, in parentheses. Example:

$$\begin{array}{ccccc} wage & = & 1.23 & + & 0.12 \quad educ \\ & & (0.56) & & (0.03) \end{array}$$

Classes of estimators



The Gauss-Markov theorem

- It can be shown that, given assumptions #1 through #4, LS estimators have the lowest variance of all linear unbiased estimators.
- They are the Best Linear Unbiased Estimators (BLUE).

Asymptotically normal

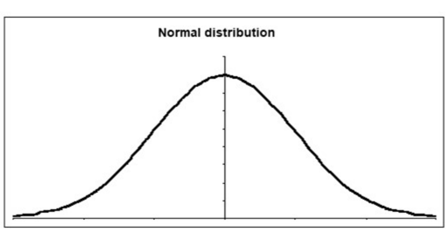
- As the sample size increases, the distribution of the LS estimators approaches a normal distribution around the true values of the coefficients.

$$\hat{\beta}_2 \overset{A}{\sim} N\left(\beta_2, \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \overset{A}{\sim} N\left(\beta_1, \frac{\hat{\sigma}^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right)$$

Why asymptotic normality is important

- This property allows easy calculation of asymptotic confidence intervals and tests.



Conclusions

Assuming homoskedasticity and no autocorrelation, LS estimators

- have variances that can be estimated using fairly simple formulas.
- are B______ L________ U__________ E__________ (Gauss-Markov theorem).
- have asymptotically ________________ distributions.

ASYMPTOTIC CONFIDENCE INTERVALS AND TESTS

ASYMPTOTIC CONFIDENCE INTERVALS AND TESTS

•How can we calculate confidence intervals and tests using only the classical assumptions?

Asymptotic distribution of least-squares estimators

- Under these four assumptions:
  - Assumption #1: $E(\varepsilon_i) = 0$.
  - Assumption #2: $E(\varepsilon_i|x_i) = 0$.
  - Assumption #3: homoskedasticity.
  - Assumption #4: no autocorrelation.

the distribution of LS estimators is asymptotically normal.

Asymptotic distribution of least-squares estimators (cont'd)

- Formally:

$$\hat{\beta}_2 \overset{A}{\sim} N\left(\beta_2, \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \overset{A}{\sim} N\left(\beta_1, \frac{\hat{\sigma}^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}\right)$$

Asymptotic confidence intervals

- We can use the asymptotic normal distribution to form confidence intervals for the true slope and intercept.
- Sample size should be reasonably large (ideally, at least _________ observations).

Formulas for asymptotic 95% confidence intervals

$$\hat{\beta}_2 \pm 1.96 \cdot SE(\hat{\beta}_2) = \hat{\beta}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}$$

$$\hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1) = \hat{\beta}_1 \pm 1.96 \cdot \sqrt{\frac{\hat{\sigma}^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}}$$

Formulas for asymptotic 90% confidence intervals

$$\hat{\beta}_2 \pm 1.645 \cdot SE(\hat{\beta}_2) = \hat{\beta}_2 \pm 1.645 \cdot \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}$$

$$\hat{\beta}_1 \pm 1.645 \cdot SE(\hat{\beta}_1) = \hat{\beta}_1 \pm 1.645 \cdot \sqrt{\frac{\hat{\sigma}^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}}$$

ASYMPTOTIC CONFIDENCE INTERVALS AND TESTS

Example 1

- We are interested in the demand for college admission.
- We estimate the relationship between between tuition level and enrollment using a sample of 100 small colleges:

$$\underset{}{\textit{enroll}} = \underset{(200.1)}{1127.4} - \underset{(0.2)}{0.3}\ \textit{tuition}$$

Example 1 (cont'd)

- The 95% asymptotic confidence interval for β_2 is -0.3 ± 1.96 (0.2) = -0.3 ± ________ or (__________,___________)
- The 90% asymptotic confidence interval for β_2 is -0.3 ± 1.645 (0.2) = -0.3 ±_______ or (__________,___________)

Asymptotic tests

- We can use the asymptotic normal distribution to test hypotheses about the true slope and intercept.
- Sample size should be reasonably large (ideally, at least 100 observations).

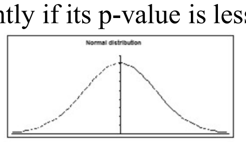
Calculating t-statistics

- Calculate the t-statistic by subtracting the hypothesized value (b) from the estimate, and then dividing by the standard error.

$$\frac{\hat{\beta}_2 - b}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - b}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}} \quad \text{or} \quad \frac{\hat{\beta}_1 - b}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - b}{\sqrt{\dfrac{\hat{\sigma}^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}}}$$

Interpreting t-statistics

- If null hypothesis is true, t-statistic has asymptotic standard normal distribution and its value is usually near zero.
- So reject null hypothesis if its value is far from zero (past the critical point) or equivalently if its p-value is less than the test size.



Example 2

- We want to know whether the demand for water increases with income—in economic terms, whether water is a normal good.
- We estimate the relationship between between income and water consumption using a sample of 500 households:

$$\textit{water} = \underset{(45.6)}{237.5} + \underset{(0.33)}{0.78}\ \textit{income}$$

ASYMPTOTIC CONFIDENCE INTERVALS AND TESTS

Example 2 (cont'd)

- We must test whether x has no effect on y, that is H_0: β_2=0 against H_1: $\beta_2 \neq 0$, at 5% significance.
- t-statistic = (0.78-0)/0.33 = _________.
- The critical points at 5% are ±1.96.
- So __________ H_0 at 5% significance.
- Note: P-value = Prob{|Z|>2.36} = 0.0091.

Hypothesized value different from zero

- Most regression software automatically computes t-statistics for H_0: β = 0.
- But sometimes a hypothesized value other than b = 0 is of interest.
- Easy to compute with calculator or spreadsheet!

Example 3

- Suppose we want to know whether the Keynesian marginal propensity to consume is exactly one.
- Using a large macroeconomic data set of national income and consumption, we estimate the Keynesian consumption function:

$$\underset{}{consumption} = \underset{(234.5)}{200.6} + \underset{(0.04)}{0.93}\ income$$

Example 3 (cont'd)

- We must test H_0: β_2=1 against H_1: $\beta_2 \neq 1$, at 5% significance.
- The t-statistic = (0.93-1)/0.04 = _________.
- The critical points at 5% are ±1.96.
- So __________ H_0 at 5% significance.
- Note: P-value = Prob{|Z|>1.75} = 0.0802.

Hypotheses about the intercept

- Once in a while, the value of the intercept is of interest.
- For example, if intercept is zero, then y is proportional to x: $y = \beta_2 x$.
- Testing H_0: $\beta_1 = 0$ is straightforward. This t-statistic is automatically computed by most regression software.

Example 4

- We are interested in whether firms enjoy constant returns to scale—that is, cost proportional to output.
- We estimate the relationship between output and total cost using a data set on firms:

$$cost = \underset{(0.70)}{1.53} + \underset{(0.06)}{0.34}\ output$$

ASYMPTOTIC CONFIDENCE INTERVALS AND TESTS

Example 4 (continued)

- We must test H_0: $\beta_1=0$ against H_1: $\beta_1 \neq 0$, at 5% significance.
- t-statistic = (1.53-0)/0.70 = ___________.
- The critical points at 5% are ±1.96.
- So ______________ H_0 at 5% significance.
- Note: P-value = Prob{|Z|>2.19} = 0.0286.

"Accepting" versus "not rejecting" H_0

$$t = \frac{\hat{\beta} - b}{SE(\hat{\beta})}$$

- Possible reasons for a low t-statistic:
  - (1) Estimated coefficient $\hat{\beta}$ is close to hypothesized value b. This supports H_0.
  - (2) Standard error is large. This does _________ support H_0. Just indicates ignorance about the true value.
- Better to say "cannot reject H_0," rather than "accept H_0."

Conclusions

- Assuming homoskedasticity and no autocorrelation,
  - LS estimators have distributions which are asymptotically ___________,
  - asymptotic confidence intervals and t-tests can be computed using the standard ____________ distribution.

PREDICTION WITH TWO-VARIABLE REGRESSION

PREDICTION WITH TWO-VARIABLE REGRESSION

- How can we predict values of y outside our sample?

What if…?

- An important use of LS estimates is “what if” or *conditional prediction*.
- Example: we have estimated the relation between tax rates and tax revenue. What if tax rates are set at some new level?
- Another example: we have estimated the relation between interest rates and investment. What if interest rates are raised?

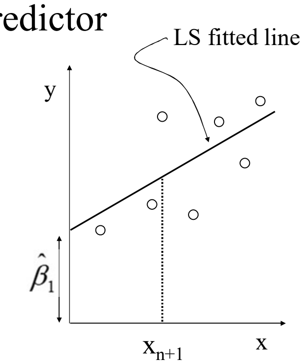
Prediction using LS

- Suppose we have estimated a linear relationship between x and y using LS.
- Given another value of x_{n+1} (not in our sample) how can we use our estimates to predict the corresponding value of y_{n+1}?

LS predictor

- LS predictor uses same formula as formula for fitted values:

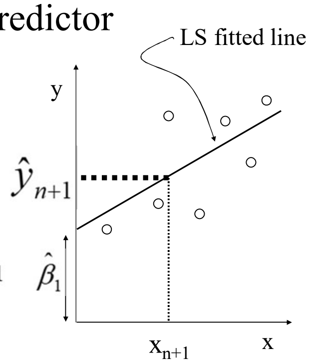
$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$$



LS predictor

- LS predictor uses same formula as formula for fitted values:

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$$



Prediction error

- Predictions are never exactly correct:
  $$\hat{y}_{n+1} \neq y_{n+1}$$
- True value: $y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$.
- LS prediction error:
  $$\hat{y}_{n+1} - y_{n+1} = \left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}\right) - \left(\beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}\right)$$

PREDICTION WITH TWO-VARIABLE REGRESSION

Sources of prediction error

- LS prediction error results from
  (1) errors in estimating β_1 and β_2
  (2) the new error term ε_{n+1}.

$$\hat{y}_{n+1} - y_{n+1}$$
$$= \left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}\right) - \left(\beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}\right)$$
$$= \left(\hat{\beta}_1 - \beta_1\right) + \left(\hat{\beta}_2 - \beta_2\right) x_{n+1} - \varepsilon_{n+1}$$

LS prediction is unbiased

- Prediction error is inevitable, but the *expected value* of LS prediction error is zero because the LS estimators are unbiased.

$$E\left(\hat{y}_{n+1} - y_{n+1}\right)$$
$$= E\left(\hat{\beta}_1 - \beta_1\right) + E\left(\hat{\beta}_2 - \beta_2\right) x_{n+1} - E\left(\varepsilon_{n+1}\right)$$

Variance of prediction error

$$Var\left(\hat{y}_{n+1} - y_{n+1}\right) = Var\left(\hat{\beta}_1\right) + Var\left(\hat{\beta}_2\right) x_{n+1}^2$$
$$+ 2Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) x_{n+1} + Var\left(\varepsilon_{n+1}\right)$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{\sum \left(x_i - \bar{x}\right)^2} + 1 \right)$$

Variance of prediction error (cont'd)

- Formula is complicated. No need to memorize.
- However, *do* memorize the implications of the formula, on the next slide.
- And remember that a large variance of prediction error means we ______________ predict y_{n+1} precisely.
- Small variance means we ________ predict y_{n+1} precisely.

The formula shows that the variance of LS prediction error is ***smaller*** when...

- ... the sample size (n) is _____________.
- ... the variation of x_i in the sample is _____________.
- ... x_{n+1} is __________ to the sample mean.
- However, the variance of LS prediction error can never be less than _____.

LS predictor is best unbiased

- It can be shown that, given assumptions #1 through #4, the LS predictor has the ________________ variance of all linear unbiased predictors.
- LS predictor is therefore called the Best Linear Unbiased Predictor (BLUP).

PREDICTION WITH TWO-VARIABLE REGRESSION

Conclusions

- The least squares predictor for y_{n+1} uses formula for fitted values, applied to x_{n+1}.
- Prediction error arises from estimation error, and the new error term ε_{n+1}.
- Assuming homoskedasticity and no autocorrelation, LS predictor for y_{n+1} is B____ L_______ U_______ P_________.

THE ASSUMPTION THAT ERROR TERMS ARE NORMALLY-DISTRIBUTED

THE ASSUMPTION THAT ERROR TERMS ARE NORMALLY-DISTRIBUTED

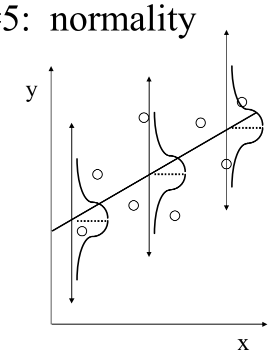
- What final assumption is useful for small samples?

Small samples

- If the sample size is small (say, less than 50) the asymptotic distribution of the LS estimators is not likely to be an accurate approximation.
- But the exact distribution can be derived if we make one more assumption.

Assumption #5: normality

- The error terms follow a normal distribution.
- $\varepsilon_i \sim N(0,\sigma^2)$.
- Recall: normal distribution has bell-shaped density function.

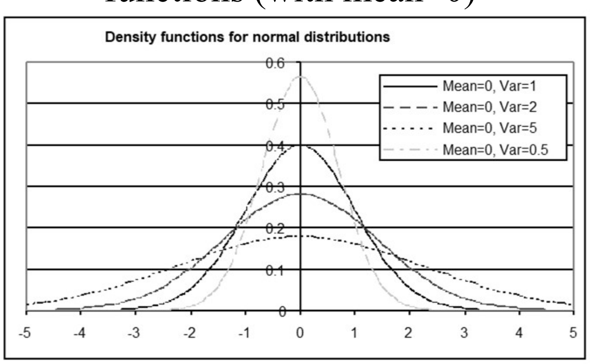


Density function for the error term

- The formula for the density function is:

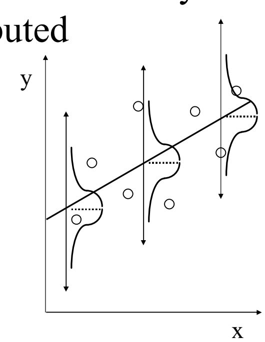
$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\varepsilon_i^{\,2}}{2\sigma^2}\right)$$

Examples of normal density functions (with mean=0)



Given x_i, y_i is also normally-distributed

- Fact: Linear functions of normally-distributed random variables are also normally-distributed.
- But $y_i = \beta_1+\beta_2 x_i+\varepsilon_i$.
- So, given x_i , $y_i \sim N(\beta_1+\beta_2 x_i, \sigma^2)$.



THE ASSUMPTION THAT ERROR TERMS ARE NORMALLY-DISTRIBUTED

Density function for the dependent variable (y_i)

- Substituting y_i - [$\beta_1 + \beta_2 x_i$] for ε_i , we derive the conditional density function for y_i, given x_i, as:

$$f(y_i|x_i) =$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y_i-[\beta_1+\beta_2 x_i])^2}{2\sigma^2}\right)$$

Conclusions

- If the sample size is small, it is useful to assume the error term ε_i follows a normal distribution with mean zero.
- In that case, the dependent variable y_i also follows a _______________ distribution with conditional mean ________________.

PROPERTIES WITH NORMALLY-DISTRIBUTED ERROR TERMS

PROPERTIES WITH NORMALLY-DISTRIBUTED ERROR TERMS

- What does the additional assumption of normally-distributed error terms buy us?

Additional assumption yields additional properties

- Under the additional assumption that the error terms ε_i are normally-distributed, we can show additional useful properties of LS:
  - LS estimators are ML estimators.
  - LS estimators are best in a broader class than just linear unbiased estimators.
  - LS estimators also follow normal distributions.

Independence of error terms from each other

- If error terms ε_i are normally distributed, they are *independent*, not just uncorrelated.
- This implies the joint density function of the error terms = the product of the individual density functions: $f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) = f(\varepsilon_1)\, f(\varepsilon_2) \ldots f(\varepsilon_n)$, where

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\varepsilon_i^{\,2}}{2\sigma^2}\right)$$

Independence of y_i from each other

- Similarly y_i are *independent*, given x_i.
- This implies the joint density function of the y_i= the product of the individual conditional density functions: $f(y_1, y_2,\ldots, y_n) = f(y_1)\, f(y_2) \ldots f(y_n)$, where

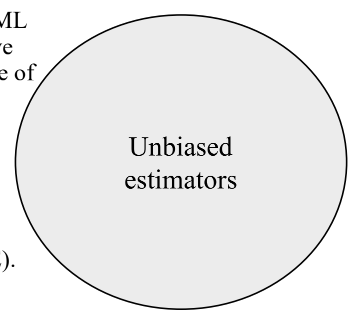
$$f(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - [\beta_1 + \beta_2 x_i])^2}{2\sigma^2}\right)$$

The maximum-likelihood principle

- It can be shown that LS estimators maximize the joint density function of the data, $f(y_1, y_2, \ldots, y_n)$.
- That is, the LS estimators follow the principle of ________________________.
- This is important because most ML estimators are consistent and, if they are unbiased, are *best unbiased*.

Lowest variance

- Because they are ML estimators, LS have the lowest variance of ALL unbiased coefficient estimators (linear or not).
- They are the Best Unbiased Estimators (BLUE).



PROPERTIES WITH NORMALLY-DISTRIBUTED ERROR TERMS

Exact distribution of LS estimators

- LS estimators are linear functions of the y_i and (by implication) of the error terms ε_i .
- This implies LS estimators are exactly normally-distributed, even in small samples.

$$\hat{\beta}_2 \sim \mathrm{N}\left(\beta_2, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right), \quad \hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2}\right)$$

Standardizing the LS estimators

- In other words, the following functions of LS estimators follow standard normal distributions:

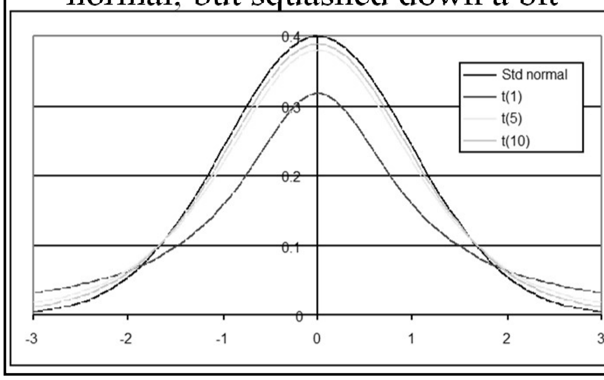
$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}}} \sim \mathrm{N}(0,1), \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2}}} \sim \mathrm{N}(0,1)$$

t-statistics follow t distributions (exactly)

- It can be shown that if σ^2 is replaced by its unbiased estimator, the resulting expressions each have a t distribution with n-2 degrees of freedom:

$$\frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \sim t_{(n-2)} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{(n-2)}$$

t-distribution is like standard normal, but squashed down a bit



Conclusions

Assuming the error terms are normally-distributed,

- LS estimators are are B________ U__________ E____________ .
- the LS estimators have exactly ________ distributions, even in small samples.
- subtracting the true values and dividing by the standard errors gives t-statistics with ________ degrees of freedom.

EXACT CONFIDENCE INTERVALS AND TESTS

EXACT CONFIDENCE INTERVALS AND TESTS

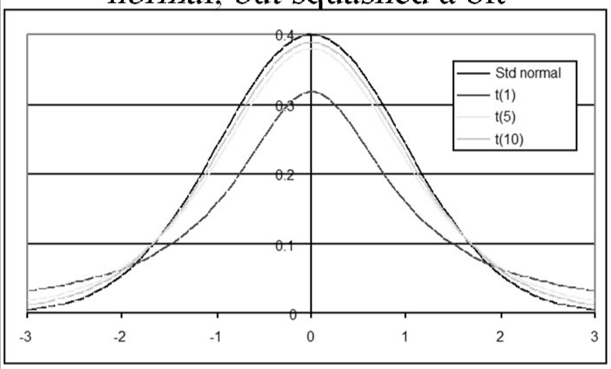
- How can we calculate confidence intervals and tests exploiting the assumption that the error term is normally-distributed?

t-statistics follow t distributions (exactly)

- Assuming the error terms ε_i are normally-distributed, we can use the t-statistics to calculate exact confidence intervals and tests that are valid even in small samples.

$$\frac{\hat{\beta}_2-\beta_2}{SE\left(\hat{\beta}_2\right)} \sim t_{(n-2)} \qquad \frac{\hat{\beta}_1-\beta_1}{SE\left(\hat{\beta}_1\right)} \sim t_{(n-2)}$$

t-distribution is like standard normal, but squashed a bit



Exact confidence intervals

- Use confidence point c from the t-distribution with n-2 degrees of freedom, at desired confidence level.

$$\hat{\beta}_2 \pm c \cdot SE\left(\hat{\beta}_2\right) = \hat{\beta}_2 \pm c \cdot \sqrt{\frac{\hat{\sigma}^2}{\sum\left(x_i-\bar{x}\right)^2}}$$

$$\hat{\beta}_1 \pm c \cdot SE\left(\hat{\beta}_1\right) = \hat{\beta}_1 \pm c \cdot \sqrt{\frac{\hat{\sigma}^2 \sum x_i^2}{n\sum\left(x_i-\bar{x}\right)^2}}$$

Example 1

- We estimate the following demand curve for a grocery item using a sample of 15 cities.
- We want to calculate a confidence interval for the coefficient of price.

$$\underset{}{quantity} = \underset{(8.7)}{45.2} - \underset{(0.2)}{0.3}\ price$$

Example 1 (cont'd)

- Since n=15, DOF = _______.
- 95% confidence interval: c = 2.160 so confidence interval is -0.3 ± 0.432, or (____________, ___________).
- 90% confidence interval: c = 1.771 so confidence interval is -0.3 ± 0.3542, or (____________, ___________).

EXACT CONFIDENCE INTERVALS AND TESTS

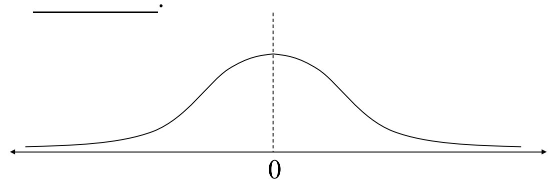
Exact tests using t-statistics

- Calculate the t-statistic by subtracting the hypothesized value (b) from the estimate, and then dividing by the standard error.

$$\frac{\hat{\beta}_2 - b}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - b}{\sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}} \quad \text{or} \quad \frac{\hat{\beta}_1 - b}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - b}{\sqrt{\frac{\hat{\sigma}^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}}}$$

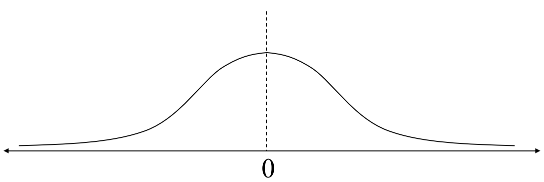
Interpreting t-statistics

- If null hypothesis is true, t-statistic follows t distribution exactly with n-2 degrees of freedom and its value is usually near _________.



Interpreting t-statistics (cont'd)

- So reject null hypothesis if its value is ______ from zero, past the critical point, or equivalently if its p-value is __________.



Example 2

- We estimate the relationship between income and energy use with a sample of 12 countries.
- We want to test the null hypothesis that income has no effect on energy use.

energy use	=	67.8	+	0.78	*income*
per capita		(12.3)		(0.33)	*per capita*

Example 2 (cont'd)

- We must test H_0: $\beta_2=0$ against H_1: $\beta_2 \neq 0$, at 5% significance.
- The t-statistic = (0.78-0)/0.33 = -2.36.
- Since n=12, DOF = _________, and the critical points at 5% are ± 2.228 .
- So ____________ H_0 at 5% significance.
- Note: P-value = Prob{|W|>2.36} = 0.040.

Example 3

- We estimate the relationship between water use and the price of water with a sample of 16 communities.
- The estimated coefficient of price is negative, but is this just sampling error?

water use	=	45.7	–	2.8	*price*
per capita		(9.5)		(1.6)	*of water*

EXACT CONFIDENCE INTERVALS AND TESTS

Example 3 (cont'd)

- We want to test the null hypothesis that price has no effect on water use, against the ___________-sided alternative that it has a negative effect.
- No one believes that price could have a *positive* effect on water use!

water use	=	45.7	–	2.8	*price*
per capita		(9.5)		(1.6)	*of water*

Example 3 (cont'd)

- We must test H_0: β_2=0 against H_1: β_2<0, at 5% significance.
- The t-statistic = (-2.8-0)/1.6 = -1.75.
- Since n=14, DOF = _________, and the one-tailed critical point at 5% is -1.782 .
- So ______________ H_0 at 5% significance.
- Note: P-value = Prob{W<-1.782} = 0.053.

Summary: What distribution to use for CIs and tests

		Distribution of error term: Normal	Distribution of error term: Unknown
Sample size	Small (<50)		
	Large (>100)		

Conclusions

- Assuming the error terms ε_i are normally-distributed, we can use the _____________ to calculate exact confidence intervals and hypothesis tests.
- These are valid even in _________ samples.

PREDICTION INTERVALS

PREDICTION INTERVALS

- How can we calculate prediction intervals exploiting the assumption that the error term is normally-distributed?

Conditional prediction

- An important use of LS estimates is "what if?" or *conditional prediction*.
- Given a new value of the x variable, we may wish to predict the value of y.
- The LS predictor $\hat{y}_{n+1}$ simply substitutes the new value x_{n+1} into the estimated equation.

LS predictor is best unbiased

- As mentioned earlier, given assumptions #1 through #4, the LS predictor has the ________________ variance of all linear unbiased predictors.
- LS predictor is therefore called the Best Linear Unbiased Predictor (BLUP).

Example 1

- Suppose we have estimated the following equation relating house size (in square feet) to selling price (in thousands of dollars) in a particular neighborhood:

$$\text{price}_i = \underset{(13.1)}{102} + \underset{(0.008)}{0.0471}\ \text{size}_i$$

Example 1: LS predictor

- Suppose we wish to predict the selling price of a house, outside our sample. So give it a new subscript, _________.
- Suppose size_{n+1} = 2000 square feet.
- LS predictor is computed by substituting this value in the estimated equation:

$$\widehat{\text{price}}_{n+1} = 102 + 0.0471\ \text{size}_{n+1}$$
$$= 102 + 0.0471\ (2000) = \_\_\_\_\_\_\_$$

Prediction error

- Predictions are never exactly correct:

$$\hat{y}_{n+1} \neq y_{n+1} .$$

- As mentioned earlier, LS prediction error results from
  (1) errors in estimating βs.
  (2) the new random error term ε_{n+1}.

PREDICTION INTERVALS

Standard error of prediction

- *Standard error of prediction error* is an estimate of the standard deviation of the prediction error.
- Earlier, a formula was given for the variance of the prediction error.
- Inserting $\hat{\sigma}^2$ for σ^2 and taking square root gives the standard error of prediction error:

$$SE(\hat{y}_{n+1} - y_{n+1}) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2} + 1 \right)}$$

Distribution of prediction error

- Assuming the error terms are normally-distributed,

$$\frac{\hat{y}_{n+1} - y_{n+1}}{SE \text{ of prediction error}} \sim t_{n-2}$$

Exact prediction intervals

- Use confidence point c from the t-distribution with n-2 degrees of freedom, at desired confidence level.
- Formula is similar to confidence interval:

$$\hat{y}_{n+1} \pm c \cdot (\text{SE of prediction error})$$

Example 1: prediction interval

- Assume we have calculated the standard error of prediction error at 7.8 .
- Suppose n=25. Since DOF = _______, a table of the t-distribution gives c at 95% = 2.069 .
- 95% prediction interval:
  \$196.2 ± 2.069 (7.8) = \$196.2 ± \$16.1,
  or (\$__________ , \$ __________).

Computing variance of prediction error: a trick

- Formula for the variance of the prediction error, given above, is tedious.
- In practice, easier to use the following trick.

Easy way to compute prediction and SE of prediction error

(1) Transform the data on the x variable by subtracting the value of x_{n+1}.

(2) Re-estimate equation using the transformed x data.

(3) Use the ________________ ($\tilde{\beta}_1$) of re-estimated equation for prediction.

(4) SE of prediction error = $\sqrt{SE(\tilde{\beta}_1)^2 + \hat{\sigma}^2}$

PREDICTION INTERVALS

Example 1: computing SE of prediction error

- Suppose we wish to predict the selling price of a house with $size_{n+1} = 2500$ square feet.

(1) Transform data: subtract 2500 from the size of all the houses in the original data:

$$\widetilde{size}_i = size_i - 2500$$

Example 1: computing SE of prediction error (cont'd)

(2) Re-estimate equation using the transformed x data:

$$\underset{}{price_i} = \underset{(3.9)}{219.8} + \underset{(0.008)}{0.0471}\, \widetilde{size}_i$$

$$\hat{\sigma}^2 = 27.04$$

- The coefficient of the x variable will ____ change, but the intercept ________ change.

Example 1: computing SE of prediction error (cont'd)

(3) Use intercept of re-estimated equation for prediction:

- By definition, $\widetilde{size}_{n+1} = 2500 - 2500 = 0$, so $\widehat{price}_{n+1} = 219.8 + 0 =$ _________.

(4) Compute SE of prediction error:

$$\sqrt{SE\left(\tilde{\beta}_1\right)^2 + \hat{\sigma}^2} = \sqrt{3.9^2 + 27.04} = \_\_\_\_\_\_\_$$

Example 1: new prediction interval

- Using these results we can quickly compute a prediction interval when $size_{n+1} = 2500$.
- Recall n=25 and DOF = 23, so c at 95% = 2.069 .
- 95% prediction interval: $219.8 ± 2.069 (6.5) = $219.8 ± $13.45, or ($__________ , $ __________).

Conclusions

- Assuming the error terms ε_i are normally-distributed, we can use the _____________ to calculate exact prediction intervals.
- These are valid even in _________ samples.
- The SE of prediction error is most easily calculated from a regression on _________________ x data.

SUMMARY OF PROPERTIES OF LEAST-SQUARES ESTIMATORS

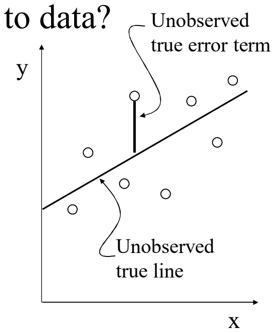
SUMMARY OF PROPERTIES OF LEAST-SQUARES ESTIMATORS

- What is so great about least-squares?

Why do we use LS to fit lines to data?

If the data (specifically, the unobserved true error term) satisfy certain assumptions,

then LS has a number of very good properties that other methods do not have.



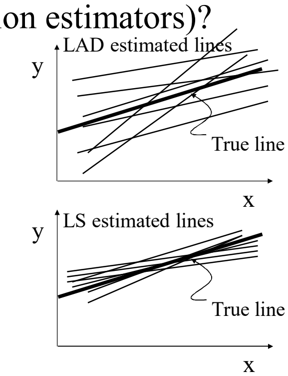
Assumptions about data	Properties of LS estimators
Start with these assumptions: #1: $E(\varepsilon_i) = 0$ #2: $E(\varepsilon_i \vert x_i) = 0$	•LS estimators are unbiased. •LS estimators are consistent.
Add these assumptions: #3: ε_i are homoskedastic #4: ε_i are not autocorrelated	•LS estimators are BLUE. •Usual formulas for SEs, confidence intervals and tests are valid for large sample.
Add one more assumption: #5: ε_i follow a normal distribution	•LS estimators are BUE. •Usual formulas for SEs, confidence intervals and tests are valid for any size sample.

Why don't we use LAD (least absolute deviation estimators)?

If assumptions #1 through #5 hold,

then LAD has _______ variance than LS (it is not _________).

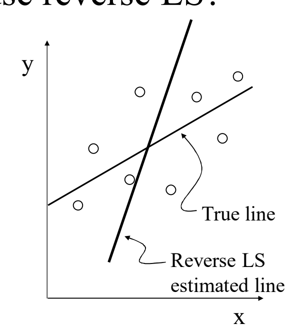
In repeated samples, the LS estimated lines tend to be closer to the unobserved true line.



Why don't we use reverse LS?

If assumptions #1 and #2 hold,

then reverse LS is ____________ and _______________.



Summary

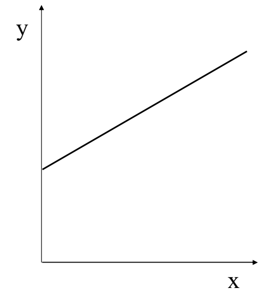
- We use LS rather than LAD, reverse LS, or some other method because, under assumptions #1 through #5, LS has ____________ properties.
- The stronger the assumptions we are willing to make about the data, the ____________ its properties compared to other methods.
- If assumptions #1 through #5 do not hold, then LS may not be better than other methods.

ALTERNATIVE FUNCTIONAL FORMS

- How can two-variable linear regression be used to fit nonlinear relationships?

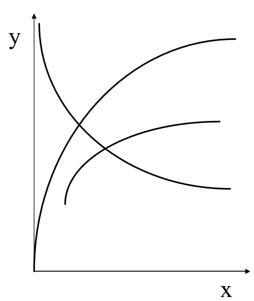
The linear functional form

- Is it realistic to assume all relationships are straight lines?
  $y = \beta_1 + \beta_2 x$.



Realism

- In fact, curved relationships are surely common.
- Examples:
  - production with diminishing marginal product.
  - demand with constant elasticity.



A trick

- Nonlinear relationships can be fitted by *transforming* x or y before fitting the linear regression.
- If f(.) and g(.) are specified, we can use ordinary LS to fit:
  $f(y) = \beta_1 + \beta_2 g(x)$.
- If f(y) or g(x) are nonlinear, then the relationship between x and y is no longer linear.

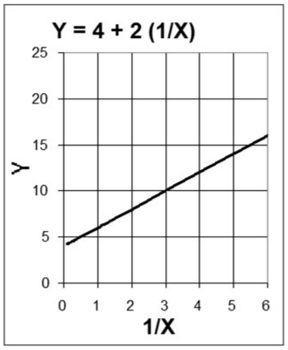
Independent variable in reciprocal form

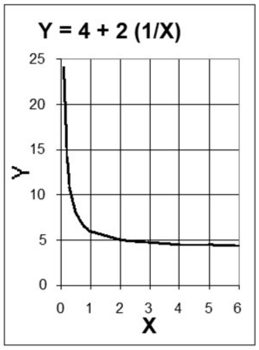
- $y = \beta_1 + \beta_2 (1/x)$.
- $dy/dx = - \beta_2 (1/x^2)$.
- As $x \to$ infinity, $y \to$ _______.
- Interpretation: Asymptotic to y-axis: ____ forms lower bound.
- Note: Discontinuous at x=0, so x data should be either strictly positive or strictly negative.

Independent variable in reciprocal form: interpretation

- Asymptotic to y-axis: β_1 forms lower bound.
- Example: Suppose we have $y = 4 + 2(1/x)$ and x is strictly positive.
- Then the lower bound for y is _______.

Independent variable in reciprocal form: shapes of curves





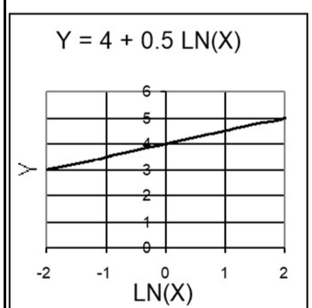
Independent variable in natural logarithm form

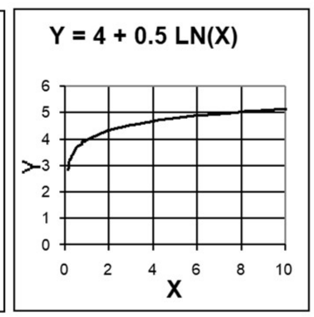
- $y = \beta_1 + \beta_2 \ln(x)$.
- dy/dx = ________.
- Note: Must have x>0.
- Asymptotic to __________.

Independent variable in natural logarithm form: interpretation

- A 1% increase in x causes a (β_2 times 0.01) increase in y.
- Example: Suppose we have $y = 4 + 0.5 \ln(x)$.
- Then a 5% increase in x causes a __________-unit increase in y.

Independent variable in natural logarithm form: shapes of curves





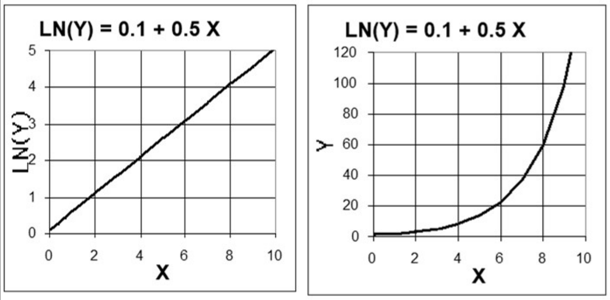
Dependent variable in natural logarithm form

- $\ln(y) = \beta_1 + \beta_2 x$.
- dy/dx = [dy/dln(y)] [dln(y)/dx] = __________.
- Note: Must have y>0.
- Asymptotic to __________.
- Good choice for "human-capital" functions, where y = earnings and x = education.

Dependent variable in natural logarithm form: interpretation

- Since $dy/dx = y\,\beta_2$, so $(dy/y)/dx = \beta_2$.
- A 1-unit increase in x causes a (100 β_2) *percent* increase in y.
- Example: Suppose we have $\ln(y) = 0.1 + 0.5x$.
- Then a 0.2-unit increase in x causes a ________ percent increase in y.

Dependent variable in natural logarithm form: shapes of curves



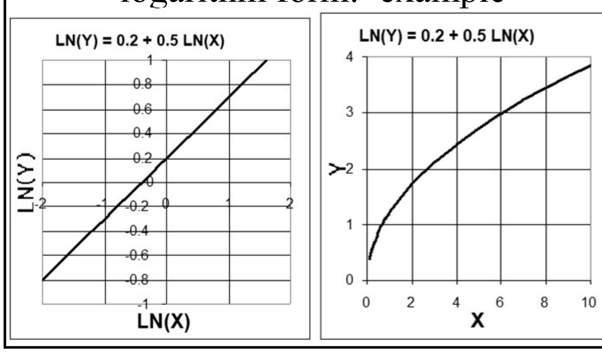
Both variables in natural logarithm form

- $\ln(y) = \beta_1 + \beta_2 \ln(x)$.
- dy/dx
  $= [dy/d\ln(y)]\,[d\ln(y)/d\ln(x)]\,[d\ln(x)/dx]$
  $= y\ \beta_2\ (1/x) =$ ____________.
- Note: Must have $y>0$ and $x>0$.
- Good choice for demand functions, supply functions, and production functions.

Both variables in natural logarithm form: interpretation

- A 1% increase in x causes a β_2 % increase in y.
- Example: Suppose we have $\ln(y) = 0.2 + 0.5 \ln(x)$.
- Then a 10 percent increase in x causes a _______ percent increase in y.
- The ______________ of y with respect to x is _______.

Both variables in natural logarithm form: example



Conclusions

- Nonlinear relationships between x and y can be fitted by transforming the data before estimation by ordinary LS.
- Common transformations include reciprocals and natural logarithms.
- If both variables are in logarithms, then β_2 is the __________ of y with respect to x.

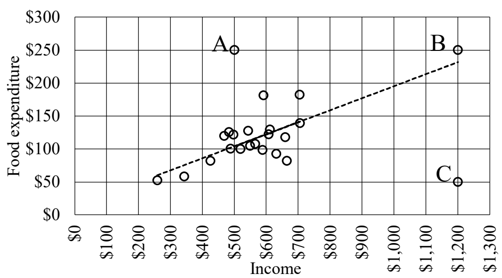
INFLUENTIAL OBSERVATIONS

- What are "influential observations"?
- Why do they merit attention?

Influential observations

- While ordinary least squares uses all of the data, some observations have more influence on the estimates than others.
- Outlier = observation whose y-value is far from the fitted line.
- High leverage point = observation whose x-value is far from the rest.

Household expenditures data with 3 new observations



Influential observations in the household expenditures data

- Observation A is an __________. It will likely increase the sum of squared residuals and lower the R-square.
- Observation B is a __________________ point. It will likely raise the R-square.
- Observation C is both a ________________ point and an __________. It will likely lower the slope and the R-square.

Actual impact of influential observations

	$\hat{\beta}_2$	t ($\hat{\beta}_2$)	R^2
Original data	0.182	3.502	0.405
Original data + A	0.182	1.863	0.155
Original data + B	0.200	6.348	0.680
Original data + C	0.011	0.242	0.003

How to find influential observations?

- Before computing LS, *always* compute descriptive statistics—mean, standard deviation, minimum and maximum.
- Do a box plot of each variable and the LS residuals.
- Print the five largest and five smallest values of each variable and the residuals.
- If the sample size is modest, do a scatter plot of y against x.

Why do influential observations occur?

- *Possibly data error.* Perhaps a zero was accidentally omitted (or inserted) when the data were collected. Perhaps a value like 999 really denotes "missing data."
- *Possibly the observation does not belong in the sample.* Perhaps the "household" is in fact a restaurant or a group home.
- *Possibly just random variation.* Perhaps the household happened to be buying food for a big party that week.

What to do about influential observations?

- Check for data errors.
- Check whether observation does not belong in sample.
- If neither of the above, do nothing.
- It is tempting to omit outliers so as to raise R-square, but then sample is no longer representative of larger population, so not a good idea.

Conclusions

- Influential observations have greater influence on regression results than other observations.
- ______________ = observation whose y-value is far from the fitted line.
- __________________________= observation whose x-value is far from the rest.

PART 3

Multiple Regression with Cross-Sectional Data

WHY INCLUDE MORE REGRESSORS?

WHY INCLUDE MORE REGRESSORS?

- Including more regressors requires more data and more computation.
- Are they worth it?

More regressors

- Two-variable regression is rarely used.
- More common is *multiple regression:*
  $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_K x_K + \varepsilon$.
- Whether our purpose is prediction or causal inference, including more regressors can help us get more useful results.

If our purpose is prediction...

- We want a model that "explains" the y_i well.
- Our model should produce predicted values $\hat{y}_i$ close to the actual values y_i.
- Adding more regressors always improves the "fit," _________ R^2 and _________ $\hat{\sigma}^2$.

Prediction example 1: dependent variable is term insurance face

Model	R^2	$\hat{\sigma}^2$
$\beta_1 + \beta_2$ income	0.0007	1.68 $\times 10^{12}$
$\beta_1 + \beta_2$ income + β_3 education	0.040	1.62 $\times 10^{12}$
$\beta_1 + \beta_2$ income + β_3 education + β_4 number in household	0.053	1.60 $\times 10^{12}$

Data from Survey of Consumer Finances. See Frees (2010) p. 70. n=500.

Prediction example 2: dependent variable is ln(food expenditure)

Model	R^2	$\hat{\sigma}^2$
$\beta_1 + \beta_2$ ln(income)	0.063	0.818
$\beta_1 + \beta_2$ ln(income) + β_3 ln(family size)	0.159	0.735
$\beta_1 + \beta_2$ ln(income) + β_3 ln(family size) + β_4 schooling	0.172	0.723

Data from Consumer Expenditure Survey 2022, Diary Survey. n=5358.

Polynomial functions sometimes improve the "fit"

- We can model y as a quadratic or possibly a cubic function of x, if we include x^2 and possibly x^3 as additional regressors.
  - Linear: ________________________
  - Quadratic: ______________________
  - Cubic: __________________________

WHY INCLUDE MORE REGRESSORS?

If our purpose is causal inference...

- We want to measure the effect of x on y, *ceteris paribus**.
- That requires measuring what happens to y when x changes, while holding constant all other factors that might influence y.
- We want unbiased estimates of the slope ________________. R^2 is unimportant.

* Latin: *other things equal.*

Omitting regressors can bias the LS slope estimators

- Suppose a variable is omitted (left in the error term) that is correlated with x.
- This violates assumption #2: $E(\varepsilon_i|x_i)=0$ or $Cov(\varepsilon_i, x_i)=0$.
- LS slope estimator will suffer from ________________________ bias.
- LS will not measure the true *ceteris paribus* effect of x.

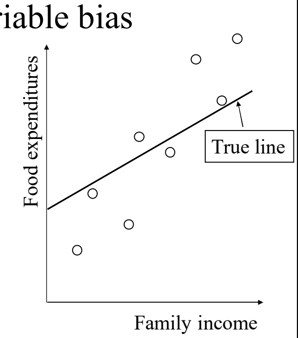
Causal inference example 1: effect of income on food expenditures

- Suppose we want to measure the effect of family income (x) on food expenditures (y) using 2-variable regression.
- Other factors like family size are left in the error term ε_i.
- But family size also affects food expenditures *and* is _______________ correlated with income: big families tend to have higher income.
- Thus $Cov(\varepsilon_i, x_i) > 0$.

Causal inference example 1: omitted variable bias

If $Cov(\varepsilon_i,x_i) > 0$, then

- observations tend to lie ________ the true line when x_i is large.
- observations tend to lie ________ the true line when x_i is small.
- LS slope estimator is biased ______.



Causal inference example 1: dependent variable is ln(food expenditure)

Model	$\hat{\beta}_2$
$\beta_1 + \beta_2$ ln(income)	0.189
$\beta_1 + \beta_2$ ln(income) + β_3 ln(family size)	0.110

Data from Consumer Expenditure Survey 2022, Diary Survey. n=5358.

Causal inference example 2: effect of schooling on earnings

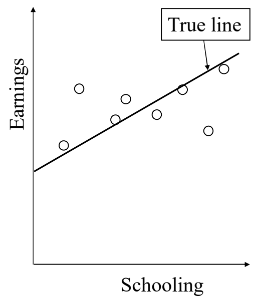
- Suppose we want to measure the effect of years of schooling (x) on earnings (y) using 2-variable regression.
- But work experience also affects earnings and is _______________ correlated with schooling.
- Thus $Cov(\varepsilon_i, x_i) < 0$.
- This means that ε_i tends to be positive when x_i is large and negative when x_i is small.

WHY INCLUDE MORE REGRESSORS?

Causal inference example 2: omitted variable bias

If Cov(ε_i, x_i) < 0, then

- observations tend to lie _________ the true line when x_i is large.
- observations tend to lie _________ the true line when x_i is small.
- LS slope estimator is biased __________.



Causal inference example 2: dependent variable is ln(earnings)

Model	$\widehat{\beta}_2$
$\beta_1 + \beta_2$ schooling	0.118
$\beta_1 + \beta_2$ schooling $+ \beta_3$ experience	0.121

Data from Current Population Survey 2023. n=68,855.

Multiple regression can estimate *ceteris paribus* relationships

- If an important regressor, correlated with the included regressor, is omitted from the regression equation, then the coefficient of the included regressor is ____________.
- Including ___________ regressors in the equation eliminates bias.

Conclusions

- Two-variable regression is often inadequate.
- For prediction, more regressors can allow more precise prediction of y, and permit ________________ functions of x.
- For causal inference, adding more regressors can prevent _________________ bias and better estimate _________________ effects.

DEFINITION OF LEAST-SQUARES WITH TWO REGRESSORS

DEFINITION OF LEAST-SQUARES WITH TWO REGRESSORS

- How can we model a relationship between y and two regressors?
- How can we estimate that relationship using least-squares?

Suppose y depends on two regressors

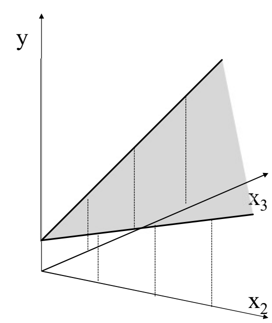
- Then $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where ε is a random error term.
- If the xs change, then the resulting change in y is given by:
- $\Delta y = \beta_2 \Delta x_2 + \beta_3 \Delta x_3$, assuming the error term does not change.

Coefficients are *ceteris paribus* effects

- If x_2 changes but x_3 is held constant, change in y is given by: $\Delta y = \beta_2 \Delta x_2$.
  - Example: If x_2 increases by 2, y increases by _______.
- If x_3 changes but x_2 is held constant, change in y is given by: $\Delta y = \beta_3 \Delta x_3$.
  - Example: If x_3 decreases by 3, y increases by _______.

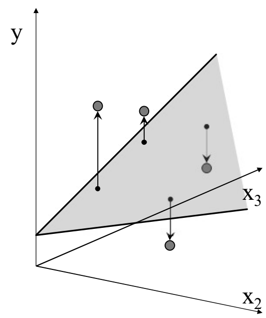
Two regressors

- Expected value of y, conditional on x_2 and x_3, forms a plane:
- $E(y|x_2,x_3) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$.
- However, we do not observe the coefficients $(\beta_1, \beta_2, \beta_3)$ of the true plane.



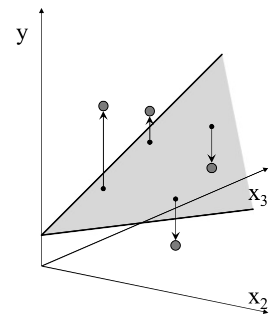
Observations rarely lie exactly on true plane

- Observations are triples: y, x_2, x_3.
- Some are above the true plane, some are below.



The least-squares principle

- How can we estimate the true plane?
- Using the available data, choose the plane that minimizes the sum of the squared vertical deviations from it.



DEFINITION OF LEAST-SQUARES WITH TWO REGRESSORS

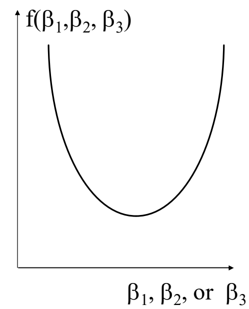
The least-squares principle (cont'd)

- In other words, find values of β_1, β_2, and β_3 that minimize the following criterion or objective function:
  $f(\beta_1, \beta_2, \beta_3) =$

$$\sum_{i=1}^{n} \left(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}] \right)^2$$

Minimizing the function

- Objective function is quadratic in β_1, β_2, or β_3 .
- Its minimum occurs where its slope = ________________



Solving for LS estimates of β_1, β_2, and β_3

- Set zero equal to derivatives of $f(\beta_1,\beta_2, \beta_3)$ with respect to β_1, β_2, and β_3 :

$$0 = \sum -2\left(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}]\right)$$
$$0 = \sum -2\left(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}]\right) x_{i2}$$
$$0 = \sum -2\left(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}]\right) x_{i3}$$

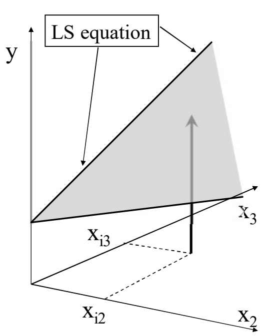
LS estimators

- These equations are first-order conditions (FONCs).
- Values of β_1, β_2, and β_3 that solve them are called the "LS estimators."
- Formulas for β_1, β_2, and β_3 are complicated but can be quickly evaluated on computers.

LS fitted values

- Fitted values are computed by inserting actual values of x_{i2} and x_{i3} into the LS estimated equation:

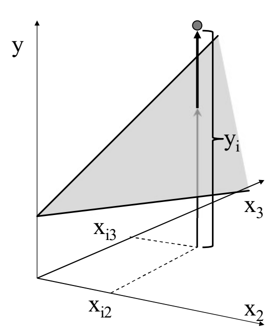
$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$$



LS residuals

- Residuals are computed as the difference between actual values of y_i in the data and the fitted values:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$



DEFINITION OF LEAST-SQUARES WITH TWO REGRESSORS

Conclusions

- Two regressors influence y if the true relationship is $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3$.
- LS estimators are values of β_1, β_2, and β_3 that minimize $\sum(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}]\ )^2$.
- LS ____________ values are defined as $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$, for i = 1,...,n.
- LS residuals are defined as $\hat{\varepsilon}_i =$ ________

ALGEBRAIC PROPERTIES OF LEAST-SQUARES WITH MULTIPLE REGRESSORS

ALGEBRAIC PROPERTIES OF LEAST-SQUARES WITH MULTIPLE REGRESSORS

- What properties of LS estimates must hold, regardless of data assumptions?

Multiple regression

- Suppose y is influenced by (K-1) xs according to the true relationship: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_K x_K$.
- This is a linear equation.
- Change in y is given by
- $\Delta y = \beta_2 \Delta x_2 + \beta_3 \Delta x_3 + \ldots + \beta_K \Delta x_K$.

Observations rarely lie exactly on the true equation

- Actual data will not lie on this equation. Some observations will lie above, some below.
- Deviations are given by $y_i - (\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_K x_{iK})$.
- Deviations may be positive or negative.

The least-squares principle

- Find values of β_1 through β_K that minimize the following quadratic objective function: $f(\beta_1,\ldots,\beta_K) =$

$$\sum_{i=1}^{n} \left(y_i - \left[\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} \right] \right)^2$$

FONCs for $\beta_1, \ldots, \beta_K$

- Set zero equal to derivatives of $f(\beta_1,\ldots,\beta_K)$ with respect to β_1 through β_K :

$$0 = \sum -2\left(y_i - \left[\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} \right] \right)$$
$$0 = \sum -2\left(y_i - \left[\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} \right] \right) x_{i2}$$
$$\vdots$$
$$0 = \sum -2\left(y_i - \left[\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} \right] \right) x_{iK}$$

LS estimators

- Values of β_1 through β_K that solve these FONCs are called the "LS estimators."
- Formulas are very complicated (unless matrix notation is used) but can be quickly evaluated on computers.

ALGEBRAIC PROPERTIES OF LEAST-SQUARES WITH MULTIPLE REGRESSORS

Fitted values and residuals

- LS fitted values defined as:
$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_K x_{iK}$$
- LS residuals defined as:
$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Note: sample means lie exactly on fitted line

- According to first FONC,
$$\begin{aligned} 0 &= \sum\left(y_i - \left[\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_K x_{iK}\right]\right) \\ &= \tfrac{1}{n}\sum\left(y_i - \left[\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_K x_{iK}\right]\right) \\ &= \tfrac{1}{n}\sum y_i - \hat{\beta}_1 - \hat{\beta}_2 \tfrac{1}{n}\sum x_{i2} - \ldots - \hat{\beta}_K \tfrac{1}{n}\sum x_{iK} \end{aligned}$$
- Therefore
$$\bar{y} = \beta_1 + \beta_2 \bar{x}_2 + \ldots + \beta_K \bar{x}_K$$

Algebraic property 1: sum of actuals = sum of fitted values

- Substituting the definition of the fitted values back into the first FONC:
$$0 = \sum\left(y_i - [\hat{y}_i]\right) = \left(\sum y_i\right) - \left(\sum \hat{y}_i\right)$$
- Thus the sum of the actual values of y equals the sum of the LS fitted values.

Algebraic property 2: sum of residuals = zero

- Alternatively, using the definition of the residuals, we can write:
$$0 = \sum\left(y_i - [\hat{y}_i]\right) = \sum \hat{\varepsilon}_i$$
- Thus the sum of LS residuals equals _____.

Algebraic property 3: sum of product of residuals and regressors = zero

- Substituting the definition of the residuals into the remaining FONCs:
$$0 = \sum\left(\hat{\varepsilon}_i\right) x_{i2}$$
$$\vdots$$
$$0 = \sum\left(\hat{\varepsilon}_i\right) x_{iK}$$
- Thus the sum of the product of the residual and any regressor equals _________.

Algebraic property 4: sum of product of fitted values and residuals = zero

$$\begin{aligned} \sum \hat{y}_i \hat{\varepsilon}_i &= \sum\left(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_K x_{iK}\right)\hat{\varepsilon}_i \\ &= \sum\left(\hat{\beta}_1 \hat{\varepsilon}_i + \hat{\beta}_2 x_{i2} \hat{\varepsilon}_i + \ldots + \hat{\beta}_K x_{iK} \hat{\varepsilon}_i\right) \\ &= \left(\sum \hat{\beta}_1 \hat{\varepsilon}_i\right) + \left(\sum \hat{\beta}_2 x_{i2} \hat{\varepsilon}_i\right) + \ldots + \left(\sum \hat{\beta}_K x_{iK} \hat{\varepsilon}_i\right) \\ &= \hat{\beta}_1\left(\sum \hat{\varepsilon}_i\right) + \hat{\beta}_2\left(\sum x_{i2} \hat{\varepsilon}_i\right) + \ldots + \hat{\beta}_K\left(\sum x_{iK} \hat{\varepsilon}_i\right) \end{aligned}$$

ALGEBRAIC PROPERTIES OF LEAST-SQUARES WITH MULTIPLE REGRESSORS

Algebraic property 5: sum-of-squares decomposition

$$\sum(y_i-\bar{y})^2=\sum(\hat{\varepsilon}_i+\hat{y}_i-\bar{y})^2$$
$$=\sum(\hat{\varepsilon}_i+[\hat{y}_i-\bar{y}])^2$$
$$=\sum\left(\hat{\varepsilon}_i^2+[\hat{y}_i-\bar{y}]^2+2\hat{\varepsilon}_i[\hat{y}_i-\bar{y}]\right)$$

Algebraic property 5: sum-of-squares decomposition (cont'd)

- But the third term is zero because:

$$\sum\hat{\varepsilon}_i[\hat{y}_i-\bar{y}]=\sum\hat{\varepsilon}_i\hat{y}_i-\sum\hat{\varepsilon}_i\bar{y}$$
$$=\left(\sum\hat{\varepsilon}_i\hat{y}_i\right)-\bar{y}\left(\sum\hat{\varepsilon}_i\right)$$

- So we have the decomposition:

$$\sum(y_i-\bar{y})^2=\sum\hat{\varepsilon}_i^2+\sum[\hat{y}_i-\bar{y}]^2$$

Algebraic property 5: sum-of-squares decomposition (cont'd)

"Residual sum of squares" or "error sum of squares"
+ "explained sum of squares" or "regression sum of squares"
= "total sum of squares."

$$\sum\hat{\varepsilon}_i^2$$
$$\underline{+\sum[\hat{y}_i-\bar{y}]^2}$$

Conclusions

- The following sums necessarily = _______:
  - LS residuals.
  - products of LS residuals and any x.
  - products of LS residuals and fitted values.
- Moreover, total sum of squares of the y
  = LS residual sum of squares
  + LS explained sum of squares.

"R^2" AND "ADJUSTED R^2"

R^2 AND ADJUSTED R^2

- How can we measure the "goodness of fit" of a regression?

Using the sum-of-squares decomposition for LS

- "Residual sum of squares" or "error sum of squares"
- + "explained sum of squares" or "regression sum of squares"
- = "total sum of squares."

$$\sum \hat{\varepsilon}_i^2 + \sum \left[\hat{y}_i - \bar{y}\right]^2 = \sum \left(y_i - \bar{y}\right)^2$$

Measuring goodness-of-fit

- A natural measure is the *fraction of the total sum of squares that is explained by the xs*.
- This is the R^2 measure:

$$R^2 = \frac{\sum \left(\hat{y}_i - \bar{y}\right)^2}{\sum \left(y_i - \bar{y}\right)^2}$$

- R^2 must lie between zero and one if the equation is estimated by LS and an intercept is included.

A second definition of R^2

- Using the decomposition property, we can alternatively rewrite R^2 as

$$R^2 = \frac{\sum \left(y_i - \bar{y}\right)^2 - \sum \hat{\varepsilon}^2}{\sum \left(y_i - \bar{y}\right)^2} = 1 - \frac{\sum \hat{\varepsilon}^2}{\sum \left(y_i - \bar{y}\right)^2}$$

Are these two definitions of R^2 always equal?

- These definitions are equal if
  - the coefficients are estimated by ordinary LS.
  - an intercept is included.
- For other methods, or if no intercept is included, the first definition may exceed one or the second definition may be negative!

Interpreting R^2

- With a little tedious algebra, it can be shown that

$$R^2 = \frac{\left(\sum \left(\hat{y}_i - \bar{y}\right)\left(y_i - \bar{y}\right)\right)^2}{\left(\sum \left(\hat{y}_i - \bar{y}\right)^2\right)\left(\sum \left(y_i - \bar{y}\right)^2\right)} = \frac{\left[\text{sample covariance}\left(\hat{y}_i, y_i\right)\right]^2}{\text{sample var}\left(\hat{y}_i\right) \times \text{sample var}\left(y_i\right)} = \left[\text{sample correlation}\left(\hat{y}_i, y_i\right)\right]^2$$

"R^2" AND "ADJUSTED R^2"

Adding more regressors

- If any new regressor is added to an equation, R^2 *must* _________. Why?
- Compare
  $y = \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + \varepsilon$
  with
  $y = \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + \beta_{K+1} x_{K+1} + \varepsilon$

Effect on R^2 of adding new regressors

- Shorter equation is a *constrained* version of the longer equation, with $\beta_{K+1} = 0$.
- General principle: Constrained minimization will ________ reach as low a value as unconstrained minimization.
- So LS applied to shorter equation will ____ reach as low a sum of squared residuals as LS applied to longer equation.

Effect on R^2 of adding new regressors (cont'd)

- Recall R^2 = 1 – (sum of squared residuals / total sum of squares).
- Since adding more regressors __________ the sum of squared residuals, it must _____________ the R^2 value

When R^2 is not useful

- So R^2 will always ____________ when new regressors are added, even if the new regressors are not really relevant.
- Conclusion: R^2 is _________ useful for comparing the fit of equations of different length.
- Reason: R^2 tends to favor ____________ equation, even if it contains irrelevant regressors.

Theil's adjusted R^2

- To level the playing field between long and short equations, Henri Theil proposed this alternative measure:
- $Adjusted\ R^2 = 1 - \frac{\frac{1}{n-K}\sum \hat{\varepsilon}_i^2}{\frac{1}{n-1}\sum (y_i - \bar{y})^2}$
- Sometimes abbreviated as $\bar{R}^2$.

Effect of adding new regressors on Theil's adjusted R^2

- Adding new regressors
  - raises K (number of βs including intercept), which lowers (n-K), which raises the whole second term, which *lowers* adjusted R^2.
  - but also lowers the sum of squared residuals, which *raises* adjusted R^2.
- So Theil's adjusted R^2 can either ___________ or __________ if new regressors are added.

"R^2" AND "ADJUSTED R^2"

Interpreting Theil's adjusted R^2

- We can write Theil's measure as:
  $Adjusted\ R^2 = 1 - \frac{\frac{1}{n-K}\sum \hat{\varepsilon}_i^2}{\frac{1}{n-1}\sum(y_i-\bar{y})^2} = 1 - \frac{\hat{\sigma}^2}{\widehat{Var}(Y_i)}$
- Numerator of second term is unbiased estimator of variance of error term.
- Denominator is unbiased estimator of variance of Y_i .
- So Theil's adjusted R^2 = 1 – (estimated variance of error term / total estimated variance of y)

Ordinary R^2 versus Theil's adjusted R^2

- We can rewrite Theil's adjusted R^2 as:
- $Adjusted\ R^2 = 1 - \left(\frac{n-1}{n-K}\right)\frac{\sum \hat{\varepsilon}_i^2}{\sum(y_i-\bar{y})^2}$
  Expression in parentheses >1 and does not appear in the definition of ordinary R^2.
- So Theil's adjusted R^2 is always __________ than ordinary R^2 and can be negative.

Conclusions

- R^2 measures the fraction of the variation in y that is explained by the regressors.
- R^2 must always lie between ________ and ________, when the line is fitted by LS.
- But R^2 always _________ when more regressors are added, even if irrelevant.
- Theil's adjusted R^2 does ______ always rise when more regressors are added.

FUNDAMENTAL ASSUMPTIONS AND RESULTING LS PROPERTIES

FUNDAMENTAL ASSUMPTIONS AND RESULTING LS PROPERTIES

- What basic statistical assumptions do we need to justify least-squares estimation of the multiple-regression model?

The linear regression model with multiple regressors

- y is influenced by (K-1) observed regressor variables and an unobserved error term (ε) according to the true or population relationship:
  $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{iK} + \varepsilon_i$.
- Here K = number of βs, or the number of regressors plus 1.

Fundamental assumptions

- Assumption #1:
  Mean of error term is zero: $E(\varepsilon_i) = 0$.
- Assumption #2:
  All regressors are uncorrelated with the error term: $E(\varepsilon_i | x_{i1}, x_{i2}, \ldots, x_{iK}) = 0$.
  - Implies $E(\varepsilon_i x_{ij})$ and $Cov(\varepsilon_i x_{ij}) = 0$, for $j=1,\ldots,K$.

Recall the "Method of Moments" principle

- Set the moments (means and covariances) of the observations equal to the formulas for the theoretical moments.
- Solve for estimators of the parameters of interest (here, the βs).

"Method of Moments" estimators for β_1 through β_k

- By assumption #1, $E(\varepsilon_i) = 0$, so set

$$0 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left[\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}\right]\right)$$

"Method of Moments" estimators for the linear regression model (cont'd)

- By assumption #2, $E(\varepsilon_i x_{ij})$ for $j=1,\ldots,K$, so set

$$0 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_{ij}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left[\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}\right]\right)x_{ij}$$

FUNDAMENTAL ASSUMPTIONS AND RESULTING LS PROPERTIES

"Method of Moments" estimators for the linear regression model (cont'd)

- Combining these results,

$$0 = \sum \left(y_i - [\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}] \right)$$
$$0 = \sum \left(y_i - [\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}] \right) x_{i2}$$
$$\vdots$$
$$0 = \sum \left(y_i - [\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}] \right) x_{iK}$$

Method of moments = least squares for the linear regression model

- These "method-of-moments" equations are ____________________ to the FONCs we derived from the least-squares principle.
- Conclusion: Least-squares estimators satisfy the "method of moments" principle.

LS estimators are unbiased

- An *unbiased* estimator has a mean equal to the true population value of the unknown parameter.
- It can be shown that LS estimators are unbiased:

$$E\left(\hat{\beta}_i\right) = \beta_i \,, \qquad \text{for } j = 1, \ldots, K$$

LS estimators are consistent

- A *consistent* estimator's distribution bunches more and more closely around the true population value, as $n \rightarrow \infty$.
- It can be shown that under assumptions #1 and #2 (and additional modest assumptions) the LS estimators in the multiple regression model are consistent:

$$\hat{\beta}_j \xrightarrow{p} \beta_j \,, \qquad \text{for } j = 1, \ldots, K$$

Variance of LS estimators

- Formulas for variance of LS estimators are too complicated to be useful without further assumptions.
- They depend on the variances of all *n* error terms and the *n(n-1)* covariances between all pairs of error terms.

Conclusions

- Fundamental assumptions are:
  - Assumption #1: $E(\varepsilon_i) = 0$.
  - Assumption #2: $E(\varepsilon_i | x_{i1}, x_{i2}, \ldots, x_{iK}) = 0$.
- They imply that LS estimators
  - satisfy ____________________ principle.
  - are ________________.
  - are ________________ (under modest additional assumptions).

ADDITIONAL ASSUMPTIONS AND RESULTING LS PROPERTIES

ADDITIONAL ASSUMPTIONS AND RESULTING LS PROPERTIES

- What additional assumptions do we need to gauge the precision of our LS estimates?

Additional assumptions yield additional properties

- Under these additional assumptions:
  - Assumption #3: homoskedasticity. $Var(\varepsilon_i) = E(\varepsilon_i)^2 = \sigma^2$.
  - Assumption #4: no autocorrelation. $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$, for i not = j.
- The same good properties we showed for two-variable regression also hold for multiple regression.

Variances of LS estimators

- Let R_j^2 denote the R^2 from regressing x_j on all the other regressors and the intercept.
- Example: If we are estimating
  $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3$, then
  R_2^2 = the R^2 from $x_2 = \alpha_1 + \alpha_2 x_3$.
  R_3^2 = the R^2 from $x_3 = \gamma_1 + \gamma_2 x_2$.

Variances of LS estimators (cont'd)

- Example: If we are estimating
  $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, then
  R_2^2 = the R^2 from $x_2 = \alpha_1 + \alpha_2 x_3 + \alpha_3 x_4$.
  R_3^2 = the R^2 from $x_3 = \gamma_1 + \gamma_2 x_2 + \gamma_3 x_4$.
  R_4^2 = the R^2 from $x_4 = \delta_1 + \delta_2 x_2 + \delta_3 x_3$.
- It can be shown that

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)\sum(x_{ij} - \bar{x}_j)^2}$$

Implications of formula for $Var(\hat{\beta}_j)$

- The larger the variance of the error term ($Var(\varepsilon_i) = \sigma^2$), the ______________ the variances of the LS estimators.
- The larger the variation of the xs around their sample means, the _____________ the variances of the LS estimators.

More implications of formula for $Var(\hat{\beta}_j)$

- The larger the sample size, the ___________ the variances of the LS estimators.
- The more correlated x_j is with the other regressors, the higher the R_j^2 value, and the _______________ the variance of the LS estimator for β_j .

ADDITIONAL ASSUMPTIONS AND RESULTING LS PROPERTIES

Estimating the variance of the error term

- The true value of σ^2 is unknown.
- An unbiased estimator of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- Here n = sample size,
  K = number of βs
  = the number of xs plus 1.

Example: two regressors

- Suppose we are estimating a 3-variable regression equation:
  $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.
- Then K = _______.
- So the unbiased estimator of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-3}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}$$

Example: three regressors

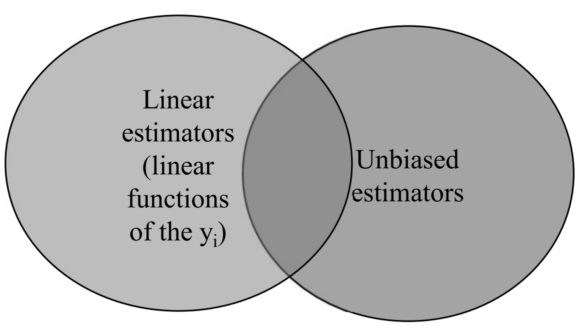
- Suppose we are estimating
  $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$.
- Then K = _______.
- So the unbiased estimator of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-4}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}$$

Standard errors of LS estimators

- Estimates of the standard deviations of the LS estimators are given by substituting $\hat{\sigma}^2$ into the variance formula given above, and taking the square root.
- SEs are automatically reported by regression software programs (including Excel).

Classes of estimators



The Gauss-Markov theorem

- Given assumptions #1 through #4, LS estimators have the lowest variance of all linear unbiased estimators.
- They are the B________ L__________ U___________ E_____________ (BLUE).

ADDITIONAL ASSUMPTIONS AND RESULTING LS PROPERTIES

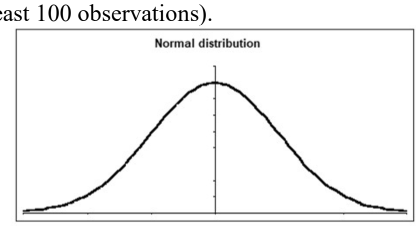
Asymptotically normal

- As the sample size increases, the distribution of the LS estimators approaches a normal distribution around the true values of the coefficients.

$$\hat{\beta}_j \overset{A}{\sim} N\left(\beta_j, SE\left(\hat{\beta}_j\right)^2\right)$$

Why asymptotic normality is important

- This property allows calculation of asymptotic confidence intervals and tests.
- Sample size should be reasonably large (ideally at least 100 observations).



Asymptotic confidence intervals

- We can use the asymptotic normal distribution to calculate confidence intervals for any of the β coefficients.
- 95% confidence interval =
$$\hat{\beta}_j \pm \_\_\_\_\_\_\_\_ \cdot SE\left(\hat{\beta}_j\right)$$
- 90% confidence interval =
$$\hat{\beta}_j \pm \_\_\_\_\_\_\_\_ \cdot SE\left(\hat{\beta}_j\right)$$

Asymptotic tests of individual coefficients

- We can use the asymptotic normal distribution to test hypotheses about the true coefficients.
- Under H_0: $\beta_j = b$, the usual "t-statistic" is asymptotically distributed as standard normal:
$$\frac{\hat{\beta}_j - b}{SE\left(\hat{\beta}_j\right)} \overset{A}{\sim} N(0,1)$$

Computing t-statistics

- In words, the t-statistic = "estimated value minus hypothesized value, divided by standard error."
- t-statistics for the hypothesis that $\beta_j=0$ are automatically computed by Excel and by statistical software programs.
- t-statistics for other hypothesized values of β_j can easily be computed using a calculator or Excel or statistical software.

Computing t-statistics: example

- This model was estimated using 2000 observations on workers:
$$\ln(\text{wage}) = \beta_1 + \beta_2 \text{ schooling} + \beta_3 \text{ work experience} + \beta_4 (\text{work experience})^2 .$$
- Excel output:

	Coefficients	Standard Error	t Stat
Intercept	4.989	0.085	58.890
Schooling	0.099	0.005	18.144
Exper	0.040	0.003	12.143
Expersq	-0.001	0.0001	-11.234

ADDITIONAL ASSUMPTIONS AND RESULTING LS PROPERTIES

Computing t-statistics: example (cont'd)

- To test the null hypothesis that the coefficient of schooling is zero, use the t-statistic given in the output: t = 18.144 .
- Since |t| > 1.96, ____________ the null hypothesis at 5% significance.

Computing t-statistics: example (cont'd)

- To test the null hypothesis that the coefficient of schooling is 0.10, compute this t-statistic:

$$t = \frac{0.099 - 0.10}{0.005} = 9.8$$

- Since |t| < 1.96, ___________________ the null hypothesis at 5% significance.

Prediction

- Suppose we have estimated a linear relationship between x and y using LS.
- Given another value of the x_{n+1} (not in our sample) how can we predict the corresponding value of y_{n+1}?
- LS predictor uses formula for fitted values:

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{2,n+1} + \ldots + \hat{\beta}_K x_{K,n+1}$$

Prediction error

- By contrast, true but unknown value is $y_{n+1} = \beta_1 + \beta_2 x_{2,n+1} + \ldots + \beta_K x_{K,n+1} + \varepsilon_{n+1}$.
- Difference is prediction error:

$$\hat{y}_{n+1} - y_{n+1} = \left(\hat{\beta}_1 - \beta_1\right) + \left(\hat{\beta}_2 - \beta_2\right) x_{2,n+1} + \ldots + \left(\hat{\beta}_K - \beta_K\right) x_{K,n+1} + \varepsilon_{n+1}$$

LS prediction is unbiased

- Note that prediction error results from estimation error and new error term ε_{n+1} .
- But the *expected value* of prediction error is zero.

$$E\left(\hat{y}_{n+1} - y_{n+1}\right) = E\left(\hat{\beta}_1 - \beta_1\right) + E\left(\hat{\beta}_2 - \beta_2\right) x_{2,n+1} + \ldots + E\left(\hat{\beta}_K - \beta_K\right) x_{K,n+1} + E\left(\varepsilon_{n+1}\right)$$

LS predictor is best unbiased

- It can be shown that, given assumptions #1 through #4, the LS predictor has the lowest variance of all linear unbiased predictors.
- LS predictor is B_______ L__________ U____________ P____________ (BLUP).

ADDITIONAL ASSUMPTIONS AND RESULTING LS PROPERTIES

Conclusions

Assuming homoskedasticity and no autocorrelation, LS estimators

- have variances that can be estimated using fairly simple formulas.
- are B______ L________ U___________ E__________ (Gauss-Markov theorem).
- have asymptotically ________________ distributions.

THE NORMALITY ASSUMPTION AND RESULTING LS PROPERTIES

THE NORMALITY ASSUMPTION AND RESULTING LS PROPERTIES

- What final assumption is useful for small samples?

Small samples

- If the sample size is small (say, less than 50) the asymptotic distribution of the LS estimators is not likely to be an accurate approximation.
- But the exact distribution can be derived if we make one more assumption.

Assumption #5: normality

- The error terms follow a normal distribution: $\varepsilon_i \sim N(0, \sigma^2)$.
- This implies that, given the x_i, the y_i also follow a normal distribution:
  $y_i \sim N(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}\ ,\ \sigma^2)$.

Additional assumption yields additional properties

- Error terms ε_i are *independent*, not just uncorrelated.
- LS estimators are "maximum-likelihood" (ML) estimators.
- LS estimators are B______ U_________ E_____________ , not merely BLUE.

Exact distribution of LS estimators

- LS estimators are linear functions of the y_i and (by implication) of the error terms ε_i .
- Thus, LS estimators are exactly normally-distributed, even in small samples.

t-statistics follow *t* distributions (exactly)

- If σ^2 is replaced by its unbiased estimator, the resulting expressions each have a *t* distribution with (n-K) degrees of freedom.

$$\frac{\hat{\beta}_j - \beta_j}{SE\left(\hat{\beta}_j\right)} \sim t(n-K)$$

- Here K = number of βs, or the number of xs plus 1.

THE NORMALITY ASSUMPTION AND RESULTING LS PROPERTIES

Exact confidence intervals

- Use confidence point c from the *t* distribution with (n-K) degrees of freedom, at desired confidence level.

$$\hat{\beta}_j \pm c \cdot SE\left(\hat{\beta}_j\right)$$

- Here, c is the value taken from the *t* table with (n-K) degrees of freedom.

Exact tests using t statistics

- Under H_0: $\beta_j = b$, the usual "t statistic" is exactly distributed as *t*, with (n-K) degrees of freedom:

$$\frac{\hat{\beta}_j - b}{SE\left(\hat{\beta}_j\right)} \sim t_{(n-K)}$$

- Here, critical point is found in *t* table with (n-K) degrees of freedom.

Testing all slope coefficients simultaneously

- Suppose we wish to test the hypothesis that none of the xs have any effect on y.
- Thus in the model
  $y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i$,
  we wish to test H_0: $0 = \beta_2 = \ldots = \beta_K$
  at significance level of, say, 5%.
- Alternative hypothesis is that ___________ slope coefficient is nonzero.

A possible approach

- We could check the t-statistic for every coefficient, and reject H_0 if *any* t-statistic were significant at level 5%.
- However, this test would have a true significance level much greater than 5%.
- Reason: Probability that *at least one* t-statistic is significant when H_0 is true is much greater than 5%.

Another possible approach

- We could check the t-statistic for every coefficient, and reject H_0 if *all* t-statistics were significant at level 5%.
- However, this test would have a true significance level much less than 5%.
- Reason: Probability that *all* t-statistics are significant *at the same time* when H_0 is true is much less than 5%.

The right way to test the joint hypothesis

- Use the following statistic.

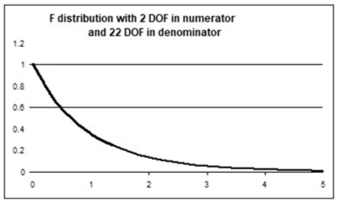
$$F = \frac{\left(\sum\left(\hat{Y}_i - \bar{Y}\right)^2\right) \div (K-1)}{\left(\sum\left(\hat{\varepsilon}_i\right)^2\right) \div (n-K)}$$

- This statistic follows the "F" distribution, with (K-1) DOF in the numerator and (n-K) DOF in the denominator.

THE NORMALITY ASSUMPTION AND RESULTING LS PROPERTIES

"THE" F Statistic

- "THE" F statistic is large if the fitted values vary a lot, compared to the residuals.
- This would occur if the regressors do help to explain y.
- So reject H_0 if F is large.



"THE" F statistic: example

- Suppose the equation
  $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$,
  is estimated on a sample of 25 observations.
- Then (K-1) = DOF in numerator = _____
  and (n-K) = DOF in denominator = _____.
- Suppose THE F statistic = 5.5.
- Critical point at 5% = 3.44, so __________ H_0 and conclude that some xs affect y.

What if errors are not normally distributed?

- THE F-statistic can still be used if the sample size is large.
- Its asymptotic critical points, as the sample size increases without bound, are usually given on the bottom row of the F-table.

Why "THE" F-statistic?

- Other test statistics we will study also happen to follow the F distribution.
- But this particular F statistic is arguably the most important, and it is computed automatically by most regression software.
- So I distinguish it, half in jest, by referring to it as "THE F-statistic."

Conclusions

- If error terms are normally-distributed, LS estimators
  - are B______ U________ E__________.
  - have __________ normal distributions, even in small samples.
- Also, t-statistics are __________.
- Finally, THE ______________ can be used for joint test of all slope coefficients.

PREDICTION AND PREDICTION INTERVALS WITH MULTIPLE REGRESSION

PREDICTION AND PREDICTION INTERVALS WITH MULTIPLE REGRESSION

- How can we compute predictions and prediction intervals with multiple regression.

Conditional prediction

- An important use of LS estimates is "what if?" or *conditional prediction*.
- Given new values of the x variables, we may wish to predict the value of y.
- The LS predictor for y simply substitutes the new values of the x variables into the estimated equation.
- Same formula as for fitted values.

Example 1: a prediction problem

- Suppose we have estimated the following relationship using data on recent college graduates:
  college $GPA_i = \beta_1 + \beta_2\, ACT_i + \beta_3\, HSGPA_i$
- We want to use our results to predict the college success of a high school senior who has not yet attended college.

Example 1: LS predictor

- Given new values of ACT_{n+1} and $HSGPA_{n+1}$, we want to predict college GPA_{n+1} .
- Subscript (n+1) emphasizes these data are for a person *not* in our original sample of n college graduates.
- LS predicted college $\widehat{GPA}_{n+1}$ =
  $$\hat{\beta}_1 + \hat{\beta}_2 ACT_{n+1} + \hat{\beta}_3 HSGPA_{n+1}$$

Definition of LS predictor and prediction error

- General formula for LS predictor:
  $\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1,2} + \hat{\beta}_3 x_{n+1,3} + \ldots + \hat{\beta}_K x_{n+1,K}$
- True value:
  $y_{n+1} = \beta_1 + \beta_2 x_{n+1,2} + \beta_3 x_{n+1,3} + \ldots + \beta_K x_{n+1,K} + \varepsilon_{n+1}$
- LS prediction error:
  $\hat{y}_{n+1} - y_{n+1}$

Sources of prediction error

- As in two-variable regression, LS prediction error in multiple regression results from
  (1) errors in estimating βs.
  (2) the new random error term ε_{n+1}.
- These two sources of error are uncorrelated if the individual we are predicting was not in our estimation sample.

PREDICTION AND PREDICTION INTERVALS WITH MULTIPLE REGRESSION

Sources of prediction error (cont'd)

$$\hat{y}_{n+1} - y_{n+1}$$
$$= \left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1,2} + \ldots + \hat{\beta}_K x_{n+1,K}\right)$$
$$- \left(\beta_1 + \beta_2 x_{n+1,2} + \ldots + \beta_K x_{n+1,K} + \varepsilon_{n+1}\right)$$
$$= \left(\hat{\beta}_1 - \beta_1\right) + \left(\hat{\beta}_2 - \beta_2\right) x_{n+1,2} + \ldots$$
$$+ \left(\hat{\beta}_K - \beta_K\right) x_{n+1,K} - \varepsilon_{n+1}$$

LS prediction is unbiased

- Prediction error is inevitable.
- But under assumptions #1 and #2, the *expected value* of LS prediction error is zero because the LS estimators are unbiased.

$$E\left(\hat{y}_{n+1} - y_{n+1}\right)$$
$$= E\left(\hat{\beta}_1 - \beta_1\right) + E\left(\hat{\beta}_2 - \beta_2\right) x_{n+1,2} + \ldots$$
$$+ E\left(\hat{\beta}_K - \beta_K\right) x_{n+1,K} - E\left(\varepsilon_{n+1}\right)$$

LS predictor is best unbiased

- It can be shown that, under the Gauss-Markov assumptions #1 through #4, the LS predictor has the *lowest* variance of all linear unbiased predictors.
- LS predictor is B_________ L_________ U___________ P____________ (BLUP).

Variance of prediction error

- The general formula for the variance of the prediction error is rather complicated.
- It depends not only on the variances of the LS estimators, but also on their covariances.
- In general, the variance of prediction error is _____________, the smaller σ^2, and the closer the x_{n+1} are to the mean values of the same variables in the estimation sample.

Computing variance of prediction error: a trick

- Advanced software programs can compute the variance of prediction error by a simple command.
- But here is a trick that will coax even the crudest statistical software (like Excel) to compute the variance of prediction error.
  (1) Transform the data.
  (2) Re-estimate equation on transformed data.
  (3) Use re-estimated equation for prediction.

(1) Transform the data

- Subtract the new values of the x's from all the corresponding x's in the original data.
- That is, create new data as follows.

$$\tilde{x}_{i,2} = x_{i,2} -$$
$$\tilde{x}_{i,3} = x_{i,3} -$$
$$\vdots$$
$$\tilde{x}_{i,K} = x_{i,K} -$$

PREDICTION AND PREDICTION INTERVALS WITH MULTIPLE REGRESSION

(2) Re-estimate equation on transformed data

- Estimate
$$y_i = \hat{\tilde{\beta}}_1 + \hat{\beta}_2 \tilde{x}_{i,2} + \hat{\beta}_3 \tilde{x}_{i,3} + \ldots + \hat{\beta}_K \tilde{x}_{i,K} + \varepsilon_i$$
- It can be shown that estimated coefficients will be unchanged, except the intercept.
- What will the new intercept be?

(3) Use re-estimated equation for prediction

- Note that inserting the new values of x makes all the transformed x's zero:
$$\tilde{x}_{n+1,2} = x_{n+1,2} - x_{n+1,2} =$$
$$\tilde{x}_{n+1,3} = x_{n+1,3} - x_{n+1,3} =$$
$$\vdots$$
$$\tilde{x}_{n+1,K} = x_{n+1,K} - x_{n+1,K} =$$

(3) Use re-estimated equation for prediction (cont'd)

- Since all other terms are zero, the new estimated intercept will be the LS predictor:
$$\hat{y}_{n+1} = \hat{\tilde{\beta}}_1 + \hat{\beta}_2 \tilde{x}_{n+1,2} + \hat{\beta}_2 \tilde{x}_{n+1,3} + \ldots + \hat{\beta}_K \tilde{x}_{n+1,K}$$
$$= \tilde{\beta}_1$$

(3) Use re-estimated equation for prediction (cont'd)

- Under assumptions #1 through #4, the variance of the LS prediction is thus
$$Var(\hat{y}_{n+1} - y_{n+1}) = Var(\tilde{\beta}_1) + Var(\varepsilon_{n+1})$$
- This can be estimated as $SE(\tilde{\beta}_1)^2 + \hat{\sigma}^2$, which are automatically computed by almost all software.
- In Excel, $\hat{\sigma}^2$ is reported as "Residual MS."

(3) Use re-estimated equation for prediction (cont'd)

- Taking the square root of the variance of the LS prediction gives:

Standard error of prediction error =
$$\sqrt{\qquad\qquad\qquad}$$

Variance of prediction error in large samples

- For reasonably large samples, SE($\tilde{\beta}_1$)^2 is usually much smaller than $\hat{\sigma}^2$.
- Here's why: Under assumptions #1 through #4, the LS estimators are consistent, so as n increases, SE($\tilde{\beta}_1$)^2 $\rightarrow$ ________.
- But $\hat{\sigma}^2 \rightarrow$ ______, a constant.
- Thus, variance of prediction error $\rightarrow \sigma^2$.

PREDICTION AND PREDICTION INTERVALS WITH MULTIPLE REGRESSION

Exact prediction interval

- Under assumption #5 (error term normally distributed) it can be shown that

$$\frac{\hat{y}_{n+1} - y_{n+1}}{\text{SEPE}} \sim t_{n-K}$$

where "SEPE" is the standard error of prediction error = $\sqrt{SE\left(\tilde{\beta}_1\right)^2 + \hat{\sigma}^2}$

Exact prediction interval (cont'd)

- Use confidence point c from the t-distribution with n-K degrees of freedom, at desired confidence level:

$$\hat{y}_{n+1} \pm c \cdot SEPE$$

Example 1: Transform data

- Suppose we wish to forecast College GPA for a high school senior with ACT_{n+1}=26 and $HSGPA_{n+1}$=3.7 .
- First we transform original data set:

$$A\tilde{C}T_{i,2} = ACT_{i,2} -$$

$$HS\tilde{G}PA_{i,3} = HSGPA_{i,3} -$$

Example 1: Re-estimate equation on transformed data

- Then we re-estimate the equation using the transformed data:

$$\widehat{College\ GPA_i} =$$

$$\hat{\tilde{\beta}}_1 + \hat{\beta}_2 A\tilde{C}T_i + \hat{\beta}_3 HS\tilde{G}PA_i$$

Example 1: Use re-estimated equation for prediction

- Suppose our new estimate of the transformed intercept is $\hat{\tilde{\beta}}_1 = 3.4$
- Thus the LS prediction for a high school senior with ACT=26 and HSGPA=3.7 is college GPA = __________.
- This will be ______________ to the prediction using the original equation.

Example 1: Use re-estimated equation for prediction (cont'd)

- Suppose our SE for the transformed intercept is $SE(\tilde{\beta}_1) = 0.1$ and our estimate of the variance of the error term is $\hat{\sigma}^2 = 0.03$
- Then the standard error of prediction error (SEPE) is

$$\sqrt{SE\left(\tilde{\beta}_1\right)^2 + \hat{\sigma}^2} = \sqrt{0.1^2 + 0.03} = \_\_\_\_$$

PREDICTION AND PREDICTION INTERVALS WITH MULTIPLE REGRESSION

Example 1: prediction interval

- Suppose there are n=400 individuals in our estimation sample. So DOF= ________ and we can use the bottom row of the t-table (DOF=∞) or standard normal table.
- For a 95% prediction interval, c=1.96 from the table.
- The prediction interval is 3.4 ± 1.96 SEPE = 3.4 ± 0.392 = (________,________).

Conclusions

- The LS predictor for y_{n+1} uses the formula for fitted values, applied to the new x's.
- Under assumptions #1 and #2, it is unbiased.
- Under assumptions #1 through #4, it is the B______ L_______ U_______ P__________.
- The standard error of prediction error can easily be computed by first transforming the data.
- Under assumption #5, the prediction interval can be computed using the points from a t-table.

THE ANALYSIS-OF-VARIANCE (ANOVA) TABLE

THE ANALYSIS-OF-VARIANCE (ANOVA) TABLE

- What do the numbers in the ANOVA table mean?

Using the sum-of-squares decomposition

"Residual sum of squares" or "error sum of squares" $\sum \hat{\varepsilon}_i^2$

+ "explained sum of squares" or "regression sum of squares" $+\sum \left[\hat{y}_i - \bar{y}\right]^2$

= "total sum of squares."

Reporting the sums-of-squares decomposition

- Most computer programs for least squares report the sums-of-squares ("variance").
- Also report additional information that can be used to compute many statistics.
- Numbers are formatted as an "analysis of variance" (ANOVA) table.
- Arrangement of rows and columns differs across computer programs.

Example of ANOVA table from Microsoft Excel

- SS = "sums of squares"
- df = "degrees of freedom"
- MS = "mean squares"

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

The "sums of squares" column

$\sum (\hat{y}_i - \bar{y})^2$ $\sum \hat{\varepsilon}_i^2$ $\sum (y_i - \bar{y})^2$

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

The "degrees of freedom" column

K-1 n-K n-1

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

THE ANALYSIS-OF-VARIANCE (ANOVA) TABLE

Interpreting the "degrees of freedom" column

- Number of observations = n = _____
- Number of βs = K = _____

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

The "mean squares" column

$$\frac{\sum(\hat{y}_i - \bar{y})^2}{K-1} \qquad \frac{\sum \hat{\varepsilon}_i^2}{n-K} \qquad \frac{\sum(y_i - \bar{y})^2}{n-1}$$

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

The unbiased estimate of the variance of ε

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-K} =$$

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

Ordinary R^2

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \_\_\_\_\_\_$$

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

Theil's adjusted R^2

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-K}\sum \hat{\varepsilon}_i^2}{\frac{1}{n-1}\sum(y_i - \bar{y})^2} = 1 - \_\_\_\_\_\_$$

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

"THE" F statistic

$$F = \frac{\frac{1}{K-1}\sum(\hat{y}_i - \bar{y})^2}{\frac{1}{n-K}\sum \hat{\varepsilon}_i^2} = \_\_\_\_\_\_$$

ANOVA	df	SS	MS
Regression	2	31.40	15.70
Residual	531	117.04	0.22
Total	533	148.44	

THE ANALYSIS-OF-VARIANCE (ANOVA) TABLE

Conclusions

- The ANOVA table has eight entries reporting
  - SS = ________________________
  - df = _________________________
  - MS = ________________________
- Using these numbers, one can compute the estimated variance of ε, R^2, Theil's adjusted R^2, and THE F statistic.

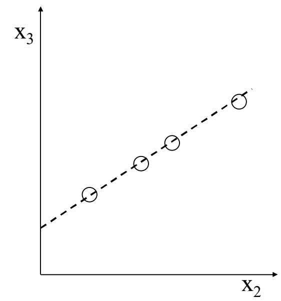
MULTICOLLINEARITY

MULTICOLLINEARITY

- What happens if regressors are perfectly or almost perfectly correlated with each other?

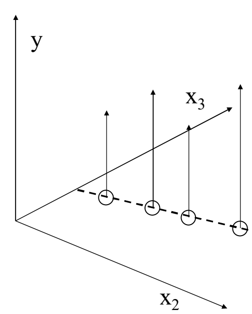
Perfect multicollinearity: definition

- One regressor is related to one or more other regressors according to a *line*:
- $x_2 = \alpha_1 + \alpha_2 x_3$
- $x_2 = \alpha_1 + \alpha_2 x_3 + \alpha_3 x_4$
- etc.
- In this graph, x_2 and x_3 are *perfectly collinear.*



Geometry of perfect multicollinearity

- One regressor is perfectly correlated with one or more others.
- In three-variable case, all observations lie in a vertical plane.



Typical causes of perfect multicollinearity

(1) Some regressor never varies in our dataset.

- Example: We estimate a demand function regressing quantity (y) on price (x), but all observations have the same price.

(2) Regressors that are supposed to be distinct are related by definition.

Example of perfect multicollinearity

- Suppose we estimate a production function, relating output to workers and machines.
- But it turns out that in our dataset each machine is always operated by two workers: workers = 2 x machines.

Another example of perfect multicollinearity

- Suppose we estimate a human capital equation, relating workers' pay to education, age, and work experience.
- But it turns out that work experience is not observed directly in our dataset. It is imputed as follows:
experience = age - education - 6.

MULTICOLLINEARITY

Least squares with multicollinearity

- One of the normal equations is now redundant, a combination of the others.
- We have only (K-1) distinct equations in K unknown βs.

$$0 = \sum \left(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}]\right)$$
$$0 = \sum \left(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}]\right) x_{i2}$$
$$0 = \sum \left(y_i - [\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}]\right) x_{i3}$$

Consequences of perfect multicollinearity

- LS coefficient estimates ____________ be computed for collinear regressors.
- They are mathematically undefined.
- However, if one of the collinear regressors is dropped from the equation, the coefficients of the non-collinear variables ________ still be computed.

What will statistical software do if there are collinear regressors?

- Most statistical software detects the problem and drops one regressor.
- Excel cannot detect the problem and tries to estimate all the coefficients. However,
  - standard errors are huge.
  - coefficient estimates are often far from reasonable.

Is LS the wrong estimation method?

- Not necessarily.
- If two regressors always move together, the "experiment" is badly designed.
- No sensible estimation method—certainly not LAD or reverse LS—could separate the effects of the two regressors using only the information in the sample.

How can we fix perfect multicollinearity? Three ways

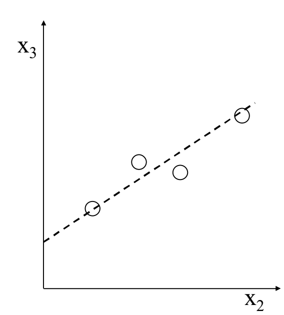
(1) Give up on measuring the effects of collinear variables. Drop one from the equation.

(2) Get more and better data, where the regressors are not perfectly correlated.

(3) Impose restrictions on the coefficients, such as that they sum to a constant.

Imperfect multicollinearity: definition

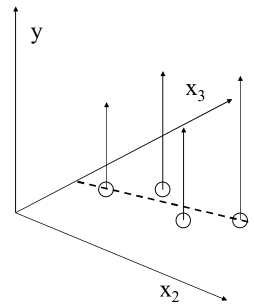
- One regressor is *approximately* related to one or more other regressors according to a line:
- $x_2 = \alpha_1 + \alpha_2 x_3$ + small error
- Here, x_2 and x_3 are *almost collinear.*



MULTICOLLINEARITY

Geometry of imperfect multicollinearity

- One regressor is closely correlated with one or more others.
- In three-variable case, all observations lie close to a vertical plane.



Typical causes of imperfect multicollinearity

(1) Some regressor hardly varies in our dataset.

- Example: We estimate a demand function regressing quantity (y) on price (x), but most observations happen to have the same price.

(2) Certain regressors tend to move together.

Example of imperfect multicollinearity

- Suppose we estimate an aggregate production function for the U.S., relating output to the labor force (x_2) and the capital stock (x_3), using time-series data.
- Unfortunately, the labor force and the capital stock have both grown steadily over time, so x_2 and x_3 are (imperfectly) collinear.

Consequences of imperfect multicollinearity

- Coefficients for collinear regressors cannot be precisely estimated.
- Estimates may even have wrong sign.
- Estimates sensitive to slight changes in data.
- Standard errors are __________ and t-statistics are __________.

Are the LS estimates "wrong"?

- Imperfect multicollinearity is _________ a violation of the classical assumptions.
- Classical properties ______________.
- Coefficient estimates of collinear regressors are simply imprecise (and the standard errors clearly warn us so).

Why are LS estimates imprecise?

- Recall $Var(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2)\ \sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$, where R_j^2 denotes the R^2 from regressing x_j on all the other regressors and the intercept.
- If x_j is closely correlated with other regressors, then R_j^2 is close to one and $\mathrm{Var}(\hat{\beta}_j)$ explodes.
- So $1/(1-R_j^2)$ is sometimes called the *variance inflation factor* for coefficient $\hat{\beta}_j$.

MULTICOLLINEARITY

How can we fix imperfect multicollinearity? Three ways

(1) Ignore it if the collinear variables are not the focus of the investigation.

(2) If they are the focus, get more and better data, where the regressors vary more and are not closely correlated.

- Example: Instead of time-series, use cross-section data on countries to estimate aggregate production function.

How can we fix imperfect multicollinearity? Three ways (cont'd)

(3) Consider whether any restrictions implied by economic theory can be assumed and imposed.

Conclusions

- Perfect multicollinearity means regressors are perfectly correlated with each other.
  - LS estimates of the coefficients of collinear regressors _______________ be computed.
- Imperfect multicollinearity means regressors are closely correlated.
  - Classical properties of LS still hold but coefficient estimates are ________________.

TESTING HYPOTHESES ABOUT COEFFICIENTS

TESTING HYPOTHESES ABOUT COEFFICIENTS

- How can we test hypotheses involving multiple slope coefficients?

Testing a restriction on a single coefficient

- We have already discussed how to test hypotheses like
  - H_0: $\hat{\beta}_2 = 0$ against H_1: $\hat{\beta}_2 \neq 0$
  - H_0: $\hat{\beta}_2 = 0.10$ against H_1: $\hat{\beta}_2 \neq 0.10$

  using t-statistics.
- Each null hypothesis is a ______________ on the value of a coefficient.

Testing a single restriction involving multiple coefficients

- The same procedure can be used to test a restriction on multiple coefficients, such as
  (1) H_0: $\hat{\beta}_2 = \hat{\beta}_3$ against H_1: $\hat{\beta}_2 \neq \hat{\beta}_3$
  (2) H_0: $\hat{\beta}_2 + \hat{\beta}_3 = 1$ against H_1: $\hat{\beta}_2 + \hat{\beta}_3 \neq 1$
- In principle, these could be tested with t-statistics like
  $\frac{\hat{\beta}_2 - \hat{\beta}_3}{SE(\hat{\beta}_2 - \hat{\beta}_3)}$ or $\frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{SE(\hat{\beta}_2 + \hat{\beta}_3)}$

An easier alternative

- However, computing $SE(\hat{\beta}_2 - \hat{\beta}_3)$ or $SE(\hat{\beta}_2 + \hat{\beta}_3)$ is somewhat involved.
- An easier way is to compute an F-statistic for the restriction.
- It requires estimating the equation with and without the restriction.

Estimating an equation with and without a restriction: example (1)

- Unrestricted equation (H_1: $\hat{\beta}_2 \neq \hat{\beta}_3$):
  consumption = $\beta_1 + \beta_2$ labor income
  + β_3 capital income + ε.
- Restricted equation (H_0: $\hat{\beta}_2 = \hat{\beta}_3$):
  earnings = β_1
  + β_2 (labor income + capital income) + ε.

Estimating an equation with and without a restriction: example (2)

- Unrestricted equation (H_1: $\hat{\beta}_2 + \hat{\beta}_3 \neq 1$):
  ln(output) = $\beta_1 + \beta_2$ ln(labor)
  + β_3 ln(capital) + ε.
- Restricted equation (H_0: $\hat{\beta}_2 + \hat{\beta}_3 = 1$):
  ln(output) = β_1 + (1-β_3) ln(labor)
  + β_3 ln(capital) + ε OR
  [ln(output)-ln(labor)] = β_1
  + β_3 [ln(capital)- ln(labor)] + ε

TESTING HYPOTHESES ABOUT COEFFICIENTS

Comparing the results

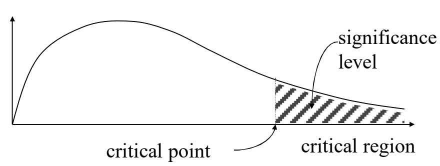
- Save the sums of squared residuals from the unrestricted and restricted regressions (USSR and RSSR).
- Note that necessarily, RSSR ___ USSR.
- If *much* greater, then the restrictions made the fit much worse, and the restrictions should be _______________.
- But how much greater?

Perform the F test of a single restriction

- Compute the statistic as $F = \frac{(RSSR-USS\ \ )}{\frac{1}{n-K}\ USSR}$ where K = number of βs in unrestricted equation (K=3 in both examples).
- Find critical point in a table of the F distribution.
- Here, DOF in numerator = 1 and DOF in denominator = n-K.
- Statistical software will often compute the critical point and/or the p-value automatically.

Perform the F test of a single restriction (cont'd)

- Reject the restriction if F statistic is greater than the critical point—that is, if the restriction increases the sum of squared residuals a lot.



Does it matter whether we use a t test or an F test on a single restriction?

- Not for a two-sided t test (as in these examples).
- It can be shown that the F test statistic of a single restriction equals the square of the t test statistic.
- Similarly, the critical point for the F test equals the square of the critical point for the t test.

Testing multiple restrictions on coefficients

- Sometimes we may wish to test multiple restrictions jointly.
- Example: we may wish to test whether a group of regressors has any effect on y.
- How can we test this type of hypothesis?
- Two test procedures are in wide use:
  - general F-test.
  - LM test.

Example (3)

- Suppose we estimate an equation intended to explain weekly earnings as a function of education, work experience, usual weekly hours, and job risk:

  earnings = $\beta_1 + \beta_2$ educ + β_3 exper + β_4 hours + β_5 risk + ε.

TESTING HYPOTHESES ABOUT COEFFICIENTS

How can we test coefficients jointly?

- Suppose we wish to test whether the human capital variables (educ and exper) are *jointly* significant?
  - H_0: $0 = \beta_2 = \beta_3$.
  - H_1: Either $\beta_2 \neq 0$ or $\beta_3 \neq 0$ (or both).
- Note that the null hypothesis is actually ___ restrictions.

A possible approach

- We could check the t-statistic for every coefficient, and reject H_0 if *either* t-statistic were significant at level 5%.
- However, this test would have significance level much greater than 5%.
- Reason: Probability that *at least one* t-statistic is significant when H_0 is true is much greater than 5%.

Another possible approach

- We could check the t-statistic for every coefficient, and accept H_0 if *both* t-statistics were insignificant at level 5%.
- However, this test would have a true significance level much less than 5%.
- Reason: Probability that *both* t-statistics are significant *at the same time* when H_0 is true is much less than 5%.

The right way to test the joint hypothesis

- Estimate the equation with and without the regressors in question.
- Unrestricted equation:
  earnings = $\beta_1 + \beta_2$ educ + β_3 exper + β_4 hours + β_5 risk + ε.
- Restricted equation (assumes β_2=β_3=____):
  earnings = $\beta_1 + \beta_4$ hours + β_5 risk + ε.

Comparing the results

- Save the sums of squared residuals from the unrestricted and restricted regressions (USSR and RSSR).
- Note that necessarily, RSSR ___ USSR.
- If *much* greater, then the restrictions made the fit much worse, and the restrictions should be ________________.
- But how much greater?

Perform the general F test of multiple restrictions

- Compute the statistic as $F = \dfrac{\frac{1}{r}(RSSR - USSR)}{\frac{1}{n-K}\, USSR}$
- r = number of restrictions (coefficients set to zero in the restricted regression).
  - In example (3), r = ______.
- K = number of βs in the unrestricted regression.
  - In example (3), K = ______.

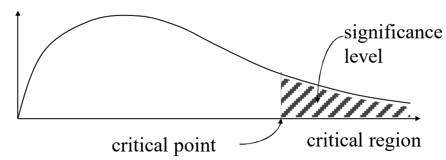
TESTING HYPOTHESES ABOUT COEFFICIENTS

Perform the general F test of multiple restrictions (cont'd)

- Find critical point in a table of the F distribution.
- Here, DOF in numerator = r and DOF in denominator = n-K.
- Statistical software will often compute the critical point and/or the p-value automatically.

Perform the general F test of multiple restrictions (cont'd)

- Again, reject the restriction if F statistic is greater than the critical point—that is, if the restriction increases the sum of squared residuals a lot.



Must the error terms be normally-distributed for the F-test to be valid?

- The F-test is an exact test (i.e., the critical points are exactly as shown in the F-table) if the errors are normally-distributed.
- However, if the sample size is large, the F-test is asympotically valid, even if the error terms are not normally-distributed.

Testing multiple restrictions on coefficients: an alternative test

- The Lagrange multiplier (LM) test has recently become popular in econometrics.
- The name comes from constrained optimization in the context of maximum-likelihood estimation.
  - Recall that if the error term is normally-distributed, LS is maximum-likelihood.
- But this test is also asymptotically valid even if the error terms are not normally-distributed.

Compute the LM test statistic

- Estimate the restricted regression and save the residuals.
- Estimate an *auxiliary regression*:
  - Dependent variable = restricted residuals.
  - Regressors = all regressors, including the excluded regressors.
- Compute ($n\,R_A^2$), where R_A^2 is computed from the auxiliary regression.

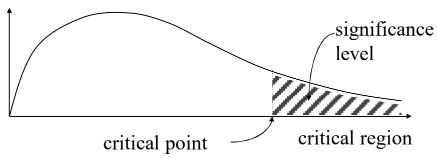
What is an "auxiliary regression"?

- An auxiliary regression is a regression used only to compute a test statistic.
- It has ______ substantive meaning.
- The coefficients do _______ correspond to parameters of any substantive model.
- But here, if they are statistically different from zero, then we can reject the restrictions (H_0).

TESTING HYPOTHESES ABOUT COEFFICIENTS

Perform the LM test of multiple restrictions

- It can be shown that under H_0, ($n R_A^2$) is distributed as chi-square with r degrees of freedom.
- So reject H_0 if ($n R_A^2$) is ___________ than the critical point.



Why the LM test works

- By the algebraic properties of LS, residuals from the restricted regression must be uncorrelated with the included regressors.
- Under H_0, the residuals should also be uncorrelated with these r excluded regressors.
- So under H_0, ($n R^2$) from the auxiliary regression should be _______________.

Applying LM test to example (3)

- In this example, restricted regression is: earnings = $\beta_1 + \beta_4$ hours + β_5 risk + ε.
- Auxiliary regression is: $\hat{\varepsilon} = \alpha_1 + \alpha_2$ educ + α_3 exper + α_4 hours + α_5 risk + ν.
- Here, ν is a new error term.

Applying LM test to example (3) (cont'd)

- Under H_0, ($n R^2$) from the auxiliary regression is distributed as chi-square with DOF=______. We reject H_0
  - if ($n R^2$) is __________ than the critical point.
  - or equivalently if P-value is ______________ than desired significance.

Conclusions

- To test a restriction involving multiple coefficients, we can use a t-test or an F-test.
- To test multiple restrictions, we can use an F-test or an LM test.
- The F-test compares the sum of squared residuals without and without the restrictions, and rejects H_0 if the change in fit is ___________.
- The LM test regresses the restricted residuals on all the regressors, and rejects H_0 if ($n R^2$) for this auxiliary regression is __________.

ALTERNATIVE FUNCTIONAL FORMS

ALTERNATIVE FUNCTIONAL FORMS

- How can multiple linear regression be used to fit nonlinear relationships?

Nonlinear relationships

- It is unrealistic to assume that all relationships are linear.
- Examples of nonlinear relationships:
  - production with diminishing returns.
  - demand with constant elasticities (not constant slopes).
  - U-shaped average cost.

A trick

- Nonlinear relationships can be fitted by transforming the variables before estimation by ordinary least squares.
- For two-variable regression, we considered reciprocals and logarithms.
- Multiple regression offers additional possibilities.

Popular transformations

(1) quadratic form.
(2) cubic form.
(3) interaction effects.

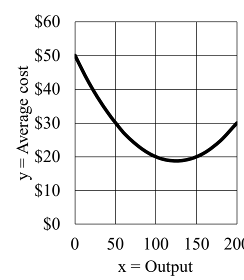
Quadratic form

- Form: $y = \beta_1 + \beta_2 x + \beta_3 (x^2)$.
- Effect of x: $dy/dx = \beta_2 + 2 \beta_3 x$.
- Effect increases if β_3 ___ 0, decreases if β_3____ 0.

Example of quadratic form

U-shaped average cost curve

Avg cost = 50
- 0.50 output
+ 0.002 output^2.



ALTERNATIVE FUNCTIONAL FORMS

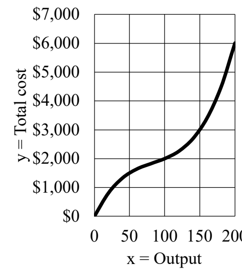
Cubic form

- Form: $y = \beta_1 + \beta_2 x + \beta_3(x^2) + \beta_4(x^3)$.
- Effect of x:
  $dy/dx = \beta_2 + 2\,\beta_3\,x + 3\,\beta_4\,x^2$.
- Effect of x first decreases and then increases if β_3 ____ 0 and β_4 _____ 0 .

Example of cubic form

Total cost function

Total cost = 50 output
– 0.50 output^2
+ 0.002 output^3 .



Interaction effects

- Form: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4(x_2 x_3)$.
- Effect of x_2: $dy/dx_2 = \beta_2 + \beta_4\,x_3$.
- Effect of x_3: $dy/dx_3 = \beta_3 + \beta_4\,x_2$.
- Effect of each regressor depends on the value of the other.

Example of interaction effects

Human capital equation

hourly wage = –8.9 + 1.8 educ + 0.5 exper
+ 0.04 (educ * exper) .

- d pay / d educ = 1.8 + 0.04 exper.
- So for someone with *exper*=10 years, an extra year of *educ* increases pay by 1.8 + 0.04(10) = $________.

Example of interaction effects (cont'd)

Human capital equation

- Also, d wage / d exper = 0.5 + 0.04 educ.
- So for someone with *educ*=16 years, an extra year of *exper* increases pay by 0.5 + 0.04(16) = $_________.

Example of interaction effects (cont'd)

Human capital equation

- Also, d wage / d exper = 0.5 + 0.04 educ.
- So for someone with *educ*=16 years, an extra year of *exper* increases pay by 0.5 + 0.04(16) = $___1.14___.

ALTERNATIVE FUNCTIONAL FORMS

Conclusions

- Nonlinear relationships can be fitted by ______________________ the variables before estimating by ordinary least squares.
- Popular transformations include reciprocals, logarithms, polynomials, and interaction effects.

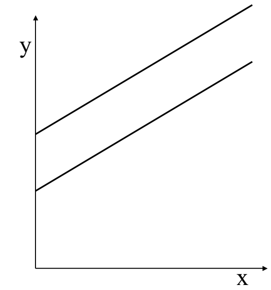
DUMMY (OR BINARY) VARIABLES

DUMMY (OR BINARY) VARIABLES

- How can we allow for a change in the intercept?

Change in intercept

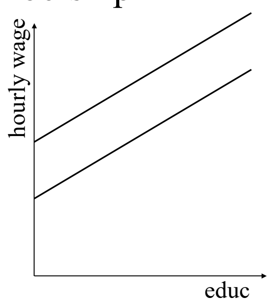
- Suppose we believe that the intercept in our equation is different for part of the sample.



Example: human capital and union membership

hourly wage
$= \beta_1 + \beta_2$ educ

- However, β_1 might be higher for members of labor unions.



Defining a dummy variable

- Define $d_i = 1$ for all workers who are members of labor unions.
- Define $d_i = 0$ for all workers who are not members of unions.
- d_i is thus a zero-one variable, called a "binary variable" or a "dummy variable."

Creating a dummy variable

Name	**Union member?**	**d_i**
J. Smith	yes	
K. Jones	no	
L. Ramirez	no	
M. Huang	yes	
etc.		

Including the dummy variable

- Now estimate
  hourly wage $= \beta_1 + \beta_2$ educ $+ \beta_3$ d .
- The coefficient β_3 measures the difference in the intercept between union and nonunion workers.
- Difference can be positive or negative.

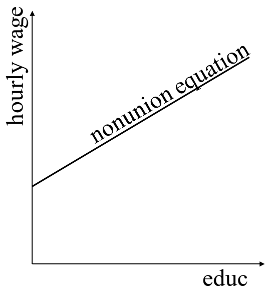
DUMMY (OR BINARY) VARIABLES

When $d_i = 0$

- For nonunion workers, $d_i = 0$, so:

hourly wage $= \beta_1 + \beta_2$ educ $+ \beta_3 0$,

- intercept $= \beta_1$
- slope $= \beta_2$.

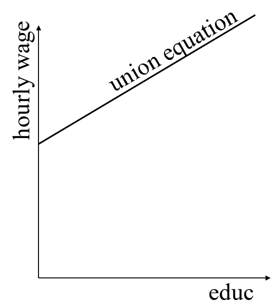


When $d_i = 1$

- For union workers, $d_i = 1$, so:

hourly wage $= \beta_1 + \beta_2$ educ $+ \beta_3 1$,

- intercept $= \beta_1 + \beta_3$
- slope $= \beta_2$.



Computing intercepts: numerical example

- Suppose hourly wage = 2.1 + 0.7 educ + 1.7 d where d = 1 for union members, = 0 for workers not members of a union.
- Nonunion intercept = ________ .
- Union intercept = 2.1 + 1.7 = ________ .
- Slope for both = ________ .
- Interpretation: union workers earn $___________ more per hour than nonunion workers with same education, on average.

Testing for different intercept

- To test whether the two groups have a different intercept, just use the t-test on β_3.
  - H_0: Groups have same intercept. $\beta_3 = 0$.
  - H_1: Groups have different intercepts. $\beta_3 \neq 0$.

Two ways to define the dummy variable

- We could have defined $d_i = 1$ for all workers who are *not* members of labor unions.
- LS estimate of β_3 would have been same magnitude but opposite sign.
- SE for β_3 would have been the same.
- t statistic would be same magnitude but opposite sign.
- Estimated intercepts of two groups would be unchanged.

More than two groups

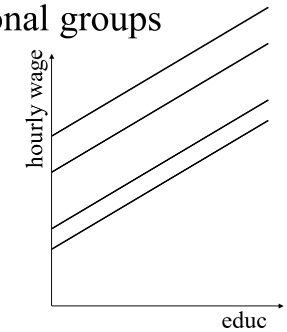
- Suppose we believe that the intercept varies across more than two groups.
- Then we need to create more dummy variables.
- For m groups, we need m-1 dummy variables.

DUMMY (OR BINARY) VARIABLES

Example: human capital and occupational groups

hourly wage
$= \beta_1 + \beta_2$ educ

- However, β_1 might be different for:
- professional workers
- managerial workers
- clerical workers
- blue-collar workers



Defining dummy variables

- Define $dprof_i = 1$ for professional workers and = 0 for all others.
- Define $dman_i = 1$ for managerial workers and = 0 for all others.
- Define $dcler_i = 1$ for clerical workers and = 0 for all others.
- No dummy variable for blue-collar workers.

Including the dummy variables

- Now estimate
  hourly wage = $\beta_1 + \beta_2$ educ + β_3 dprof + β_4 dman + β_5 dcler .
- The coefficients β_3 , β_4 , and β_5 measure the difference in the intercept between these workers and blue-collar workers.
- Differences can be positive or negative.

Each group now has own intercept

- Professional workers' intercept = _______.
- Managerial workers' intercept = ________.
- Clerical workers' intercept =___________.
- Blue-collar workers' intercept = ________.

Why only three dummy variables for four groups

- Suppose we defined a 4th dummy variable
  - $dblue_i = 1$ for blue-collar workers and = 0 for all others.
- Since every worker is a member of one and only one group,
  $dprof_i + dman_i + dcler_i + dblue_i = 1$.
- Including $dblue_i$ in the regression would cause perfect ______________________.

One reference group and m-1 dummy variables

- For m groups, we have one "reference group" or "base group" and m-1 dummy variables.
- Here, blue-collar workers are the reference group.
- The three coefficients β_3, β_4, and β_5 measure differences from the reference group.

DUMMY (OR BINARY) VARIABLES

Testing for different intercepts

- To test whether the four groups have a different intercepts, just use an F-test on β_3, β_4, and β_5 .
- H_0: Groups have same intercept. $\beta_3 = \beta_4 = \beta_5 = 0$.
- H_1: Groups have different intercepts. $\beta_3 \neq 0$, or $\beta_4 \neq 0$, or $\beta_5 \neq 0$.

Applying the F-test

- Unrestricted equation: hourly wage = $\beta_1 + \beta_2$ educ + β_3 dprof + β_4 dman + β_5 dcler .
- Restricted equation (assumes $0=\beta_3=\beta_4=\beta_5$): hourly wage = $\beta_1 + \beta_2$ educ.
- Here, r = number of restrictions = ____.
- K = number of βs in unrestricted equation = ____.

Formula for the F-statistic

- More generally, r = number of dummy variables = number of groups minus one.
- K = number of βs, including dummy coefficients.

$$F = \frac{\frac{1}{r}\ (RSSR - USSR)}{\frac{1}{n-K}\ \ USSR}$$

F-test: numerical example

- Suppose previous equation estimated on 100 observations: n=100.
- Sum of squared residuals with dummy variables (unrestricted) = 487.2
- Sum of squared residuals without dummy variables (restricted) = 743.1.
- K = 1 intercept + 1 slope coefficient + 3 dummy variables = _______.

F-test: numerical example (cont'd)

- r = three dummy coefficients = 3.

$$F = \frac{\frac{1}{3}\left(743.1-487.2\right)}{\frac{1}{100-5}\left(487.2\right)} = 16.6$$

- Critical point for F(3,95) at 1% significance is about 4.01. Easily ____________ null hypothesis of common intercept.

F-test: numerical example (cont'd)

- r = three dummy coefficients = 3.

$$F = \frac{\frac{1}{3}\left(743.1-487.2\right)}{\frac{1}{100-5}\left(487.2\right)} = 16.6$$

- Critical point for F(3,95) at 1% significance is about 4.01. Easily __REJECT__ null hypothesis of common intercept.

DUMMY (OR BINARY) VARIABLES

Many ways to define the dummy variables

- Any group can be the reference group.
- So choose the reference group to make economic interpretation as easy as possible.
- Choice of reference group does *not* change intercept estimates, or the F-test statistic.

Interpreting dummy coefficients when dependent variable is in logs

- Recall: whenever dependent variable is in natural log, slope coefficient = *percent change* in dependent variable from one-unit change in regressor.
- Coefficient of dummy thus shows *percent difference* between groups, holding other regressors constant.

When dependent variable is in logs: numerical example

- Suppose ln(wage) = 0.40 + 0.09 educ + 0.12 d where d = 1 for union members, = 0 for workers not members of a union.
- One more year of education (Δ educ = 1) causes the wage to rise by about _____ percent.
- Union members (d=1) enjoy wages about _______ percent higher than nonunion workers (d=0).

Conclusions

- Differences in the intercept across groups of observations can be permitted by including dummy (zero-one) variables.
- If there are only two groups, just _______ dummy variable is needed.
- If there are m groups, then _________ dummy variables are needed.

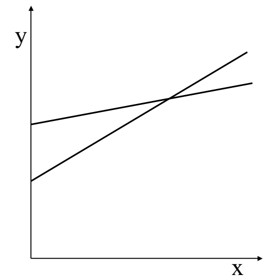
STRUCTURAL CHANGE

STRUCTURAL CHANGE

- How can we allow for a change in both the intercept and the slope?

Change in intercept and slope

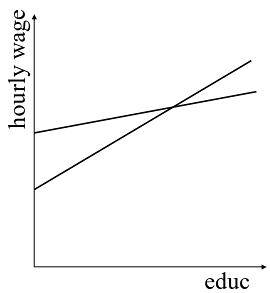
- Suppose we believe that the intercept *and* the slope in our equation are different for part of the sample.



Example: human capital and union membership

hourly wage
$= \beta_1 + \beta_2$ educ

- However, we suspect β_1 might be higher and β_2 might be lower for members of labor unions.



Defining a dummy variable and an interaction

- As before, define $d_i = 1$ for union members and $d_i = 0$ non-union members.
- But also define an interaction variable: ($d_i \times$ educ$_i$). Thus:
  - ($d_i \times$ educ$_i$) = educ$_i$ for union members.
  - ($d_i \times$ educ$_i$) = 0 for non-union members.

Creating an interaction variable

Name	Union dummy	Education	Union × Education
J. Rodriguez	0	14	
S. Aiello	1	16	
J. Wang	0	18	
R. Patel	1	12	
etc.			

Including the dummy variable and the interaction

- Now estimate: hourly wage $= \beta_1 + \beta_2$ educ $+ \beta_3$ d $+ \beta_4$ (d × educ).
- Coefficient β_3 measures the difference in the _______________ between union and nonunion workers.
- Coefficient β_4 measures the difference in the _______________ between union and nonunion workers.

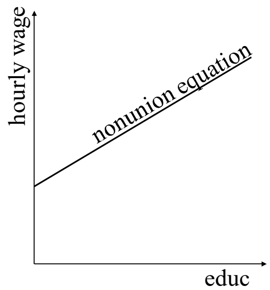
STRUCTURAL CHANGE

When $d_i = 0$

- For nonunion workers, $d_i = 0$, so:

hourly wage = $\beta_1 + \beta_2$ educ + $\beta_3\,0 + \beta_4\,0$,

- intercept = β_1
- slope = β_2 .

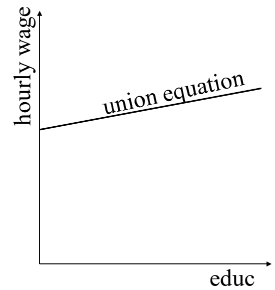


When $d_i = 1$

- For union workers, $d_i = 1$, so:

hourly wage = $\beta_1 + \beta_2$ educ + $\beta_3\,1 + \beta_4$ educ .

- intercept = $\beta_1 + \beta_3$
- slope = $\beta_2 + \beta_4$.



Numerical example

hourly wage = 6.07 + 0.74 educ + 2.01 d – 0.22 (d × educ).

- Nonunion workers (d=0):
  - Intercept = ________.
  - Slope = ________.
- Union workers (d=1):
  - Intercept is 6.07 + 2.01 = ________.
  - Slope = 0.74 – 0.22 = ________.

Testing for different slope

- To test whether the two groups have a different slope, just use the t-test on β_4.
  - H_0: Groups have same slope. $\beta_4 = 0$.
  - H_1: Groups have different slopes. $\beta_4 \neq 0$.

Testing for different intercept or different slope or both

- To test whether the two groups have a different intercepts and/or slope, must test β_3 and β_4 *jointly*.
  - H_0: Groups have same intercept and slope. $\beta_3 = 0$ and $\beta_4 = 0$.
  - H_1: Groups have different intercepts and/or different slopes. $\beta_3 \neq 0$ and/or $\beta_4 \neq 0$

Restricted versus unrestricted equations

- Unrestricted equation:
  hourly wage
  $= \beta_1 + \beta_2$ educ + β_3 d + β_4 (d × educ).
- Restricted equation (assumes $0=\beta_3=\beta_4$):
  hourly wage = $\beta_1 + \beta_2$ educ.
- Here, r = # of restrictions = _____.
- K = # of βs in unrestricted equation = ____.

STRUCTURAL CHANGE

Applying the test

- Here, H_0 means that both groups lie on the ___________ regression line.
- H_1 means they lie on ____________ lines.
- Intuition: If the sum of squared residuals _______________ substantially when we try to fit the same line to both groups, we should reject the null hypothesis.

Compute the F-statistic

- So reject H_0 if F-statistic is sufficiently high.

$$F = \frac{\frac{1}{r}\,(RSSR - USSR)}{\frac{1}{n-K}\ \ USSR}$$

"Chow test"

- Alternative hypothesis H_1 is sometimes called "structural change."
- The F-test for structural change is sometimes called the "Chow test."
- Gregory Chow first used this test in 1960 (but he did not call it a "Chow test").

Gregory Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, Vol. 28 (1960), pp. 591-605.

Two ways to define the dummy variable and interaction

- We could have defined $d_i = 1$ for non-union members.
- LS estimates of β_3 and β_4 would have been same magnitudes but opposite signs.
- SEs for β_3 and β_4 would have been the same.
- So definition does not change results, properly interpreted.

More than two slope coefficients

- Suppose we wish to test for "structural change" in a bigger equation:
- hourly wage = $\beta_1 + \beta_2$ educ + β_3 exper .
- We must define another interaction:
  - $(d_i \times \text{exper}_i) = \text{exper}_i$ for union members.
  - $(d_i \times \text{exper}_i) = 0$ for non-union members.

Restricted versus unrestricted equations

- Unrestricted equation:
  hourly wage = $\beta_1 + \beta_2$ educ + β_3 exper
  + β_4 d + β_5 (d × educ) + β_6 (d × exper).
- Restricted equation (assumes $0=\beta_4=\beta_5=\beta_6$):
  hourly wage = $\beta_1 + \beta_2$ educ + β_3 exper .
- Here, r = number of restrictions = ____.
- K = number of βs in unrestricted equation = ____.

STRUCTURAL CHANGE

Again use F-statistic

- For general Chow test,
  - K = total number of βs, including coefficients of dummies and interactions.
  - r = number of restrictions = K/2.

$$F = \frac{\frac{1}{r}\,(RSSR - USSR)}{\frac{1}{n-K}\;USSR}$$

Conclusions

- Differences in the intercept and slope across groups of observations can be permitted by including dummy (zero-one) variables and __________________.
- The dummy variable should be interacted with every variable whose coefficient is thought to differ across groups.
- Test for differences using an ___________.

INFLUENTIAL OBSERVATIONS

INFLUENTIAL OBSERVATIONS

- What are “influential observations”?
- Why do they merit attention?

Influential observations

- While ordinary least squares uses all of the data, some observations have more influence on the estimates than others.
- Outlier = observation whose y-value is far from the fitted line.
- High leverage point = observation whose x-values are far from the rest.

How to find high leverage points in multiple regression?

- Before computing LS, *always* compute descriptive statistics of x variables—mean, standard deviation, minimum and maximum.
- Do a box plot of each x variable.
- Print the five largest and five smallest values of each x variable.
- But these methods might not work because leverage depends on *combinations* of xs.

Formal definition of leverage

- It can be shown (using matrix algebra) that each LS fitted value $\hat{y}_i$ is a linear function of all the actual y_is:
  $\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n$
  where each h_{ij} depends on all the xs.
- Then h_{ii} is called the *leverage* of the ith observation.
- h_{ii} can easily be computed by statistical software.

Properties of leverage

- It can be shown that necessarily
  $\frac{1}{n} \leq h_{ii} \leq 1$ and $\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{K}{n}$
  where K = total number of βs, including the intercept.
- Conventionally, an observation is a called a *high leverage point*, if $h_{ii} >$ _______.

How to find outliers in multiple regression?

- After computing LS, do a box plot of LS residuals.
- Print the five largest and five smallest values of the residuals.

INFLUENTIAL OBSERVATIONS

Leave-one-out measures of influence

- An obvious way of finding influential observations is to _________________ an observation from the data and recompute LS estimates and/or fitted values.
- If they change a lot, the observation is influential.
- Most statistical software can easily do this for all n observations.

Measuring changes in fitted values: Cook's distance

- How much would the LS fitted values change if hypothetically an observation were *left out* of the data?
- Let $\hat{y}_{j(i)}$ denote the fitted or predicted value for observation *j* using LS estimates that *leave out* observation *i*.
- Observation i is influential if all the $\hat{y}_{j(i)}$ are far from the usual LS fitted values $\hat{y}_j$.

Formal definition of Cook's distance

- $D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{K\ \hat{\sigma}_i^2}$, where K = total number of βs, including the intercept.
- A typical value of D_i is about (______). A much higher value indicates an influential observation.
- Note that we would expect D_i to decrease as the sample size n increases, for then any individual observation should have less and less influence.

What to do about influential observations?

- Why do influential observations occur?
  - Possibly data error.
  - Possibly observation does not belong in sample.
  - Possibly just random variation.
- What to do?
  - Check for data errors.
  - Check whether observation does not belong in sample.

Conclusions

- Influential observations have greater influence on regression results than other observations.
- *Leverage* and *Cook's distance* are measures for finding influential observations in multiple regression.
- Influential observations can occur because of _________________ or because an observation does ________ belong in the sample.

SELECTION OF REGRESSORS FOR PREDICTION

SELECTION OF REGRESSORS FOR PREDICTION

- How can we find the right regressors if our purpose is prediction?

If our purpose is prediction...

- We want a model that "explains" the y_i well.
- Our model should produce predicted values $\hat{y}_i$ close to the actual values y_i.
- Adding more regressors always improves the "fit," _________ R^2 and _________ $\hat{\sigma}^2$.

Measures of prediction success for linear models (higher is better)

- $R^2 = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum (y_i - \bar{y})^2}$
- $Adjusted\ R^2 = 1 - \frac{\frac{1}{n-K}\sum \hat{\varepsilon}_i^2}{\frac{1}{n-1}\sum (y_i - \bar{y})^2}$

$$= 1 - \frac{\hat{\sigma}^2}{\widehat{Var(Y_i)}}$$

where K = number of βs including intercept.

Measures of prediction success for linear models (lower is better)

- $AIC = Akaike\ Information\ Criterion$
$$= \text{n}\ln(2\pi) + n + n\ ln\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right) + 2(K+1)$$
- $BIC = Bayesian\ Information\ Criterion$
$$= \text{n}\ln(2\pi) + n + n\ ln\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right) + (K+1)\ln(n)$$

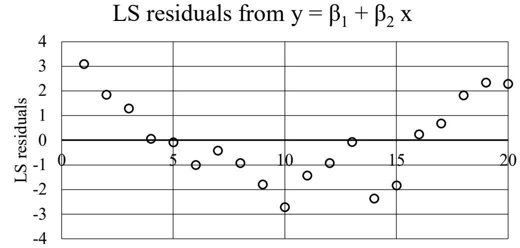
where K = number of βs including intercept.

How to choose regressors to improve prediction

A. Analysis of residuals.
- Plot residuals against included regressors.
- Plot residuals against potential regressors.

B. Automated selection of regressors.
- Stepwise algorithms.
- Best regression search algorithm.

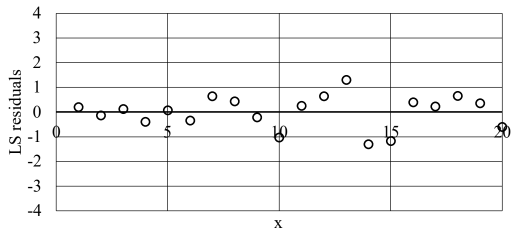
Example of plotting residuals against included regressor: problem?



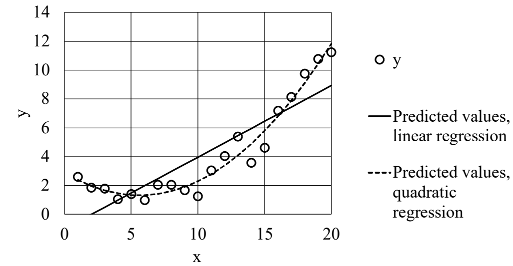
SELECTION OF REGRESSORS FOR PREDICTION

Example of plotting residuals against included regressor: solution?

LS residuals from $y = \beta_1 + \beta_2 x + \beta_3 x^2$



Raw data for example above



Automated selection of regressors

- If there are many potential regressors, there are even more potential models.
- In general, if there are m potential regressors, then there are 2^m potential models (including the model with no regressors and the model will all).
- How to find the best one?

Stepwise algorithms

- Instead of estimating all 2^m potential models, stepwise methods add or eliminate regressors one by one.
- *Forward selection* adds regressors until none of the remaining possibilities make a sufficient contribution to fit.
- *Backward elimination* (aka backward selection) subtracts regressors until all of the remaining regressors are too important to eliminate.

Stepwise algorithms: forward selection

- Estimate the m models with ONE regressor.
- Choose best model, by some criterion (t statistic, adjusted R^2, AIC, etc.).
- Now estimate m-1 models that add a second regressor.
- Choose best model, by some criterion. Repeat!
- Stop when no potential models show substantial improvement, by some predetermined criterion.

Stepwise algorithms: backward elimination

- Begin with ALL potential regressors.
- Estimate m models that drop one regressor.
- Choose best model, by some criterion (t statistic, adjusted R^2, AIC, etc.).
- Now estimate m-1 models that drop a second regressor.
- Choose best model, by some criterion. Repeat!
- Stop when no potential models show substantial improvement, by some predetermined criterion.

SELECTION OF REGRESSORS FOR PREDICTION

Caution about stepwise algorithms

- Sequential tests are not strictly valid.
- With forward selection, early t tests are computed on the *wrong model*—some regressors are missing—which violates LS assumptions.
- With backward selection, later t tests are computed *conditional on* the variable surviving prior t-tests, which means the true size* is actually much larger than 5%.

* Probability of mistakenly rejecting the null hypothesis.

Best regression algorithm

- Given m potential regressors, decide how many regressors to include. Call that number p.
- Estimate all $\binom{m}{p} = \frac{m!}{p!\,(m-p)!}$ potential models with p regressors.
- Choose best model, by some criterion (F statistic, adjusted R^2, AIC, etc.).

Caution about all automated search algorithms

- Search algorithms can "overfit," finding a model that fits the sample extremely well, but predicting poorly out-of-sample.
- Why? Because, for example, using a t test with 5% significance means rejecting the null hypothesis by mistake 1 in 20 times.
- But repeating the t test increases the chance of mistakenly rejecting the null hypothesis.

Model validation by splitting the sample

- Ideal approach, if there is sufficient data, is to divide sample into two:
  - one sample for selection and estimation
  - one sample for prediction.
- Models should be evaluated on how they perform in the prediction sample.

Conclusions

- If our purpose is prediction, we select regressors to improve the fit, as measured by R^2 , etc.
- Plotting residuals against included or omitted regressors can help find useful regressors.
- Automated methods for selection of regressors include forward selection, backward elimination, and overall best regression algorithms.
- However, automated methods can sometimes "over-fit," predicting poorly out-of-sample.

SELECTION OF REGRESSORS FOR CAUSAL INFERENCE

SELECTION OF REGRESSORS FOR CAUSAL INFERENCE

- How can we find the right regressors if our purpose is causal inference?

If our purpose is causal inference...

- We want to measure the causal effect of x on y, *ceteris paribus**.
- That requires measuring what happens to y when x changes, while holding constant all other factors that might influence y.
- We want unbiased estimates of the slope ________________. (R^2 is unimportant.)

* Latin: *other things equal.*

Purpose determines selection of regressors: example 1

- Consider equation: health = $\beta_1 + \beta_2$ schooling, where health = summary measure of health status and schooling = years of schooling.
- Our purpose might be to *predict* health—perhaps to help price health insurance or life insurance.
- In that case, we keep schooling in the equation only if it improves the fit (high t-statistic, low p-value, etc.).

Purpose determines selection of regressors: example 1 (cont'd)

- Alternatively, our purpose might be to measure the *causal effect* of schooling on health—perhaps to measure the benefits of public policy requiring high school students to stay in school longer.
- In that case, we keep schooling in the equation regardless.
- What other variables should we include?

Laboratory data versus observational data

- In some fields, laboratory experiments are used to measure causal effects.
- In a lab, one can *control* the other factors that might influence y. In that case, two variable regression is unbiased.

Laboratory data versus observational data (cont'd)

- Outside a lab, we cannot literally control other factors. We can only *observe* them.
- For example, we cannot control all factors of peoples' lives that might affect their health.
- But sometimes we can statistically control for these other factors with extra regressors.

SELECTION OF REGRESSORS FOR CAUSAL INFERENCE

Example 1: health status and schooling (cont'd)

- Parents' income might positively affect health status, because people with wealthier parents likely enjoyed better health care as children.
- At the same time, people with wealthier parents likely received more schooling.
- If parents' income is omitted from the regression, then the estimated coefficient of schooling will pick up some of the effect of parents' income.

Example 1: control variables

- To avoid omitted variable bias, we instead estimate:
  health = $\beta_1 + \beta_2$ schooling + β_3 parents' income .
- Our focus is on β_2 . Schooling is sometimes called the "treatment variable."
- Parents' income is included *not* to improve the fit, but to insure the estimate of β_2 is *unbiased*.
- Parents' income is called a "control variable."

Example 1: bad controls

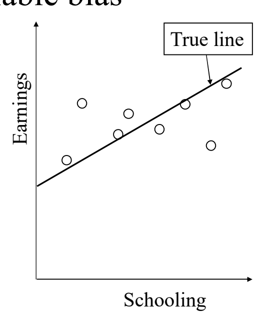
- Measures of the person's healthy habits (diet and exercise) would likely be statistically significant.
- But education improves people's health in part by encouraging healthy habits.
- So including healthy habits as controls would bias down the estimated effect of education on health.
- In general, anything caused by the treatment is a bad control and should not be used.

Example 2

- Suppose our purpose is to measure the causal effect of schooling on earnings.
- We estimate this equation:
  earnings = $\beta_1 + \beta_2$ schooling,
  where schooling = years of schooling.
- But work experience also affects earnings and is negatively correlated with schooling.

Causal inference example 2: omitted variable bias

- Omitting work experience results in omitted variable bias (aka selection bias).
- Estimated coefficient of schooling is biased ________.



Example 2: control variables

- To avoid omitted variable bias, we instead estimate:
  earnings = $\beta_1 + \beta_2$ schooling + β_3 experience.
- Our focus is on β_2 . Schooling is sometimes called the "treatment variable."
- Experience is included *not* to improve the fit, but to insure the estimate of β_2 is *unbiased*.
- Experience is called a "control variable."

SELECTION OF REGRESSORS FOR CAUSAL INFERENCE

Example 2: bad controls

- Occupation dummy variables would also likely be statistically significant.
- But education increases people's earnings in part by giving access to higher-paying occupations.
- So including occupation dummy variables as controls would bias down the estimated effect of education on earnings.
- In general, anything caused by the treatment is a bad control and should not be used.

Conclusions

- If our purpose is causal inference, we select regressors, called controls, to ensure our estimate of the coefficient of the treatment variable is unbiased.
- Good controls have an effect on the dependent variable, are correlated with the treatment variable, but are not themselves caused by the treatment variable.
- Bad controls are caused by the treatment variable.

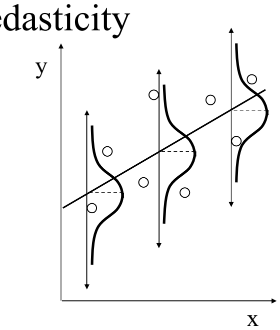
HETEROSKEDASTICITY: DEFINITION AND CONSEQUENCES

HETEROSKEDASTICITY: DEFINITION AND CONSEQUENCES

- What happens to the LS estimators if assumption #3 is violated?

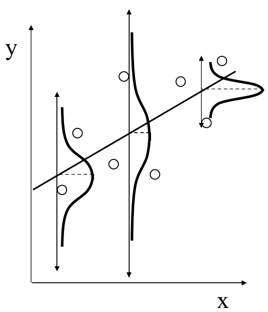
Recall assumption #3: homoskedasticity

- Error term has constant variance.
- Var (ε_i) = σ^2.
- Variance is the ______ for all observations i = 1, …, n.



Definition of heteroskedasticity

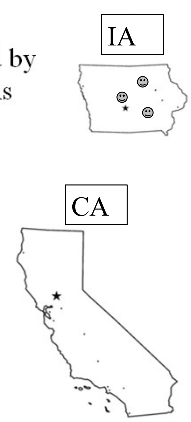
- Error term has changing variance
- Var (ε_i) = σ_i^2.
- Variance is ________ for each observation.



What causes heteroskedasticity?

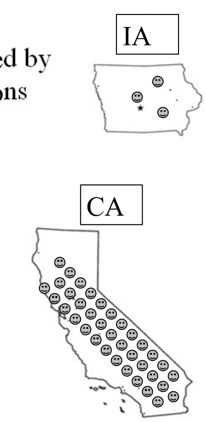
- Error term represents unobserved random variables that affect y.
- If the variance of these unobserved variables is ________ constant, then heteroskedasticity occurs.
- But why would the variance not be constant?

Heteroskedasticity can be caused by differences in size of observations



- Cross-section datasets often contain observations of very different size.
- Example: states of the U.S. are of vastly different size.
- The population of California (CA) is about ______ times population of Iowa (IA).

Heteroskedasticity can be caused by differences in size of observations



- Cross-section datasets often contain observations of very different size.
- Example: states of the U.S. are of vastly different size.
- The population of California (CA) is about __12__ times population of Iowa (IA).

HETEROSKEDASTICITY: DEFINITION AND CONSEQUENCES

Heteroskedasticity related to population: example

- Suppose we estimate a consumption function using 50 observations on states: $cons_i = \beta_1 + \beta_2\, inc_i + \varepsilon_i$.
- $Var(\varepsilon_i)$ can be either proportional or inversely proportional to state population, depending on whether $cons_i$ and inc_i are *totals* or *averages* (per capita).
- Here is why.

Heteroskedasticity when the data are TOTALS

- Suppose $cons_i$ = total consumption and inc_i = total income for the entire population of each state.
- Then ε_i must = ______________ unobserved factors for all people in that state:

$$\varepsilon_i = \sum_{j=1}^{POP_i} a_j$$

where a_j = the unobserved factor for person j in state i, whose total population is POP_i.

Heteroskedasticity when the data are TOTALS (cont'd)

- Suppose $Var(a_j) = \sigma_a^2$, constant, and the a_j are uncorrelated. Then

$$Var(\varepsilon_i) = Var\left(\sum_{j=1}^{POP_i} a_j\right) = \sum_{j=1}^{POP_i} \sigma_a^2 = POP_i\left(\sigma_a^2\right)$$

- So the variance of the error term is not constant. It is ________________________________ to the population of the state.
- In particular, $Var(\varepsilon_{CA})$ = ______ × $Var(\varepsilon_{IA})$.

Heteroskedasticity when the data are AVERAGES

- Alternatively, suppose $cons_i$ = average consumption and inc_i = average income per capita in each state.
- Then ε_i must = ______________ unobserved factor per capita in each state:

$$\varepsilon_i = \left(\frac{1}{POP_i}\right)\sum_{j=1}^{POP_i} a_j$$

where a_j = the unobserved factor for person j in state i, whose total population is POP_i.

Heteroskedasticity when the data are AVERAGES (cont'd)

- Then $Var(\varepsilon_i) = Var\left(\left(\frac{1}{POP_i}\right)\sum_{j=1}^{POP_i} a_j\right) = \left(\frac{1}{POP_i}\right)^2 \sum_{j=1}^{POP_i} \sigma_a^2$

$$= \left(\frac{1}{POP_i}\right)^2 POP_i\left(\sigma_a^2\right) = \frac{\sigma_a^2}{POP_i}$$

- So the variance of the error term is not constant. It is ________________________________ to the population of the state.
- In particular, $Var(\varepsilon_{CA})$ = ________ × $Var(\varepsilon_{IA})$.

What properties still hold when the error term is heteroskedastic?

- LS estimators are still *unbiased*.
- LS estimators are still *consistent* (under modest assumptions).
- LS estimators are still *method-of-moments* estimators.

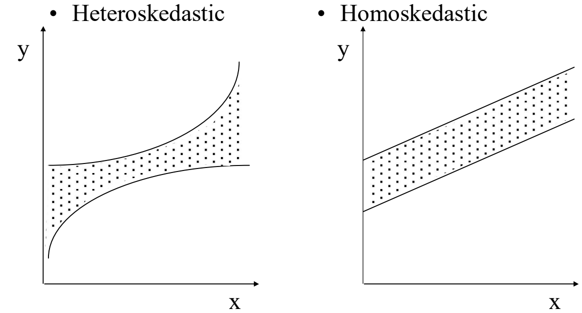
HETEROSKEDASTICITY: DEFINITION AND CONSEQUENCES

What properties no longer hold?

- LS estimators are no longer the B____ L______ U________ E_________ (Gauss-Markov theorem).
- Standard formulas for variance of LS estimators (and standard errors) are no longer correct. Why not?

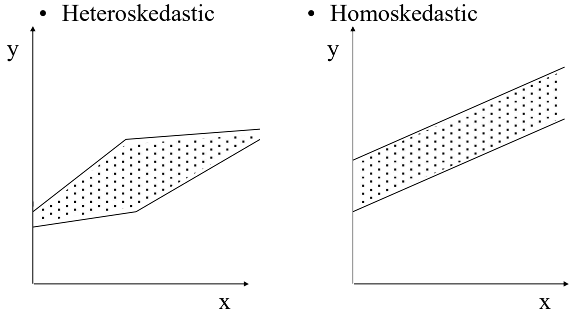
Intuition: Sometimes heteroskedasticity can *increase* the variance of the LS estimators

- Heteroskedastic
- Homoskedastic



Intuition: Sometimes heteroskedasticity can *decrease* the variance of the LS estimators

- Heteroskedastic
- Homoskedastic



Consequences of heteroskedasticity: intuition

- Heteroskedasticity can increase *or* decrease the variance of the LS estimators.
- Samples with the same average error variance might yield more *or* less precise LS estimators for slope and intercept.
- Standard errors computed without recognizing heteroskedasticity can be either too large *or* too small.

Formula for the variance of the LS slope estimator

- Assuming no autocorrelation, we found that the variance of the LS slope estimator was given by the following formula:

$$Var\left(\hat{\beta}_2\right) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 Var\left(\varepsilon_i\right)}{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^2}$$

Implications of homoskedasticity

- Assuming homoskedasticity, the formula could be simplified:

$$Var\left(\hat{\beta}_2\right) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 Var\left(\varepsilon_i\right)}{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^2}$$

HETEROSKEDASTICITY: DEFINITION AND CONSEQUENCES

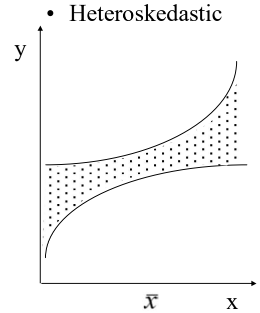
Variance of LS slope estimator with homoskedasticity

…leading to the usual formula:

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

What if the error variance is *larger* when x is far from its mean?

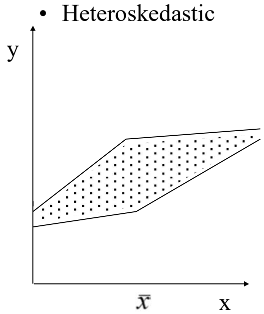
- Heteroskedastic



- Then
$$\sum (x_i - \bar{x})^2 Var(\varepsilon_i) > \sum (x_i - \bar{x})^2 \sigma^2$$
- So usual formula is too small:
$$Var\left(\hat{\beta}_2\right) > \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

What if the error variance is *smaller* when x is far from its mean?

- Heteroskedastic



- Then
$$\sum (x_i - \bar{x})^2 Var(\varepsilon_i) < \sum (x_i - \bar{x})^2 \sigma^2$$
- So usual formula is too large:
$$Var\left(\hat{\beta}_2\right) < \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Consequences of heteroskedasticity: formal results

- Usual standard error formulas are biased.
- If variance is smaller when x is close to its mean, usual formulas are biased down.
  - Standard errors are too ___________.
- If variance is smaller when x is far from its mean, usual formulas are biased up.
  - Standard errors are too ___________.

Other consequences

- Any calculations based on the usual standard errors are also biased.
- Usual formulas for confidence intervals are either too large or too small.
- Test statistics (t-tests, F-tests, etc.) are inaccurate.

Conclusions

- If the error term is heteroskedastic, LS estimators are still _____________ and _______________, and can still be justified by the *method-of-moments* principle.
- However, the usual formulas for __________________, confidence intervals, and test statistics are invalid.

TESTING FOR HETEROSKEDASTICITY

TESTING FOR HETEROSKEDASTICITY

- How can we test for heteroskedasticity?

Testing for heteroskedasticity

- Many tests have been proposed for heteroskedasticity.
- Here we cover the two most popular tests.
  - Breusch-Pagan test, with modification by Koenker.
  - White test.

Breusch, T.S., and A.R. Pagan, "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica*, Vol. 47, (1979), pp. 987-1007.
Koenker, R., "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics*, Vol. 17 (1981), pp.107-112.
White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica,* Vol 48, (1980), pp. 817-838.

Homoskedasticity versus heteroskedasticity

- $y_i = \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + \varepsilon_i$.
- H_0: Homoskedasticity (no heteroskedasticity). $\text{Var}(\varepsilon_i) = \sigma^2$. Variance of error term is ______ for all observations.
- H_1: Heteroskedasticity. $\text{Var}(\varepsilon_i) = \sigma_i^2$. Variance is ______________ for different observations.

BP (Breusch-Pagan) test for heteroskedasticity: motivation

- Suppose we suspect variance of the error term depends on one or more observed variables $z_1, z_2, ... , z_G$.
- Thus, suspect $\text{Var}(\varepsilon_i) = f(z_1,z_2, ... , z_G)$.
- zs can be the regressors (xs) or variables not included in regression, but not y.

BP test for heteroskedasticity: procedure

- Tests for a relationship between variance of error term and the zs.
- Save residuals from ordinary LS regression.
- Square them and use them as dependent variables in an "auxiliary regression":
  $\hat{\varepsilon}^2 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + ... + \alpha_G z_G + \nu$.
- Here, ν is a new error term.

What is an "auxiliary regression"?

- An auxiliary regression is a regression used only to compute a test statistic.
- It has ______ substantive meaning.
- The coefficients do ________ correspond to parameters of any model.
- But here, if they are statistically different from zero, we can reject homoskedasticity.

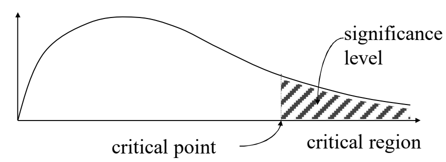
TESTING FOR HETEROSKEDASTICITY

BP test for heteroskedasticity: test statistic

- Could use F-test on auxiliary equation.
- More common to apply so-called LM test.
- Compute $Y = n R_A^2$, where R_A^2 is computed from the auxiliary regression.
- Under H_0, Y asymptotically distributed as chi-square with K_A-1 degrees of freedom, where K_A = number of α's in auxiliary regression.

Perform the BP test for heteroskedasticity

- So reject H_0 if $(n R_A^2)$ is _____________ than the critical point.



BP test for heteroskedasticity: example

- Suppose we are estimating the relationship between traffic accidents and the speed limit, using 50 cross-section observations on states:

  accident rate$_i = \beta_1 + \beta_2$ speed limit$_i + \varepsilon_i$.
- However, we suspect $\text{Var}(\varepsilon_i)$ is not constant, but related to state population and state GDP.

BP test for heteroskedasticity: example (cont'd)

- We estimate the accident equation and save the residuals $\hat{\varepsilon}_i^2$.
- Then we estimate an auxiliary regression:

  $\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 \text{ pop}_i + \alpha_2 \text{ state GDP}_i$

  and find $R_A^2 = 0.13$.
- BP test statistic = $n R_A^2 = 50(0.13)$ = _______.
- 5% critical value for chi-square with 2 degrees of freedom = 5.99.
- So __________ H_0 : homoskedasticity.

BP test for heteroskedasticity: intuition

- The auxiliary regression is intended to test for a relationship between the zs and the variance of each error term σ_i^2.
- We do not observe σ_i^2 so we use the squared residuals $\hat{\varepsilon}_i^2$ instead.
- R_A^2 measures the strength of this relationship. Reject H_0 (homoskedasticity) if R_A^2 is sufficiently large.

But why do we care about heteroskedasticity? What problems does it cause?

(1) Standard errors, t-tests, and F-tests are invalid.

- Reason: Usual formulas assume variance of error term σ_i^2 is unrelated to the x_is.

(2) LS estimators for coefficients are still unbiased and consistent, but not as precise as they could be: they are not B_____ L_______U_________E_______.

TESTING FOR HETEROSKEDASTICITY

White test for heteroskedasticity: motivation

- White (1980) proposed test that focused on first problem: invalid standard errors and tests caused by a relationship between variance of error term σ_i^2 and the x_is.

White test for heteroskedasticity: motivation

- Tests for relationship between variance of the error term term σ_i^2 and the xs (and their squares and interactions).
- White test can be viewed as special case of BP test (though proposed independently).
- White test statistic is also computed with an auxiliary regression.

White test for heteroskedasticity: procedure

- Save residuals from original ordinary LS regression.
- Square them and use them as dependent variables in an auxiliary regression.
- Regressors in auxiliary regression are original regressors, their squares, and their interactions.

White test for heteroskedasticity: example

- Suppose original equation is:
  $\text{quantity}_i = \beta_1 + \beta_2 \,\text{price}_i + \beta_3 \,\text{income}_i + \varepsilon_i$
- Then White's auxiliary equation would be:
  $\hat{\varepsilon}_i^2 = \alpha_1 + \alpha_2 \,\text{price}_i + \alpha_3 \,\text{income}_i$
  $+ \alpha_4 \,\text{price}_i^2 + \alpha_5 \,\text{income}_i^2$
  $+ \alpha_6 \,(\text{price}_i \times \text{income}_i) + \nu_i$.

White test for heteroskedasticity: test statistic

- Could use F-test on auxiliary equation.
- More common to apply so-called LM test.
- Compute $Y = n\, R_A^2$, where R_A^2 is computed from the auxiliary regression.
- Under H_0, Y asymptotically distributed as chi-square with K_A-1 degrees of freedom, where K_A = number of α's in auxiliary regression.

White test for heteroskedasticity: potential practical issue

- If original equation has many regressors, White's auxiliary equation will have a *huge number* of regressors—perhaps too many to be estimated.
- For example, suppose original regression has 10 regressors.
- Auxiliary regression uses same 10 regressors, their 10 squares, and 9+8+7+6+5+4+3+2+1=______ interactions, a total of G = ______ regressors plus a constant term!
- Auxiliary regression requires at least ______ observations just to estimate!

TESTING FOR HETEROSKEDASTICITY

Conclusion

- For heteroskedasticity related to several variables (zs) the BP test can be applied.
  - Save residuals from original regression.
  - In an auxiliary regression, regress squared residuals on z's.
  - Test statistic is (_______).
- White test is similar, but zs are the original xs, their squares, and interactions.

CORRECTING FOR HETEROSKEDASTICITY

CORRECTING FOR HETEROSKEDASTICITY

- How can we modify the regression procedure to correct for heteroskedasticity?

Why is heteroskedasticity a problem?

- Standard errors, t-tests, and F-tests are invalid.
- LS estimators for coefficients are still unbiased and consistent, but not as precise as they could be: not B____ L_______U_______E__________.

Two approaches to correcting heteroskedasticity

1) *Robust inference* corrects standard errors and test statistics so they are still valid in presence of heteroskedasticity.
2) *Weighted least squares* re-estimates the whole equation so coefficient estimates are BLUE *and* standard errors are correct. More powerful but requires more information.

1) Robust inference

- White (1980) derived formulas for standard errors that are valid under homoskedasticity or heteroskedasticity.
- Formulas are asymptotic—only valid for large samples.
- White's formulas are available in most statistical software (but not Excel).

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica,* Vol 48, (1980), pp. 817-838.

White's standard error for two-variable regression

- Recall usual formula, valid with no heteroskedasticity: $SE(\hat{\beta}_2) = \sqrt{\dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}$
- White's formula, asymptotically valid with or without heteroskedasticity: $\begin{matrix} Robust \\ SE(\hat{\beta}_2) \end{matrix} = \sqrt{\dfrac{\sum (x_i - \bar{x})^2 \hat{\varepsilon}_i^2}{\left(\sum (x_i - \bar{x})^2\right)^2}}$

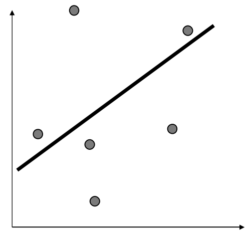
2) Weighted least squares (WLS)

- Suppose we know the pattern of the heteroskedasticity.
- Var (ε_i) = α z_i, where α = unknown constant, and z_i is observed variable.
- Using this information, we can *weight* the data before applying least squares, and thereby restore Gauss-Markov assumptions.

CORRECTING FOR HETEROSKEDASTICITY

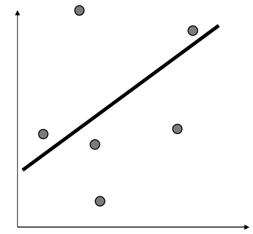
Weighted least squares: intuitive motivation

- Observations with lower variance are closer to true population regression line, on average.
- These observations should be more heavily weighted in computing estimates.



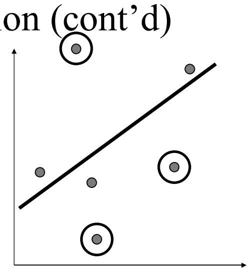
Weighted least squares: intuitive motivation (cont'd)

- Observations with higher variance are farther from true population regression line, on average.
- These observations should be discounted in computing estimates.



Weighted least squares: intuitive motivation (cont'd)

- Observations with higher variance are farther from true population regression line, on average.
- These observations should be discounted in computing estimates.



Weighted least squares: transforming the equation

- $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$.
- Assume $\mathrm{Var}(\varepsilon_i) = \alpha z_i$, where α = unknown constant, and z_i is observed variable.
- So multiply original equation by $(1/z_i^{1/2})$ to get _____________ equation:
  $(y_i/z_i^{1/2}) = \beta_1(1/z_i^{1/2}) + \beta_2(x_{i2}/z_i^{1/2})$
  $+ \beta_3(x_{i3}/z_i^{1/2}) + \nu_i$, where $\nu_i = \varepsilon_i/z_i^{1/2}$.

Weighted least squares: why heteroskedasticity is eliminated

- Formally, by definition, $\nu_i = \varepsilon_i/z_i^{1/2} = (1/z_i^{1/2})\, \varepsilon_i$.
- So $\mathrm{Var}(\nu_i) = (1/z_i)\, \mathrm{Var}(\varepsilon_i) = (1/z_i)\, \alpha z_i = \alpha$, constant.*
- Heteroskedasticity is eliminated!
- Intuitively, WLS "discounts" observations with high variance (high z_i).

* Recall that $\mathrm{Var}(aX) = a^2\, \mathrm{Var}(X)$.

Weighted least squares: choosing z_i

- If y_i is a *total* variable (total consumption, total output, total crimes, etc.) then z_i = population is usually a good choice.
- If y_i is an *average* variable (consumption per capita, output per capita, crimes per capita, etc.) then z_i = (1/population) is usually a good choice.
- Can check choice of z_i using BP test.

CORRECTING FOR HETEROSKEDASTICITY

Weighted least squares: using software

- If using Excel, must transform the data by hand, dividing each data column by $z_i^{1/2}$.
- In other software, an option for WLS is available—usually "weight = *variable.*"
- *variable* should be inversely proportional to the variance: *variable* = $1/z_i$.
- Software then multiplies data by *variable*$^{1/2}$.

Weighted least squares: interpreting results

- Assuming the original equation is correctly specified and $\mathrm{Var}(\varepsilon_i) = \alpha\, z_i$, then transformed equation is homoskedastic.
- So WLS yields
  - BLUE estimates of coefficients in original equation.
  - valid standard errors, t-tests, and F-tests of multiple coefficients.

Conclusions

- *White's robust standard error formulas* for ordinary LS are valid even in presence of heteroskedasticity of unknown form, but the coefficient estimates are not ___________.
- *Weighted Least Squares* coefficient estimates are BLUE and standard errors are valid, but WLS requires knowledge of the variable driving heteroskedasticity (z_i).

PART 4

Univariate Time Series Models

TIME SERIES DATA AND MODELS

TIME SERIES DATA AND MODELS

- What is different about time-series data and models?
- What is a "white noise" process?

Time-series datasets

- Same individual (person, firm, country) is observed repeatedly over time.
- Frequency might be weekly, monthly, quarterly, or annual.

Obs. #	Year	Unempl. rate	Inflation (CPI)	RGDP g.r. per capita
1	2000	4.0	3.4	2.5
2	2001	4.7	2.8	-0.6
3	2002	5.8	1.6	1.1

Why time-series data are different from cross-section data

- Observations have a natural ordering.
- Direction of causality: past can influence future, but future cannot influence past.
- Generally, cannot be viewed as a random sample.
- Instead, best viewed as a *stochastic* (i.e., random) *process:* variables evolving in a random way from some initial values.

Notation: subscripts

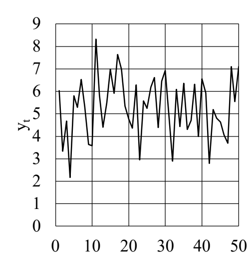
- It is conventional to index time-series observations by t (instead of by i).
- $t = 1$ for the first observation.
- $t = T$ for the last observation (instead of n).
- Observed variable is y_t .
- Unobserved error is ε_t .

White noise

- The simplest process is simply an independent, identically distributed (IID) random variable.
- Mean and variance are constant.
- Sometimes called "white noise" in a time-series context.

Plot of a white noise process

- Plot should show no particular trend or pattern.
- Mean could be different from zero.
- In this example, mean = 5.



TIME SERIES DATA AND MODELS

Forecasting: definitions

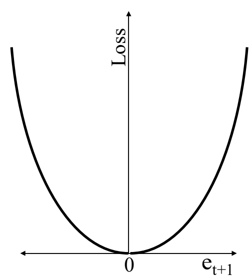
- Suppose at time *t* we wish to forecast a particular variable y (e.g., GDP, interest rate, inflation rate, etc.) one period into the future: y_{t+1}.
- This is a _______-*step ahead* forecast.
- We will use all information available to us at time *t*, including the current value y_t: this is called the _______________ *set* at time *t*, denoted I_t .

Forecast errors

- Let f_t denote the one-step ahead forecast—that is, the forecast of y_{t-1} at time t.
- Forecasts are almost never perfect.
- Define the *forecast error* = $e_{t+1} = y_{t+1} - f_t$.

Loss from errors

- The choice of forecasting method depends on the criterion or *loss function*.
- Most popular loss function is the *quadratic* (or squared-error) *loss function*: loss = $\alpha\ (e_{t+1})^2$.



Properties of quadratic loss function

- Symmetric: the loss from e_{t+1}= -10 is same as the loss from e_{t+1}= +10.
- Quadratic:
  - loss from $e_{t+1} = \pm 2$ is _____ times the loss from $e_{t+1} = \pm 1$.
  - loss from $e_{t+1} = \pm 3$ is _____ times the loss from $e_{t+1} = \pm 1$.

Implications of quadratic loss function

- Of course, e_{t+1} is not known in advance, so it must be treated as random.
- So we must choose a forecasting method that minimizes *expected* squared error, conditional on the information set:
  $$E(e_{t+1}^2 \mid I_t) = E((y_{t+1} - f_t)^2 \mid I_t).$$
- From probability theory we know that expected squared error is minimized if the forecast is chosen to be the *conditional mean*:
  $$f_t = E(y_{t+1} \mid I_t).$$

Forecasting time series

- In this section of the course our sole purpose is forecasting—that is, prediction in a time-series context.
- With a model like $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$, the LS predictor of y_{T+1} is $\hat{y}_{T+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{T+1}$.
- However, this requires us to first predict x_{T+1} , which is often difficult.
  - Exceptions: when x_t represents a time trend or a seasonal dummy variable.

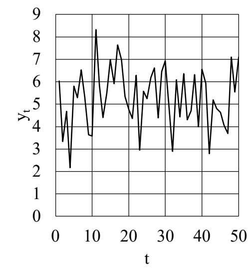
TIME SERIES DATA AND MODELS

Forecasting time series using own past values

- So we will explore methods of forecasting y_t from its own past values.
- That is, we wish to forecast y_{T+1}, y_{T+2}, y_{T+3}, etc. where the information set is $y_1, \ldots, y_T$.
- No xs will be used, except possibly time trends or seasonal dummy variables.

White noise process: example

Suppose our model is $y_t = \beta_1 + \varepsilon_t$, where ε_t is not observed but is assumed to be IID with mean zero and constant variance σ^2.



White noise process: estimation

These assumptions imply

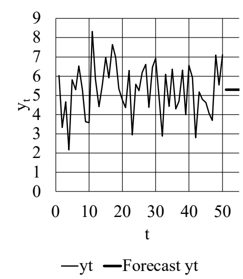
- The BLUE estimator of β_1 is just the sample mean of y_t, that is, $\hat{\beta}_1 = \frac{1}{T}\sum_{t=1}^{T} y_t$.
- The unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(y_t - \hat{\beta}_1\right)^2$$

Here, $\hat{\beta}_1 = 5.29$ and $\hat{\sigma}^2 = 1.86$.

White noise process: point forecasts

- Because ε_t is a white-noise process, our forecast for $\hat{y}_{T+1}$, $\hat{y}_{T+2}$, etc., is just the sample mean, $\hat{\beta}_1 = 5.29$.



White noise process: forecast error

- Important to report size of likely forecast error.
- Forecast error $= y_T - \hat{y}_{T+1}$
  $= (\beta_1 + \varepsilon_{T+1}) - \hat{\beta}_1$
  $= \left(\beta_1 - \hat{\beta}_1\right) + \varepsilon_{T+1}$
- Expected value of forecast error is zero because $\hat{\beta}_1$ is unbiased and ε_t has mean zero.

White noise process: forecast error variance

- Forecast error variance $= \mathrm{Var}(\hat{\beta}_1) + \mathrm{Var}(\varepsilon_{T+1})$.
- Note ________ sources of forecast error variance: (1) error in estimating $\hat{\beta}_1$ and (2) brand-new error term ε_{T+1} .
- $\mathrm{Var}(\hat{\beta}_1)$ shrinks as sample size increases, but $\mathrm{Var}(\varepsilon_{T+1})$ does not.
- There is _______ covariance because $\hat{\beta}_1$ was computed from $y_1,\ldots,y_T$, which are uncorrelated with ε_{T+1} since ε_t are assumed IID.

TIME SERIES DATA AND MODELS

White noise process: standard error of forecast

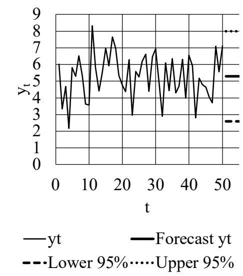
- Here, estimated forecast error variance = $\text{Var}(\hat{\beta}_1) + \hat{\sigma}^2$= 0.037 + 1.86 = __________.
- The standard error of the forecast is simply the square root :
  SE of forecast = $\sqrt{1.897}$ = __________.

White noise process: forecast interval

- Assume ε_t is normally-distributed.
- Then the 95% forecast interval is

  $= \hat{\beta}_1 \pm 1.96\ SE$
  $= 5.29 \pm 1.96(1.377)$

  __________________.



Conclusions

- Time series data have a natural ordering, a time direction of causality, and should be viewed as a _______________ process, not a random sample.
- In this section of the course, we explore how to forecast a time series using only its own past values.
- The simplest stochastic process is ____________, where observations on y_t are IID.

Conclusions

- Time series data have a natural ordering, a time direction of causality, and should be viewed as a <u>stochastic</u> process, not a random sample.
- In this section of the course, we explore how to forecast a time series using only its own past values.
- The simplest stochastic process is <u>white noise</u>, where observations on y_t are IID.

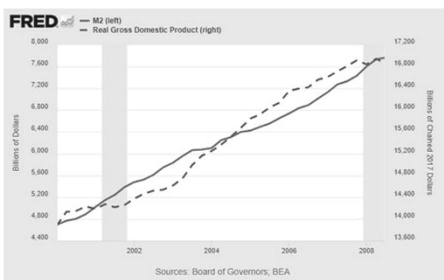
TIME TRENDS

TIME TRENDS

- What is a time trend?
- What is the difference between a linear trend and an exponential trend?

Most series cannot plausibly be modeled as white noise

Most series cannot plausibly be modeled as white noise

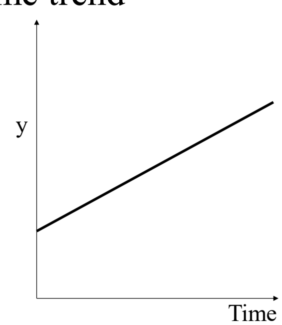


How to model a trended series?

- The simplest trended process adds a time trend regressor: $y_t = \beta_1 + \beta_2 t + \varepsilon_t$, where ε_t is not observed but is assumed to be IID with mean zero and constant variance σ_ε^2.
- Depending on the form of the dependent variable, the time trend may be called either "linear" or "exponential."

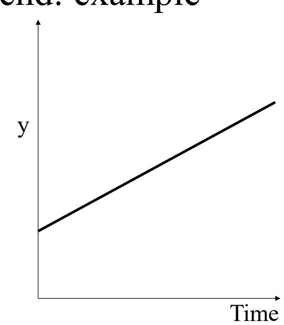
Linear time trend

- A linear time trend occurs if the dependent variable is in levels: $y_t = \beta_1 + \beta_2 t + \varepsilon_t$.
- Then y_t increases by β_2 units from one time period to the next (if the error term does not change).



Linear time trend: example

- Suppose we have estimated the following: $y_t = 27.3 + 3.4\,t$.
- Then y_t increases by ______ units from one time period to the next (if the error term does not change).



TIME TRENDS

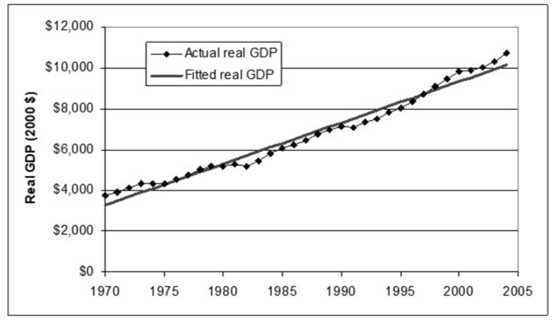
Fitting a linear trend to US real GDP

- Using data from 1970 to 2004, the following estimates were obtained (with standard errors in parentheses).

$$\text{Real GDP} = \underset{(113.8)}{3085.7} + \underset{(5.5)}{201.6}\ \text{trend}$$

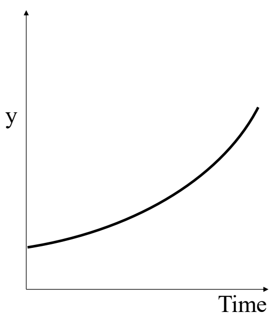
- Thus real GDP increased every year by about $_________ billion, on average.

Fitting a linear trend to US real GDP: actual and fitted values



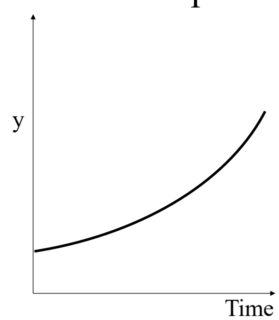
Exponential time trend

- An exponential time trend occurs if the dependent variable is in logarithms.
- $\ln(y)_t = \beta_1 + \beta_2 t + \varepsilon_t$, or $y_t = \exp(\beta_1+\beta_2 t+\varepsilon_t)$.
- Then y_t increases by $100\beta_2$ *percent* from one observation to the next.



Exponential time trend: example

- Suppose we have estimated the following: $\ln(y_t) = 5.3 + 0.07\,t$.
- Then y_t increases by _____________ from one observation to the next (if the error term does not change).



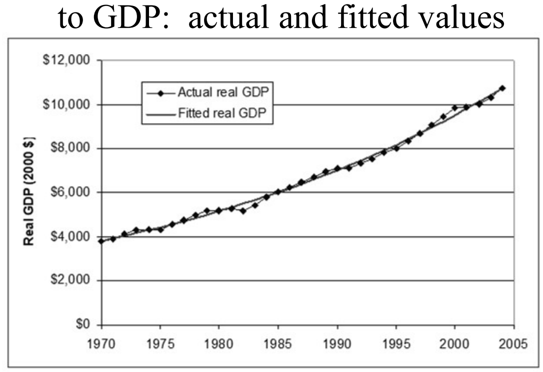
Fitting an exponential trend to US real GDP

- Using same data from 1970 to 2004, the following estimates were obtained (with standard errors in parentheses).

$$\ln(\text{real GDP}) = \underset{(0.0074)}{8.2148} + \underset{(0.00036)}{0.0305}\ \text{trend}$$

- Thus real GDP increased every year by about _________ %, on average.

Fitting an exponential trend to GDP: actual and fitted values



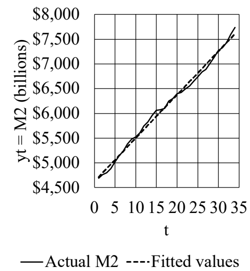
TIME TRENDS

Forecasting time series

- In this section of the course our purpose is forecasting a single time series from its own past values.
- That is, we wish to forecast y_{T+1}, y_{T+2}, y_{T+3}, etc. where the information set is $y_1, \ldots, y_T$.
- No xs.

Linear trend: example

Suppose our model is $y_t = \beta_1 + \beta_2 t + \varepsilon_t$, where ε_t is not observed but is assumed to be IID with mean zero and constant variance σ^2.



Linear trend: estimation

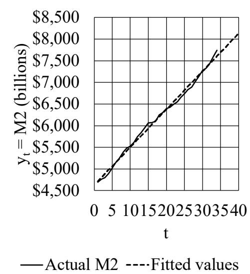
These assumptions imply

- BLUE estimators of β_1 and β_2 are just the LS estimators.
- Unbiased estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{T-2}\sum_{t=1}^{T}\hat{\varepsilon}_t^2$

Here, $\hat{\beta}_1$= 4623.8, $\hat{\beta}_2$= 87.7, and $\hat{\sigma}^2$=4294.8 .

Linear trend: point forecasts

- Our forecasts for $\hat{y}_{T+1}$, $\hat{y}_{T+2}$, etc., are just
  $\hat{\beta}_1 + \hat{\beta}_2\,(T+1)$
  $\hat{\beta}_1 + \hat{\beta}_2\,(T+2)$
  etc.



Linear trend: one-step-ahead (T+1) forecast error

- Forecast error $= y_T - \hat{y}_{T+1}$
  $= [\beta_1 + \beta_2(T+1) + \varepsilon_{T+1}] - [\hat{\beta}_1 + \hat{\beta}_2(T+1)]$
  $= (\beta_1 - \hat{\beta}_1) + (\beta_2 - \hat{\beta}_2)(T+1) + \varepsilon_{T+1}$
- Note that expected value is zero because $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased and ε_t has mean zero.
- As before, forecast error is due to (1) errors in estimating coefficients and (2) the brand-new error term ε_{T+1} . There is no covariance if ε_t is IID.

Linear trend: one-step-ahead (T+1) forecast interval

- Using formula given in "Two Variable Regression" in earlier slideshow on "Prediction Intervals," we have

$$SE = \sqrt{\hat{\sigma}^2\left(\frac{1}{T} + \frac{([T+1]-\bar{t})^2}{\sum_{t=1}^{T}(t-\bar{t})^2} + 1\right)}$$

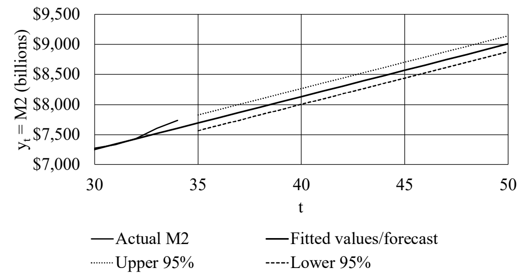
  where $\bar{t} = average\ t = \frac{1}{T}\sum_{t=1}^{T} t = \frac{T+1}{2}$.
- Assuming ε_t is normally-distributed, then 95% forecast interval is $\hat{\beta}_1 + \hat{\beta}_2\,(T+1) \pm 1.96\,SE$

TIME TRENDS

Linear trend: forecasting h-steps-ahead (T+h)

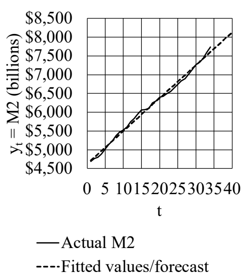
- Point forecast: $\hat{\beta}_1 + \hat{\beta}_2 (T+h)$
- $SE = \sqrt{\hat{\sigma}^2 \left(\frac{1}{T} + \frac{([T+h]-\bar{t})^2}{\sum_{t=1}^{T}(t-\bar{t})^2} + 1 \right)}$
  where $\bar{t} = average\ t = \frac{1}{T}\sum_{t=1}^{T} t = \frac{T+1}{2}$.
- 95% forecast interval is
  $\hat{\beta}_1 + \hat{\beta}_2 (T+h) \pm 1.96\, SE$

Linear trend: forecast interval



R^2 and adjusted R^2 values often *very high* in time series regressions

- Variation in the fitted values is typically much greater than variation in residuals.
- In the example just given modeling M2, $R^2 = 0.995$.



Moreover, R^2 and adjusted $R^2 \to 1$ as sample size $T \to \infty$

- To see this, assume y_t is trended and the variance of the error term is constant.
- Then $\left(\frac{1}{T-K}\right)\sum \hat{\varepsilon}^2 \xrightarrow{P} \sigma^2$, a constant.
- But $\left(\frac{1}{T-1}\right)\sum \left(y_i - \bar{y}\right)^2$ grows without bound.

Moreover, R^2 and adjusted $R^2 \to 1$ as sample size $T \to \infty$ (cont'd)

- Theil's adjusted R^2:

$$\bar{R}^2 = 1 - \frac{\left(\frac{1}{T-K}\right)\sum \hat{\varepsilon}^2}{\left(\frac{1}{T-1}\right)\sum \left(y_i - \bar{y}\right)^2}$$

- Clearly the second term must approach zero, so the adjusted R^2 must approach one.

Conclusions

- Many time series show clear linear or exponential time trends.
- The time trend is _____________ if the dependent variable is in levels, and ________________ if the dependent variable is in logs.
- Trends may be modeled in two-variable regression with a time trend variable and an IID error term.

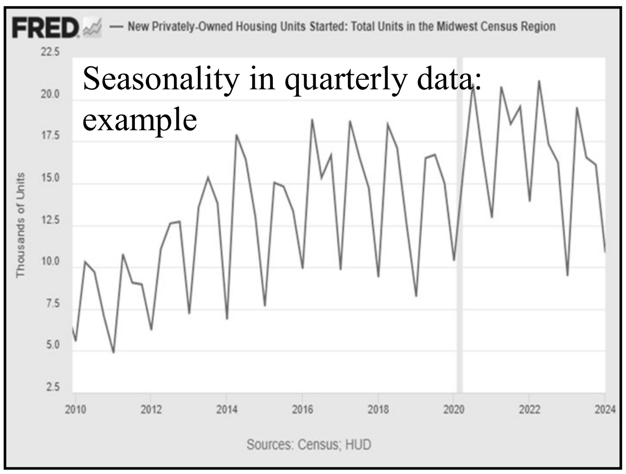
SEASONALITY

SEASONALITY

- What is seasonality in time-series data?
- What problems does it cause?
- How can they be solved?

Seasonal fluctuations in time series

- Many monthly or quarterly time series show seasonal fluctuations. Examples:
  - Housing starts are higher in summer than winter.
  - Electricity demand is higher in summer and winter than spring or fall.
  - Unemployment (and employment) rise in June.
  - Retail sales are higher in fourth quarter than other quarters (due to Christmas).



Seasonal adjustment

- *Seasonal adjustment* means removing the effects of fluctuations that follow an annual cycle.
- Some data published by the government appear to show no seasonal fluctuations because they are already seasonally-adjusted.
- Example: GDP and related data published by the Bureau of Economic Analysis.

http://www.bea.gov/

Seasonal adjustment (cont'd)

- Other data are published both with and without seasonal adjustment.
- Example: Employment and related data published by the Bureau of Labor Statistics.
- However, data that we collect ourselves (company sales, local economic activity, etc.) are _______ likely to be seasonally-adjusted.

http://www.bls.gov/

Problems caused by seasonality

- If seasonality in data is not controlled for, problems may result.
- The error term will be serially correlated (often negatively), violating assumption #4 (no serial correlation). This would invalidate standard errors, tests, and forecast intervals.
- Unrelated series may appear correlated if they both have similar seasonal patterns. This is because assumption #2 (exogeneity) is likely violated.

SEASONALITY

Controlling for seasonality

- An easy way to control for seasonality is to include dummy variables for each "season."
- Example: Suppose we wish to estimate $y_t = \beta_1 + \beta_2 t + \varepsilon_t$ using quarterly data.
- Define $d1_t = 1$ for all observations in the first quarter, = 0 otherwise.
- Similarly, define $d2_t$ and $d3_t$.

Quarterly dummy variables in a data spreadsheet

t	Date	y_t	t	$d1_t$	$d2_t$	$d3_t$
1	2010 first quarter	1.7	1	1	0	0
2	2010 second quarter	2.3	2	0	1	0
3	2010 third quarter	2.3	3	0	0	1
4	2010 fourth quarter	2.0	4	0	0	0
5	2011 first quarter	1.6	5			
6	2011 second quarter	2.4	6			
7	2011 third quarter	2.2	7			
	etc.					

Seasonal dummy variables

- Then estimate the regression $y_t = \beta_1 + \beta_2 t + \beta_3 d1_t + \beta_4 d2_t + \beta_5 d3_t + \varepsilon_t$
- Note that only ___________ dummies are used for four seasons.
- If a fourth dummy ($d4_t$) were included, then the sum of the dummy variables would always be $d1_t+d2_t+d3_t+d4_t=1$ for every observation: perfect ___________________.

Seasonal dummy variables: estimation

If ε_t are IID and the usual assumptions hold, then

- BLUE estimators of β_1 through β_5 are just the LS estimators.
- Unbiased estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{T-2}\sum_{t=1}^{T}\hat{\varepsilon}_t^2$

Interpreting seasonal dummy coefficients

- Since the fourth-quarter dummy is omitted, the fourth-quarter intercept equals β_1 .
- The value of β_3 shows how much higher y_t is in the first quarter than the fourth quarter, *ceteris paribus*—that is, purely because of seasonal effects.
- Similarly for β_4 and β_5 .

Example: housing starts

- This regression was estimated using quarterly data (2010 Q1-2024 Q1) for Midwest region:

$$\ln(\text{housing starts}) = \underset{(0.053)}{2.227} + \underset{(0.001)}{0.013}\, t$$

$$- \underset{(0.055)}{0.454}\, d1 + \underset{(0.056)}{0.180}\, d2 + \underset{(0.056)}{0.119}\, d3$$

Source: FRED, "new Privately-Owned Housing Units Started: Total Units in the Midwest Census Region, Thousands of Units, Quarterly, Not Seasonally Adjusted.

SEASONALITY

Example: interpretation of coefficients

- On average, the number of housing starts increased by ______% each quarter.
- However, housing starts are on average
  - ________% lower in the first quarter
  - ________% higher in the second quarter
  - ________% higher in the third quarter than in the fourth quarter.

(using the log approximation for percent changes)

Forecasting with seasonal dummies

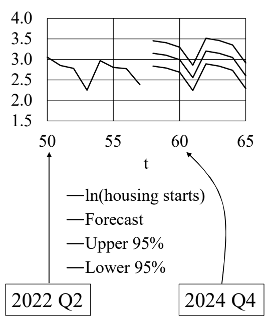
- Point forecasts simply insert appropriate values of time trend and seasonal dummies.
- As before, forecast error is due to (1) errors in estimating coefficients and (2) the brand-new error term ε_{T+1} . There is no covariance if ε_t is IID.
- The usual formulas for SE of prediction error apply.*

*See slideshow, "Prediction and Prediction Intervals with Multiple Regression"

Forecast intervals with seasonal dummies

Assuming ε_t is normally-distributed, then the 95% forecast interval is

$$\hat{\beta}_1 + \hat{\beta}_2 (T+1) + \hat{\beta}_3 d1_{T+1} + \hat{\beta}_4 d2_{T+1} + \hat{\beta}_5 d3_{T+1} \pm 1.96\ SE$$



Seasonality in monthly data

- With monthly data, __________ dummy variables are needed:

$y_t = \beta_1 + \beta_2 t + \beta_3 djan_t + \beta_4 dfeb_t$
$+ \beta_5 dmar_t + \beta_6 dapr_t + \beta_7 dmay_t$
$+ \beta_8 djun_t + \beta_9 djul_t + \beta_{10} daug_t$
$+ \beta_{11} dsep_t + \beta_{12} doct_t + \beta_{13} dnov_t + \varepsilon_t$

Conclusions

- Monthly and quarterly data often show "seasonal" fluctuations over an annual cycle.
- To control for seasonal fluctuations, include dummy variables for "seasons":
  - _____ dummy variables for quarterly data.
  - _____ dummy variables for monthly data.

STATIONARY AND WEAKLY DEPENDENT TIME SERIES

STATIONARY AND WEAKLY DEPENDENT TIME SERIES

- What is a "stationary time series?
- What is a "weakly-dependent" series?
- What is a "trend-stationary" series?

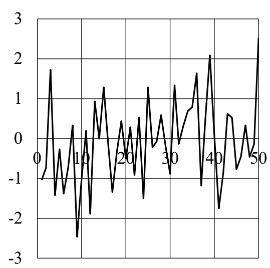
Strict exogeneity and no serial correlation

- Till now, we have assumed that error terms were IID.*
- This assumption led to strong conclusions—that LS was unbiased, BLUE, etc.
- They also made forecasting relatively simple.
- However these assumptions are usually unrealistic.

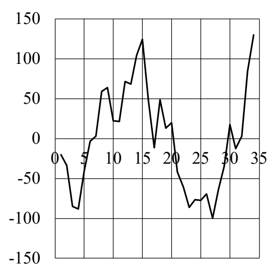
*Independent identically distributed.

Compare!

IID series

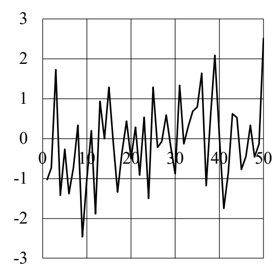


Residuals from regression of M2 on a time trend

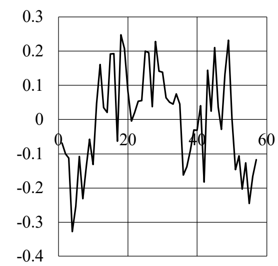


Compare!

IID series



Residuals from regression of ln(housing starts) on time trend and seasonal dummies

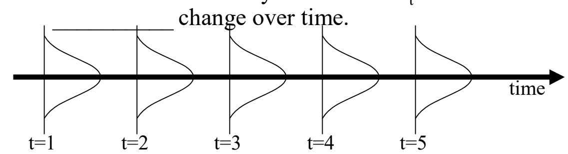


Serially-correlated processes

- Here we consider stochastic processes (or series) that are not IID.
- But they do satisfy weaker assumptions under which LS might still work pretty well.
- We define
  - Stationary time series.
  - Weakly dependent time series.

Stationary time series: definition

- A *stationary time series* is one whose distribution does _________ change over time.
- For example, if u_t is a stationary time series, then u_t has the same distribution as u_{t-1}, u_{t+1}, u_{t+2} and u_{t+h} (where h is any integer).
- This means the density function of u_t does ___________ change over time.



STATIONARY AND WEAKLY DEPENDENT TIME SERIES

Moments of a stationary time series

- If a time series is stationary, its (unconditional) moments do not change over time.
- $E(u_t) = E(u_{t+1}) = E(u_{t+2}) = E(u_{t+h})$.
- $Var(u_t) = Var(u_{t+1}) = Var(u_{t+2}) = Var(u_{t+h})$.
- But u_t could still be serially-correlated.
- We need some terminology to describe serial correlation.

Autocovariances of a time series: definition

- *Autocovariance* = covariance between u_t and one of its own past values.
  - First autocovariance = $Cov(u_t,u_{t-1})$.
  - Second autocovariance = $Cov(u_t,u_{t-2})$.
  - etc.
- In general, the first autocovariance is not equal to the second.

Autocovariance and serial correlation

- Recall that correlation is related to covariance by definition:
- Corr(X,Y) =
- A series which has nonzero auto*covariances* must also have nonzero auto*correlations*.
- Also called "serially correlated."

Autocovariances of a stationary time series

- The autocovariances of a stationary time series do not change over time.
- $Cov(u_t,u_{t-1}) = Cov(u_{t+1},u_t) = Cov(u_{t+h},u_{t+h-1})$.
- $Cov(u_t,u_{t-2}) = Cov(u_{t+1},u_{t-1}) = Cov(u_{t+h},u_{t+h-2})$.
- Thus, the covariance between the first and second u_t is the same as the covariance between the 99th and ________ u_t , and between the 999th and ________ u_t .

Covariance-stationary series: definition

- If the means, variances, and autocovariances do not change over time, the series is called *covariance-stationary*.
- *Covariance-stationarity* is a weaker condition than *stationarity* in that it does not require that the whole density function be constant over time—just means, variances, and autocovariances.

IID random variables are a trivial example of a stationary time series

- Earlier we considered error terms ε_t that satisfied $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma^2$, and $Cov(\varepsilon_t,\varepsilon_s) = 0$, obviously all constant with respect to *t*.
- Under these assumptions, ε_t is a ________________________________ process.
- If in addition we assume ε_t is independent normally-distributed, then ε_t is a ____________ process.

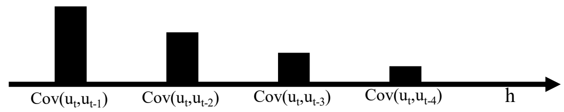
STATIONARY AND WEAKLY DEPENDENT TIME SERIES

Weakly dependent time series: definition

- A *weakly dependent times series* is one where u_t and u_{t+h} become "more independent" as h gets larger.
- Thus they become "more independent" the ____________ apart they are in time.
- This definition is obviously not very precise. More precise definitions are used, but they vary according to context.

Asymptotically uncorrelated process: definition

- One precise definition of weak dependence requires autocovariances and autocorrelations to converge to zero as observations are farther and farther apart.
- u_t is said to be *asymptotically uncorrelated* if $\mathrm{Corr}(u_t,u_{t+h}) \rightarrow$ ___ as $h \rightarrow$ ________.

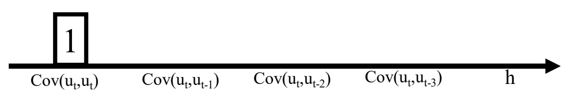


Weak dependence versus random sampling: intuition

- In a __________ sample, each observation is independent and "fresh," It contributes completely new information to the sample.
- In a ____________________ process, each observation is not completely fresh. But it does contribute some new information to the sample. It is not simply a duplicate of a previous observation.

IID random variables are a trivial example of a weakly dependent series

- In the last section, we sometimes assumed the error term ε_t was independent identically-distributed.
- Independent random variables obviously and trivially satisfy the definition of "weakly dependent" because $\mathrm{Corr}(u_t,u_{t+h}) =$ ____ for all $h \neq 0$.



Other weakly dependent time series

- Starting with independent identically-distributed random variables ε_t, we can define other series that are also weakly dependent.
- Examples (see next presentations):
  - Moving average (MA) process.
  - Autoregressive (AR) process.

A series can be stationary but not weakly dependent

- Suppose all observations in a series are equal to each other.
- That is, $u_t = u_{t+1} = u_{t+2} = u_{t+3} = \ldots = u_{t+h} + \ldots$
- All observations share the _________ distribution, so this (admittedly strange) series is stationary.
- But $\mathrm{corr}(u_t, u_{t+h}) = 1$ for *all* values of h, so this series is ________ weakly dependent.

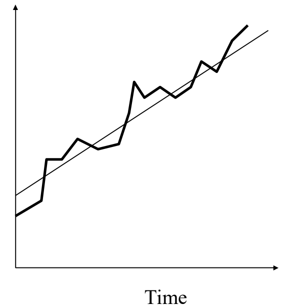
STATIONARY AND WEAKLY DEPENDENT TIME SERIES

A series can be weakly dependent but not stationary

- Suppose all observations in a series are independent random variables, but with different variances (i.e., heteroskedastic).
- For example, $Var(u_t)=3$, $Var(u_{t+1})=17$, $Var(u_{t+2})=5$, $Var(u_{t+3})=13$, etc.
- $Corr(u_t, u_{t+h}) =$ ______, for all $h \neq 0$, so the series is weakly dependent.
- But each observation has a different variance, so the series is _________ stationary.

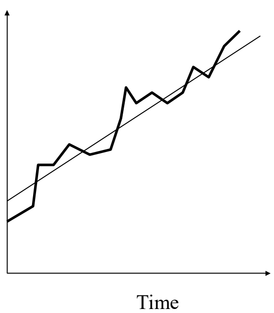
A trended series cannot be stationary

- A trended series cannot be stationary because its mean (the time trend) is changing over time.
- However, it may be stationary *around its time trend.*
- That is, $(u_t$ - trend) might be stationary.



But a trended series *can* be weakly dependent

- A trended series can be weakly dependent.
- Example: linear trend $u_t = \alpha_1 + \alpha_2\, t + \varepsilon_t$, where ε_t is independent.
- Then $Corr(u_t, u_{t+h})$ $= Corr(\varepsilon_t, \varepsilon_{t+h})$ $= 0$ for all $h > 0$.



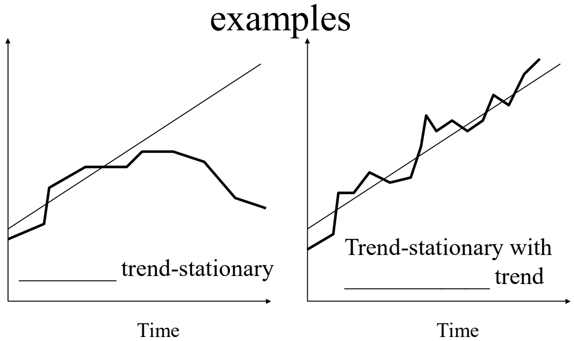
Trend-stationary series: definition

- A *trend-stationary series* is a trended series that is stationary *around* its trend and weakly dependent.
- In other words, subtract the trend and you have a stationary, weakly dependent series:

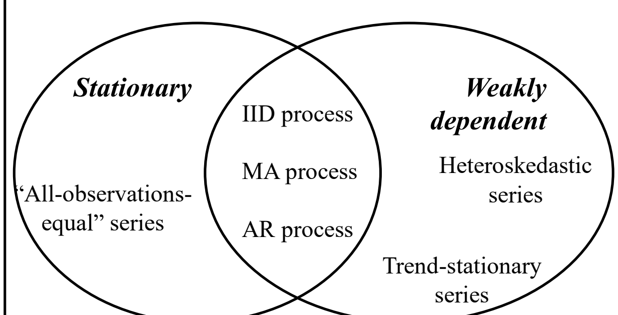
$$u_t = \text{trend} + \varepsilon_t ,$$

where ε_t is stationary and $Corr(\varepsilon_t, \varepsilon_{t+h}) \rightarrow 0$ as $h \rightarrow$ infinity.

Trend-stationary series: examples



Summary: stationarity versus weak dependence



STATIONARY AND WEAKLY
DEPENDENT TIME SERIES

Conclusions

- A _________________ series or process is one whose distribution does not change over time.
- A ________________________ series is one where u_t and u_{t+h} become "more independent" as $h \rightarrow$ infinity.
- An asymptotically uncorrelated series is one where $Corr(u_t,u_{t+h}) \rightarrow 0$ as $h \rightarrow$ infinity.
- A ____________________ series is a trended series that is stationary *around its trend* and weakly dependent.

FIRST-ORDER MOVING AVERAGE PROCESS

FIRST-ORDER MOVING AVERAGE PROCESS

- What is an "MA(1)" process?
- Why is it always stationary and weakly dependent?

Modeling serial correlation

- In the real world, most time-series are serially-correlated.
- For accurate estimation and forecasting, we need models of serial correlation that
  - fit the data reasonably well, and
  - are not excessively complex.

Popular models of serial correlation

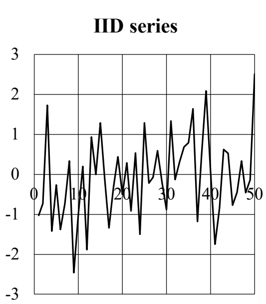
- Most popular models minimize complexity by building models of serial correlated random variables (y_t) up from other latent (unobserved) random variables that are *not* serially correlated (ε_t):
- Moving average process.
- Autoregressive process.

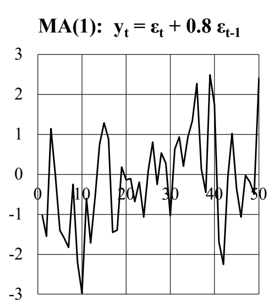
Definition of first-order moving average process (MA(1))

- Suppose ε_t is an IID* series with $E(\varepsilon_t)=0$ and $Var(\varepsilon_t)=\sigma_\varepsilon^2$.
- Thus $Cov(\varepsilon_t,\varepsilon_s) = 0$ whenever $t \neq s$.
- Let $y_t = \varepsilon_t + \alpha\, \varepsilon_{t-1}$, where α is a constant.
- We now show that y_t is stationary, serially correlated, and weakly dependent.

*Independent identically distributed.

Compare!





Mean of MA(1)

- $E(y_t) = E(\varepsilon_t+\alpha\varepsilon_{t-1}) = E(\varepsilon_t) + \alpha\, E(\varepsilon_{t-1})$
  $= 0 + \alpha\, 0 =$ ______.

FIRST-ORDER MOVING AVERAGE PROCESS

Variance of MA(1)

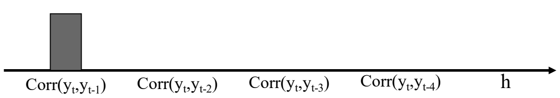
- $\mathrm{Var}(y_t) = E[(y_t - Ey_t)^2] = E[y_t^2]$
  $= E[(\varepsilon_t + \alpha\varepsilon_{t-1})^2] = E[\varepsilon_t^2 + 2\varepsilon_t\alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-1}^2]$
  $= E[\varepsilon_t^2] + 2\alpha\, E[\varepsilon_t\varepsilon_{t-1}] + \alpha^2\, E[\varepsilon_{t-1}^2]$
  $= \mathrm{Var}(\varepsilon_t^2) + 2\alpha\, \mathrm{Cov}(\varepsilon_t,\varepsilon_{t-1}) + \alpha^2\, \mathrm{Var}(\varepsilon_{t-1}^2).$
  $= \sigma_\varepsilon^2 + \alpha^2\, \sigma_\varepsilon^2$
  $= \boldsymbol{(1+\alpha^2)\, \sigma_\varepsilon^2}$**.**

Autocovariances of MA(1)

- $\mathrm{Cov}(y_t,y_{t-1}) = E(y_t y_{t-1})$
  $= E(\varepsilon_t + \alpha\varepsilon_{t-1})(\varepsilon_{t-1} + \alpha\varepsilon_{t-2})$
  $= E[\varepsilon_t\varepsilon_{t-1}] + \alpha E[\varepsilon_{t-1}\varepsilon_{t-1}] + \alpha E[\varepsilon_t\varepsilon_{t-2}] + \alpha^2 E[\varepsilon_{t-1}\varepsilon_{t-2}]$
  $= \alpha\, \mathrm{Var}(\varepsilon_{t-1}) = \boldsymbol{\alpha\, \sigma_\varepsilon^2}$**.**
- $\mathrm{Cov}(y_t,y_{t-2}) = E(y_t y_{t-2})$
  $= E(\varepsilon_t + \alpha\varepsilon_{t-1})(\varepsilon_{t-2} + \alpha\varepsilon_{t-3})$
  $= E[\varepsilon_t\varepsilon_{t-1}] + \alpha E[\varepsilon_{t-1}\varepsilon_{t-2}] + \alpha E[\varepsilon_t\varepsilon_{t-3}] + \alpha^2 E[\varepsilon_{t-1}\varepsilon_{t-3}]$
  $= \mathbf{0}$.
- Clearly $\mathrm{Cov}(y_t,y_{t-h}) = 0$ for $h>1$.

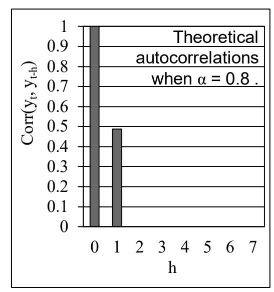
Autocorrelations of MA(1)

- Autocorrelation of a stationary series is defined as autocovariance divided by variance.
- Using results on the last two slides gives the following.
- $\mathrm{Corr}(y_t,y_{t-1}) = [\alpha\, \sigma_\varepsilon^2] / ([1+\alpha^2]\, \sigma_\varepsilon^2) = \alpha/(1+\alpha^2)$.
- $\mathrm{Corr}(y_t,y_{t-2}) =$ _______.
- Clearly $\mathrm{Corr}(y_t,y_{t-h}) =$ _______ for $h>1$.



Autocorrelations of MA(1) process drop abruptly to zero

- Like the autocovariances, the autocorrelations decay for an MA(1) process drop abruptly to zero after 1st autocorrelation.
- y_t and y_{t-2} are uncorrelated.



Stationarity and weak dependence of MA(1) process

- We have shown that $E(y_t)$, $\mathrm{Var}(y_t)$, and the autocovariances do *not* depend on t .
- So y_t is covariance _____________.
- We have shown that $\mathrm{Cov}(y_t,y_s) =$ _____ whenever t and s are more than one period apart.
- So y_t is *asymptotically uncorrelated* (a form of weak dependence).

Estimating an MA(1) model

- To fit actual data, add a constant term:
  $y_t = \beta_1 + \varepsilon_t + \alpha\, \varepsilon_{t-1}$,
- Estimation is rather complicated but is automated in statistical software.

FIRST-ORDER MOVING AVERAGE PROCESS

Forecasting with an MA(1) model

- One-step-ahead forecast inserts estimated coefficients, last residual, and zero (the expected value) for ε_{T+1}:
  $$\hat{y}_{T+1} = \hat{\beta}_1 + \varepsilon_{T+1} + \alpha\,\hat{\varepsilon}_T$$
- h-steps-ahead forecast inserts zero for all future values of ε_{t+h}:
  $$\hat{y}_{T+h} = \hat{\beta}_1 + \varepsilon_{T+h} + \alpha\,\varepsilon_{T+h-1}$$

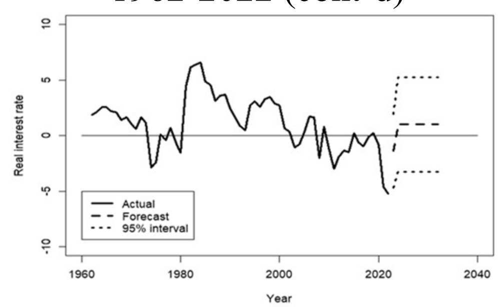
Forecast intervals with an MA(1) model

- SEs for coefficients are usually relatively tiny—it is common to ignore them when computing forecast intervals.
- Assuming ε_t is normally-distributed, then 95% forecast intervals are as follows.
- One-step-ahead: $\hat{y}_{T+1} \pm 1.96\sqrt{\hat{\sigma}_\varepsilon^2}$
- h-steps-ahead, h>1: $\hat{y}_{T+h} \pm 1.96\sqrt{(1+\hat{\alpha}^2)\hat{\sigma}_\varepsilon^2}$

Example: real interest rate 1962-2022

- Definition: Treasury one-year rate minus inflation rate (CPI).
- Sample mean = 1.058 (percent).
- $y_t = \underset{(0.387)}{1.023} + \varepsilon_t + \underset{(0.075)}{0.799}\;\varepsilon_{t-1}$
- $\hat{\sigma}_\varepsilon^2 = 2.862$

Example: real interest rate 1962-2022 (cont'd)



Extension: MA(q)

- $y_t = \beta_1 + \varepsilon_t + \alpha_1\,\varepsilon_{t-1} + \alpha_2\,\varepsilon_{t-2} + \ldots + \alpha_q\,\varepsilon_{t-q}$, where β_1 and αs are constants.
- Autocovariances $Cov(y_t, y_{t-h})$ and autocorrelations $Corr(y_t, y_{t-h})$ depend on h but not t, and are zero for h>q.
- MA(q) process is therefore stationary and weakly dependent.

Conclusions

- The MA(1) process is defined as $y_t = \varepsilon_t + \alpha\,\varepsilon_{t-1}$, where ε_t is an IID process with mean ________.
- The MA(1) process is always stationary.
- Autocovariances and autocorrelations are nonzero for one period's lag, but _________ thereafter.
- Point forecasts and forecast intervals are constant beginning two steps ahead.

FIRST-ORDER AUTOREGRESSIVE PROCESS

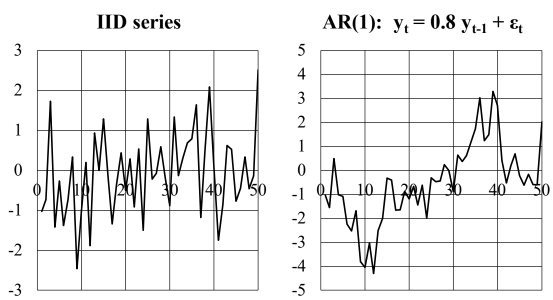
FIRST-ORDER AUTOREGRESSIVE PROCESS

- What is an "AR(1)" process?
- When is it weakly dependent?

Definition of first-order autoregressive process (AR(1))

- Suppose ε_t is an independent identically-distributed (IID) series with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_\varepsilon^2$.
- Thus $Cov(\varepsilon_t,\varepsilon_s) = 0$ whenever $t \neq s$.
- Let $y_t = \phi\, y_{t-1} + \varepsilon_t$. Assume $|\rho| < 1$.
- Assume the initial value y_0 is independent of all the ε_t , for $t > 0$.

Compare!



Properties of AR(1)

- Note that y_t depends on the current ε_t and depends (through y_{t-1}) on past ε_t , but does *not* depend on future values of ε_t .
- Thus $Cov(y_t, \varepsilon_s)$ = _____ for all s>t.
- If $|\phi| < 1$, it can be shown that the AR(1) process is stationary.
- We assume stationarity and show that y_t is serially correlated and weakly dependent.

Mean of AR(1)

- Now $E(y_t) = E(\phi\, y_{t-1} + \varepsilon_t) = \phi\, E(y_{t-1}) + E(\varepsilon_t)$ $= \phi\, E(y_{t-1})$, since $E(\varepsilon_t) = 0$.
- By stationarity, $E(y_t) = E(y_{t-1})$ so $E(y_t) = \phi\, E(y_t)$
- So $(1-\phi)\, E(y_t) = 0$. Therefore $E(y_t)$ = ____.

Variance of AR(1)

- Now $Var(y_t) = Var(\phi\, y_{t-1} + \varepsilon_t)$ $= \phi^2\, Var(y_{t-1}) + Var(\varepsilon_t) + 2\,\phi\, Cov(y_{t-1},\varepsilon_t)$.
- Now $Cov(y_t, \varepsilon_s) = 0$ for all s>t, so $Var(y_t) = \phi^2\, Var(y_{t-1}) + Var(\varepsilon_t)$.
- By stationarity, $Var(y_{t-1}) = Var(y_t)$, so $Var(y_t) = \phi^2\, Var(y_t) + \sigma_\varepsilon^2$.
- So $(1-\phi^2)\, Var(y_t) = \sigma_\varepsilon^2$.
- Therefore $Var(y_t) = \sigma_\varepsilon^2 / (1-\phi^2)$.

FIRST-ORDER AUTOREGRESSIVE PROCESS

Autocovariances of AR(1)

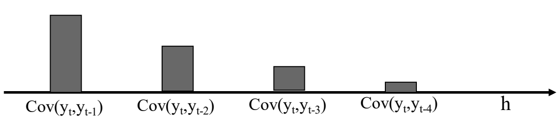
- Begin by writing $y_{t+h} = \phi\, y_{t+h-1} + \varepsilon_{t+h}$
  $= \phi\,(\phi\, y_{t+h-2} + \varepsilon_{t+h-1}) + \varepsilon_{t+h}$
  $= \phi^2\, y_{t+h-2} + \phi\, \varepsilon_{t+h-1} + \varepsilon_{t+h}$
  $= \phi^2\,(\phi\, y_{t+h-3} + \varepsilon_{t+h-2}) + \phi\, \varepsilon_{t+h-1} + \varepsilon_{t+h}$
  $= \phi^3\, y_{t+h-3} + \phi^2\, \varepsilon_{t+h-2} + \phi\, \varepsilon_{t+h-1} + \varepsilon_{t+h}$
  …
  $= \phi^h y_t + \phi^{h-1}\varepsilon_{t+1} + \ldots + \phi^2\varepsilon_{t+h-2} + \phi\varepsilon_{t+h-1} + \varepsilon_{t+h}$

Pattern of autocovariances

- Now $\text{Cov}(y_t, \varepsilon_s) = 0$ for all $s > t$.
- So for $h \geq 0$,
  $\text{Cov}(y_t, y_{t+h}) = \text{Cov}(y_t, \phi^h\, y_t) = \phi^h\, \text{Var}(y_t)$.
- Assuming $|\,\phi\,| < 1$, then the autocovariances decrease (in absolute value) as h increases.
- In other words, the longer the time interval between two observations y_t and y_{t-h}, the ___________ their covariance (in absolute value).
- But the covariance never reaches _________.

Pattern of autocovariances: example

- Suppose $\text{Var}(y_t) = 5$ and $\phi = 0.5$.
- Then $\text{Cov}(y_t,y_{t+1}) = (0.5)^1 \times 5 = 5/2$,
  $\text{Cov}(y_t,y_{t+2}) = (0.5)^2 \times 5 =$ _______,
  $\text{Cov}(y_t,y_{t+3}) = (0.5)^3 \times 5 =$ _______,
  $\text{Cov}(y_t,y_{t+4}) = (0.5)^4 \times 5 =$ _______, etc.

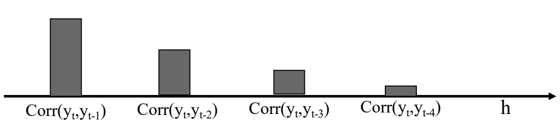


Autocorrelations of AR(1)

- Autocorrelation of a stationary series is defined as autocovariance divided by variance.
- So here, the h-th autocorrelation
  $\text{Corr}(y_t, y_{t+h}) = \text{Cov}(y_t, y_{t+h}) / \text{Var}(y_t)$.
- We already showed that for $h \geq 0$,
  $\text{Cov}(y_t, y_{t+h}) = \phi^h\, \text{Var}(y_t)$.
- So for $h \geq 0$,
  $\text{Corr}(y_t, y_{t+h}) = [\phi^h\, \text{Var}(y_t)] / \text{Var}(y_t) = \boldsymbol{\phi^h}$.

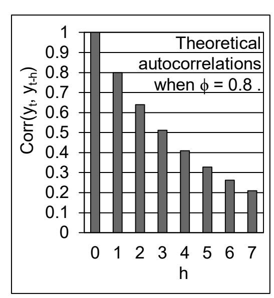
Pattern of autocorrelations: example

- Again, suppose $\phi = 0.5$.
- Then $\text{Corr}(y_t,y_{t+1}) = (0.5)^1 = 0.5$,
  $\text{Corr}(y_t,y_{t+2}) = (0.5)^2 =$ _________,
  $\text{Corr}(y_t,y_{t+3}) = (0.5)^3 =$ _________,
  $\text{Corr}(y_t,y_{t+4}) = (0.5)^4 =$ _________, etc.



Autocorrelations of AR(1) decay toward zero

- Like the autocovariances, the autocorrelations decay with *h*, the time interval between y_t and y_{t+h} .
- The decay factor is ϕ.



FIRST-ORDER AUTOREGRESSIVE PROCESS

Weak dependence of AR(1) process

- Assuming ϕ is less than one in absolute value, then as h $\to$ infinity,
  - $Cov(y_t, \phi^h y_t) = \phi^h Var(y_t) \to$ _____.
  - $Corr(y_t, y_{t+h}) = \phi^h \to$ _____.
- So y_t is *asymptotically uncorrelated* (a form of weak dependence).

Estimating an AR(1) model

- To fit actual data, add a constant:
  $y_t = \beta_1 + \phi\, y_{t-1} + \varepsilon_t$.
- Can be estimated by ordinary least squares after dropping the first observation (because y_0 is not observed).
- Other estimation methods keep the first observation and impute y_0 somehow.

Forecasting with an AR(1) model

- One-step-ahead forecast inserts estimated coefficients, last actual y_T and zero (the expected value) for ε_{T+1}:
  $$\hat{y}_{T+1} = \hat{\beta}_1 + \hat{\varphi} y_T + \varepsilon_{T+1}$$
- Two-steps ahead forecast inserts first forecast:
  $$\hat{y}_{T+2} = \hat{\beta}_1 + \hat{\varphi}\, \hat{y}_{T+1}$$
- h-steps-ahead forecast recursively inserts prior forecasts:
  $$\hat{y}_{T+h} = \hat{\beta}_1 + \hat{\varphi}\, \hat{y}_{T+h-1}$$

Forecasting with an AR(1) model

- One-step-ahead forecast inserts estimated coefficients, last actual y_T and zero (the expected value) for ε_{T+1}:
  $$\hat{y}_{T+1} = \hat{\beta}_1 + \hat{\varphi} y_T + \varepsilon_{T+1}$$
- Two-steps ahead forecast inserts first forecast:
  $$\hat{y}_{T+2} = \hat{\beta}_1 + \hat{\varphi}\, \hat{y}_{T+1}$$
- h-steps-ahead forecast recursively inserts prior forecasts:
  $$\hat{y}_{T+h} = \hat{\beta}_1 + \hat{\varphi}\, \hat{y}_{T+h-1}$$

Forecast intervals with an AR(1) model

- SEs for coefficients are usually relatively tiny—it is common to ignore them when computing forecast intervals.
- Assuming ε_t is normally-distributed, then 95% forecast intervals are as follows.
- One-step-ahead: $\hat{y}_{T+1} \pm 1.96\sqrt{\hat{\sigma}_\varepsilon^2}$

Forecast intervals with an AR(1) model: two steps ahead

- Substitution shows that
  $$\begin{aligned} y_{T+2} &= \beta_1 + \phi\, y_{T-1} + \varepsilon_{T+2} \\ &= \beta_1 + \phi\, (\beta_1 + \phi\, y_T + \varepsilon_{T+1}) + \varepsilon_{T+2} \\ &= \beta_1 (1+\phi) + \phi^2 y_T + \varepsilon_{T+2} + \phi\, \varepsilon_{T+1} \end{aligned}$$
- So $Var(\hat{y}_{T+2}) = Var(\varepsilon_{T+2} + \varphi\varepsilon_{T+1})$
  $= \sigma_\varepsilon^2(1+\varphi^2)$
- Two-steps-ahead interval:
  $$\hat{y}_{T+2} \pm 1.96\sqrt{\hat{\sigma}_\varepsilon^2(1+\hat{\varphi}^2)}$$

FIRST-ORDER AUTOREGRESSIVE PROCESS

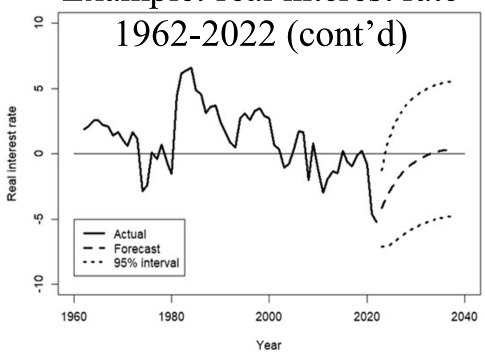
Forecast intervals with an AR(1) model: h steps ahead

- Repeated substitution shows that
  $y_{T+h} = \beta_1 (1+ \phi + \phi^2 + ... + \phi^h) + \phi^h y_T$
  $+ \varepsilon_{T+h} + \phi\, \varepsilon_{T+h-1} + \phi^2\, \varepsilon_{T+h-2} + ... + \phi^h\, \varepsilon_{T+1}$
- So $Var(\hat{y}_{T+h})$
  $= Var(\varepsilon_{T+h} + \varphi\varepsilon_{T+h-1} + \cdots + \varphi^h \varepsilon_{T+1})$
  $= \sigma_\varepsilon^2 (1 + \varphi^2 + \cdots + \varphi^{2h})$
- h-steps-ahead interval:
  $$\hat{y}_{T+h} \pm 1.96\sqrt{\hat{\sigma}_\varepsilon^2 (1 + \hat{\varphi}^2 + \cdots + \hat{\varphi}^{2h})}$$

Example: real interest rate 1962-2022

- Definition: Treasury one-year rate minus inflation rate (CPI).
- Sample mean = 1.058 (percent).
- $y_t = \underset{(1.023)}{0.697} + \underset{(0.078)}{0.824}\; y_{t-1} + \varepsilon_t$
- $\hat{\sigma}_\varepsilon^2 = 2.218$

Example: real interest rate 1962-2022 (cont'd)



Extension: AR(p)

- $y_t = \beta_1 + \phi_1\, y_{t-1} + \phi_2\, y_{t-2} + ... + \phi_p\, y_{t-p} + \varepsilon_t$, where β_1 and the ϕs are constants.
- Autocovariances $Cov(y_t, y_{t-h})$ and autocorrelations $Corr(y_t, y_{t-h})$ depend on h but not on t, and approach zero as h → infinity if $\sum_{i=2}^{p+1} |\varphi_i|$.*
- AR(p) process is therefore stationary and weakly dependent.

*This condition is sufficient. Necessary conditions are weaker but harder to check.

Conclusions

- The AR(1) process is defined as $y_t = \phi\, y_{t-1} + \varepsilon_t$, where ε_t is an IID process with mean __________.
- If $|\phi| < 1$, the AR(1) process is stationary.
- Autocovariances and autocorrelations are never zero, but they decay with factor ϕ, so AR(1) is ______________________.
- Point forecasts are recursive and converge gradually to sample mean. Forecast intervals are bounded.

ARMA(p,q) PROCESS

ARMA(p,q) PROCESS

- What is an ARMA(p,q) process?
- How can we determine p and q?

Definition of ARMA(1,1) process

- Suppose ε_t is an independent identically-distributed (IID) series with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_\varepsilon^2$.
- Let $y_t = \beta_1 + \phi\, y_{t-1} + \varepsilon_t + \alpha\, \varepsilon_{t-1}$.
- Assume $|\phi| < 1$ and $|\alpha| < 1$.
- Clearly, AR(1) and MA(1) are special cases when $\alpha=0$ or $\phi=0$.

Definition of ARMA(p,q) process

- Again, ε_t is an independent identically-distributed (IID) series with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_\varepsilon^2$.
- Let $y_t = \beta_1 + \phi_1\, y_{t-1} + ... + \phi_p\, y_{t-p}$
  $+ \varepsilon_t + \alpha_1\, \varepsilon_{t-1} + ... + \alpha_q\, \varepsilon_{t-q}$.
- Assume $\sum_{i=1}^{p}|\varphi_i| < 1$ and $\sum_{i=1}^{q}|\alpha_i| < 1$.
- Clearly, AR(p) and MA(q) are special cases.

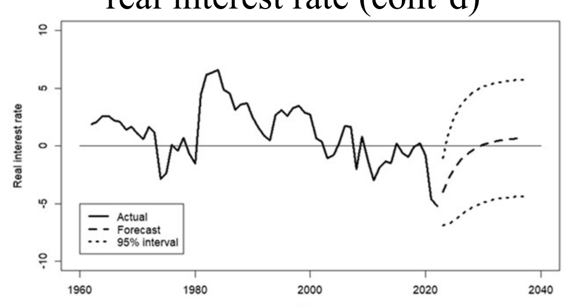
Estimation and forecasting with an ARMA(p,q) model

- Estimation ($\hat{\beta}_1$, $\hat{\varphi}$s, and $\hat{\alpha}$s) is complicated, but statistical software handles this.
- Point forecasts converge gradually to the sample mean.
- Forecast intervals are bounded.

Example: ARMA(1,1) model of real interest rate, 1962-2022

- Definition: Treasury one-year rate minus inflation rate (CPI).
- Sample mean = 1.058 (percent).
- $y_t = \underset{(0.885)}{0.794} + \underset{(0.128)}{0.758}\, y_{t-1} + \varepsilon_t + \underset{(0.211)}{0.171}\, \varepsilon_{t-1}$
- $\hat{\sigma}_\varepsilon^2 = 2.193$

Example: ARMA(1,1) model of real interest rate (cont'd)



ARMA(p,q) PROCESS

How to "identify" an ARMA(p,q) process

How can we determine which ARMA(p,q) model best fits our data? Several methods.

(1) Plot autocorrelation function and partial autocorrelation function. This is the approach originally suggested by Box and Jenkins.

G.E.P. Box and G.M. Jenkins, *Time Series Analysis, Forecasting, and Control,* Holden-Day, 1976, pp. 173-186.

How to "identify" an ARMA(p,q) process (cont'd)

(2) In practice, a good fit usually can be obtained with p and q each $\leq$ than 3. So estimate ARMA(3,3) and drop statistically insignificant coefficients, starting with α_3 and ϕ_3 .

(3) Estimate all combinations of p and q. Choose model with lowest Akaike Information Criterion (AIC).

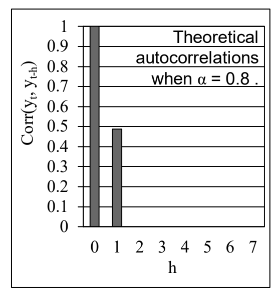
If two models seem to fit the data equally well, choose the simpler model.

(1) Plot autocorrelation function and partial function

- We have seen that the autocorrelation functions are quite different for an MA(1) process versus an AR(1) process.

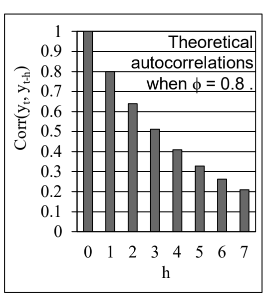
Autocorrelations of ARMA(0,q)=MA(q) process drop abruptly to zero

- Autocorrelations for an MA(1) process drop abruptly to zero after 1st autocorrelation.
- Autocorrelations for an MA(q) process drop abruptly after q^{th} autocorrelation.

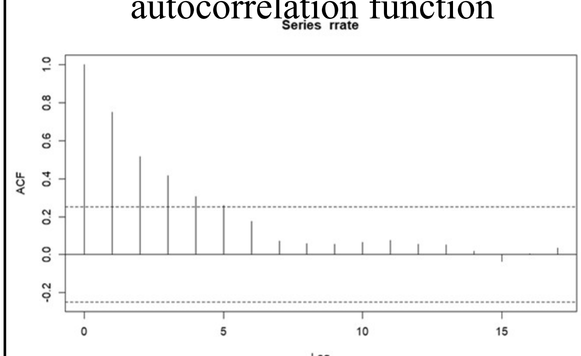


Autocorrelations of ARMA(p,0)=AR(p) process decay toward zero

- Autocorrelations for an AR(1) process decay toward zero with constant factor.
- Autocorrelations for an AR(p) process also decay but with waves.



Example: real interest rate autocorrelation function



ARMA(p,q) PROCESS

Limitations of autocorrelation functions

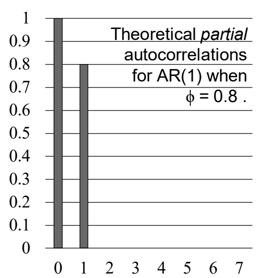
- Autocorrelations cannot be estimated exactly in finite sample. So need to check significance.
- Autocorrelations for AR(p) and ARMA(p,q) models look similar, decaying possibly with waves. Need another tool.

Partial autocorrelation function

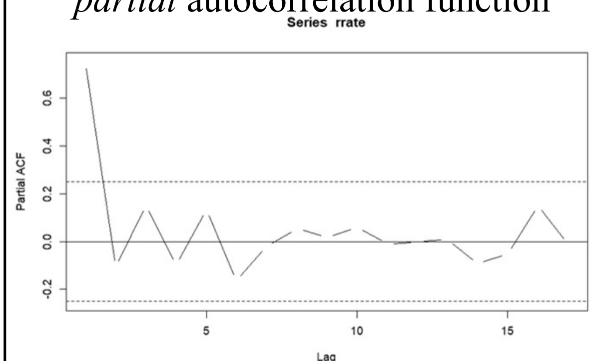
- If autocorrelations decay slowly, check the *partial* autocorrelation function (PAC).
- The first PAC is the coefficient of y_{t-1} in $y_t = \beta_1 + \phi_1 y_{t-1} + \varepsilon_t$.
- The second PAC is the coefficient of y_{t-2} in $y_t = \beta_1 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$.
- Etc.

Interpreting the PAC

- PAC for an AR(1) process drops abruptly to zero after 1st partial autocorrelation.
- PACs for an AR(p) process drop abruptly after pth partial autocorrelation.
- PACs decay slowly for ARMA(p,q).



Example: real interest rate *partial* autocorrelation function



(3) Choose model with lowest AIC

	(1)	(2)	(3)	(4)	(5)	(6)
ar1	0.824***	0.895***			0.758***	0.026
	(0.078)	(0.127)			(0.128)	(0.108)
ar2		-0.093				0.621***
		(0.133)				(0.110)
ma1			0.799***	0.947***	0.171	1.000***
			(0.075)	(0.114)	(0.211)	(0.067)
ma2				0.260***		
				(0.097)		
intercept	0.697	0.769	1.023***	1.012**	0.794	0.716
	(1.023)	(0.918)	(0.387)	(0.450)	(0.885)	(0.968)
Observations	61	61	61	61	61	61
Log Likelihood	-111.425	-111.178	-119.138	-116.020	-111.069	-108.826
sigma2	2.218	2.201	2.862	2.586	2.193	1.963
Akaike Inf. Crit.	**228.849**	**230.356**	**244.276**	**240.039**	**230.137**	**227.651**

Note: *p<0.1; **p<0.05; ***p<0.01

Conclusions

- An ARMA(p,q) process combines an AR(p) process with a MA(q) process, and is defined as
$$y_t = \beta_1 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q} .$$
- To determine p and q,
  1) Plot autocorrelation and partial autocorrelation functions of the data (__________________).
  2) Check t-statistics of estimated coefficients.
  3) Check AIC values of estimated models.

HIGHLY PERSISTENT TIME SERIES

HIGHLY PERSISTENT TIME SERIES

- What kinds of time series are NOT weakly dependent?
- What is a "random walk"?

Highly persistent series

- Highly persistent (or strongly dependent) time series show some sort of dependence between y_t and y_{t+h} that does ________ disappear as h increases.
- When these variables are used in regression analysis, the LS properties of consistency and asymptotic normality do ________ necessarily apply.

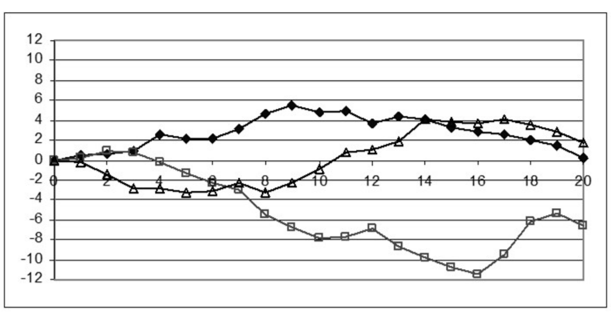
Random walk process

- A simple example of a highly persistent series is the *random walk process*:
  $$y_t = y_{t-1} + \varepsilon_t$$
  where ε_t is an independent identically-distributed series with $E(\varepsilon_t)=0$ and $Var(\varepsilon_t)=\sigma_\varepsilon^2$ (constant).
- This is like an AR(1) process with ρ = ___.

Behavior of a random walk process

- As the name suggests, a random walk process wanders randomly.
- Each value y_t is equal to the prior value y_{t-1} plus a random "step" ε_t.
- Thus each value y_t is just an accumulation of random steps, some positive and some negative, from a starting value (y_0):
  $$y_t = y_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t.$$

Three realizations of a random walk with $\varepsilon_t \sim N(0,1)$ and $y_0 = 0$



Mean of random walk is constant

- $E(y_t) = E(y_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t)$
  $= E(y_0) + E(\varepsilon_1) + E(\varepsilon_2) + \ldots + E(\varepsilon_t)$
  $= E(y_0) + 0 + 0 + \ldots + 0$
  $= E(y_0)$.
- So $E(y_t)$ is _______________ for all *t*.

HIGHLY PERSISTENT TIME SERIES

Variance of random walk is ever-increasing

- $Var(y_t) = Var(y_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t)$
  $= Var(y_0) + Var(\varepsilon_1) + Var(\varepsilon_2) + \ldots + Var(\varepsilon_t)$
  $= Var(y_0) + t\,\sigma_\varepsilon^2$.
- Usually it is assumed that y_0 is nonrandom, in which case $Var(y_t) = t\,\sigma_\varepsilon^2$.
- Because a random walk's variance increases with *t*, a random walk process is _______ stationary.

The persistent effects of each y_t on future y's

- For any positive integer *h*, we can write
  $y_{t+h} = y_t + \varepsilon_{t+1} + \varepsilon_{t+2} + \ldots + \varepsilon_{t+h}$.
- So $E(y_{t+h}|y_t)$
  $= y_t + E(\varepsilon_{t+1}) + E(\varepsilon_{t+2}) + \ldots + E(\varepsilon_{t+h})$
  = _____.
- Contrast this with an AR(1) process, for which $E(y_{t+h}|y_t)$ gradually decays back to its unconditional mean $E(y_{t+h})$. An AR(1) process does ________ wander off.

Unit root process

- A random walk is special case of a *unit root process.*
  - Name comes from AR(1) with ρ=1.
- Any unit root process can also be expressed as $y_t = y_{t-1} + \varepsilon_t$, but now ε_t need not be independent and need not have $E(\varepsilon_t)=0$.
- Instead, ε_t can be *any* weakly dependent process. For example ε_t itself could be AR(1) or MA(1).

Properties of a unit root process

- The random walk is just one example of a unit-root process.
- Other unit-root processes have different formulas for $E(y_t)$ and $Var(y_t)$.
- However, all unit root processes are highly persistent (or strongly dependent).
- The effect of y_t on future y_{t+h} does _______ disappear as $h \to$ infinity.

Random walk with drift

- *Random walk with drift:*
  $$y_t = \beta_1 + y_{t-1} + \varepsilon_t.$$
  where β_1 is a constant and ε_t is an independent identically-distributed series with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_\varepsilon^2$.
- Here, β_1 is called the "drift term."
- On average, y_t increases by _____ from one period to the next.

Mean of random walk with drift is not constant

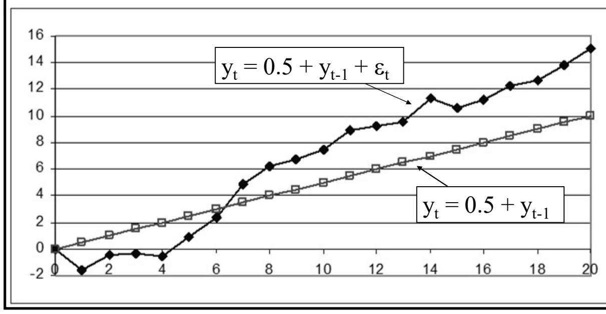
- Here, y_t is an accumulation of random steps, plus constant steps, from a starting value: $y_t = y_0 + \beta_1 t + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t$.
- Thus $E(y_t) = y_0 + \beta_1 t$ (if y_0 is nonrandom).
- Also, $E(y_{t+h}|y_t) = y_t + \beta_1 h$.

HIGHLY PERSISTENT TIME SERIES

Variance of random walk with drift is ever-increasing

- $\mathrm{Var}(y_t) = \mathrm{Var}(y_0 + \beta_1 t + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t)$
  $= \mathrm{Var}(y_0) + \mathrm{Var}(\alpha t) + \mathrm{Var}(\varepsilon_1) + \mathrm{Var}(\varepsilon_2) + \ldots + \mathrm{Var}(\varepsilon_t)$
  $= \mathrm{Var}(y_0) + t\,\sigma_\varepsilon^2$.
- Ever-increasing, like a random walk without drift.
- Usually it is assumed that y_0 is nonrandom (sometimes 0) in which case $\mathrm{Var}(y_t)$ = _______.
- Because the mean and variance of a random walk with drift depend on *t*, it is ________ stationary.

One realization of a random walk with drift: $y_t = 0.5 + y_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim N(0,1)$ and $y_0 = 0$

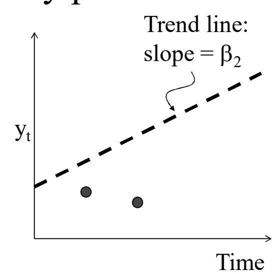


Trends versus unit roots

- Unit-root series, such as random walks, are "highly persistent" or "strongly dependent."
- They do not revert to any fixed mean value.
- But there are other series which are weakly dependent and yet have the same property of not reverting to any fixed value.

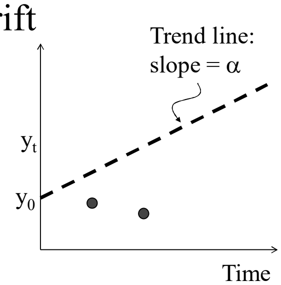
Trend-stationary process

- Example: $y_t = \beta_1 + \beta_2 t + \varepsilon_t$ where ε_t is weakly dependent.
- It does not revert back to any fixed value.
- But it keeps reverting back to $\beta_1 + \beta_2 t$, its _______________.

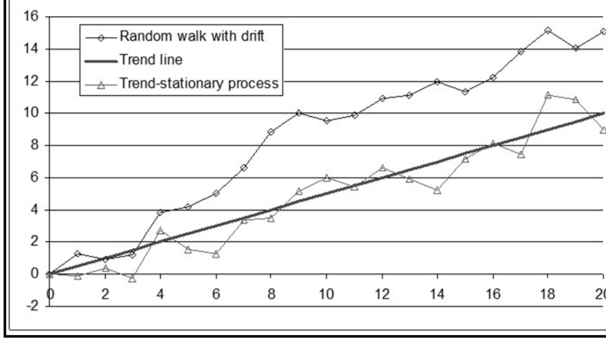


Contrast with random walk with drift

- A random walk with drift is given by: $y_t = \beta_1 + y_{t-1} + \varepsilon_t$.
- It also has a "trend": $y_0 + \beta_1 t$.
- But it gradually _______________ from its trend.



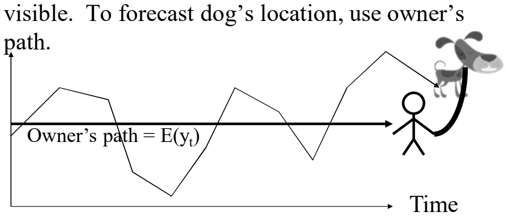
Random walk with drift versus trend-stationary process: examples



HIGHLY PERSISTENT TIME SERIES

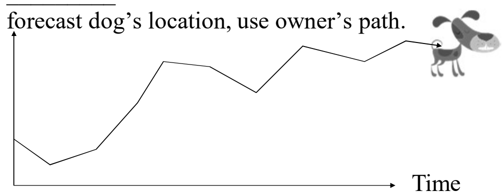
Stationary weakly-dependent process: intuition

- If y_t is stationary and weakly-dependent, it keeps reverting back to its mean.
- Like a dog on a leash, whose owner's path is visible. To forecast dog's location, use owner's path.



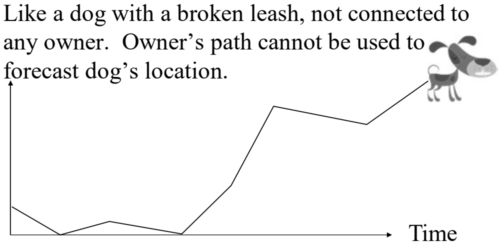
Trend-stationary process: intuition

- If y_t is trend-stationary, it keeps reverting back to a fixed path, which can be estimated.
- Like a dog on a leash, whose owner's path is _________ visible but can be inferred. To forecast dog's location, use owner's path.



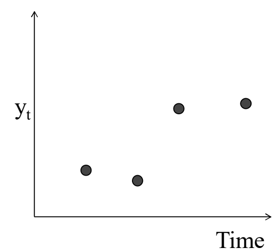
Unit-root process: intuition

- If y_t has a unit root, it wanders randomly over time, sometimes far from its mean.
- Like a dog with a broken leash, not connected to any owner. Owner's path cannot be used to forecast dog's location.



Trend-stationary or unit-root? Hard to tell from data!

- Unfortunately, it is hard to tell if a process is trend-stationary or unit-root just by looking at data.
- Without seeing the trend line, it is difficult to tell whether the series even has one.



Why unit roots matter for forecasting

- If a process has a unit root, it gradually wanders randomly away.
- A random walk wanders away randomly from its initial value _______.
- A random walk with drift wanders away randomly from its trend ___________.

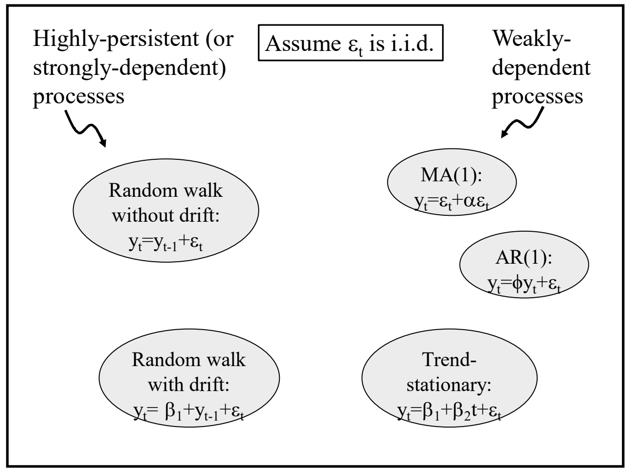
Why unit roots matter for forecasting (cont'd)

- So it is very difficult to forecast a unit-root process, even if we know its trend.
- __________-term forecasts are completely unreliable.
- Only very ________-term forecasts are reliable but they require very recent data.
- Example: stock prices.

HIGHLY PERSISTENT TIME SERIES

Why unit roots matter for economic policy

- If GDP has a unit root, for example, then any changes in GDP persist ____________.
- The economy is "permanently scarred" by recessions and "permanently strengthened" by booms.
- By contrast, if GDP is ________________, then the effects of recessions and booms eventually disappear.



Conclusions

- Highly persistent series show dependence between y_t and y_{t+h} that does ________ disappear as h increases.
- Examples include the *random walk* and the *random walk with drift*.
- *Unit root series* are a broad class of highly persistent series.
- Unit root series can look like _________________ series but they are much more difficult to forecast.

RANDOM WALK

RANDOM WALK

- How can we recognize, estimate, and forecast a random walk process?

Random walk and random walk with drift

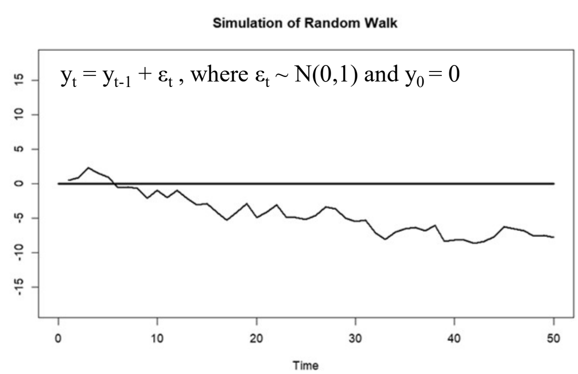
- Let ε_t be an independent identically-distributed (IID) series with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_\varepsilon^2$ (constant).
- Simple random walk: $y_t = y_{t-1} + \varepsilon_t$.
- Random walk with drift: $y_t = \beta_1 + y_{t-1} + \varepsilon_t$.

How can we recognize a random walk or a random walk with drift?

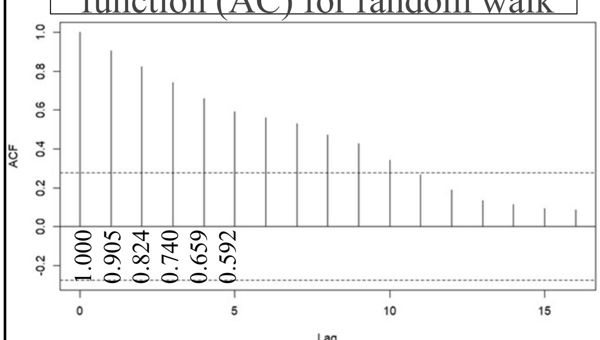
(1) Plot autocorrelation function. If series is nonstationary, will be very high and decrease slowly.

(2) Formal test: Dickey-Fuller.

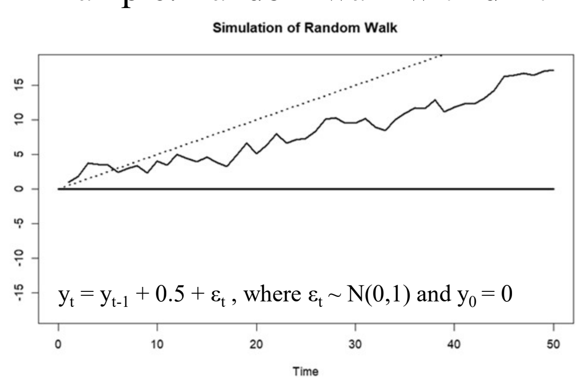
Example: simple random walk



Example: autocorrelation function (AC) for random walk

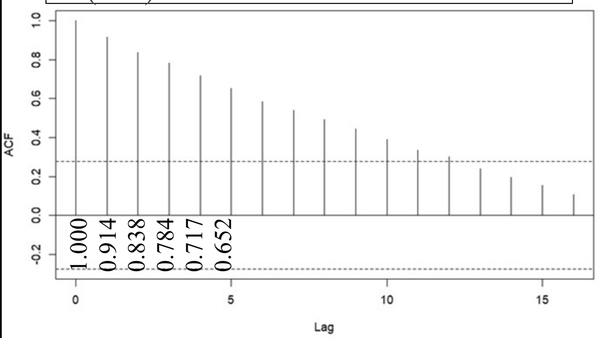


Example: random walk with drift



RANDOM WALK

Example: autocorrelation function (AC) for random walk with drift



Caution about AC plots

- Estimated autocorrelations are biased down if true autocorrelations are high.
- So suspect random walk if first autocorrelation > 0.85 or 0.90.
- However, *trend stationary* processes also produce very high autocorrelations that decrease slowly.

Dickey-Fuller test

- Consider $y_t = \beta_1 + \beta_2 t + \beta_3 y_{t-1} + \varepsilon_t$.
  - Simple random walk: $0 = \beta_1 = \beta_2$ and $\beta_3 = 1$.
  - Random walk with drift: $0 = \beta_2$ and $\beta_3 = 1$.
  - Trend stationary process: $0 = \beta_3$.
  - AR(1): $0 = \beta_2$ and $\beta_3 < 1$.
- We seek to test H_0: $\beta_3 = 1$ (random walk).
- But it turns out that LS $\hat{\beta}_3$ is not consistent under H_0.

Dickey-Fuller test (cont'd)

- So instead subtract y_{t-1} from both sides and estimate by LS:
  $\Delta y_t = \beta_1 + \beta_2 t + \gamma\, y_{t-1} + \varepsilon_t$, where $\gamma = (\beta_3 - 1)$.
- Then test H_0: $\gamma = 0$ (nonstationary).
- It turns out LS $\hat{\gamma}$ *is* consistent, but not normally distributed (even asymptotically).
- Dickey and Fuller worked out the distribution and critical values of $\hat{\gamma}$.

Example: Dickey-Fuller test

- For the simple random walk example above, test statistic = -3.155, p-value = 0.111.
  - So cannot reject H_0: $\gamma = 0$ (nonstationary).
- For the random walk with drift example above, test statistic = -2.665, p-value = 0.308.
  - Again, cannot reject H_0: $\gamma = 0$ (nonstationary).

Caution about Dickey-Fuller test

- Not a powerful test.
- If series is actually stationary, test often still fails to reject H_0: $\gamma = 0$ (nonstationary).

RANDOM WALK

How can we estimate a random walk model?

- Random walk with drift:
  $y_t = \beta_1 + y_{t-1} + \varepsilon_t$.
- Subtract y_{t-1} from both sides:
  $\Delta y_t = y_t - y_{t-1} = \beta_1 + \varepsilon_t$.
- $\hat{\beta}_1$= sample mean of Δy_t .
  $\hat{\sigma}_\varepsilon^2$= sample variance of Δy_t .
- For simple random walk, $\hat{\beta}_1$= 0.

How can we forecast a simple random walk?

- Suppose at time t we want to forecast y_{t+1} .
- Now, $y_{t+1} = y_t + \varepsilon_{t+1}$.
- At time t we know y_t but not ε_{t+1} .
- Our best forecast is the conditional mean
  $E(y_{t+1}|y_t) = y_t + 0 = y_t$.

How can we forecast a simple random walk? (cont'd)

- Suppose at time t we want to forecast y_{t+h} .
- Similarly, $y_{t+h} = y_t + \varepsilon_{t+1} + ... + \varepsilon_{t+h}$.
- Best forecast is conditional mean
  $E(y_{t+h}|y_t) = y_t + E(\varepsilon_{t+1}) + ... + E(\varepsilon_{t+h})$
  $= y_t + 0 + ... + 0 = y_t$.
- At time t, our best forecast of y_{t+h} is simply the current value y_t , no matter how far we look into the future.

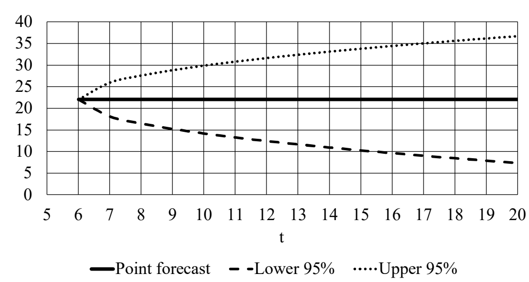
Example: forecasting a simple random walk

- Suppose at time $t = 6$, $y_6 = 22$.
- Best forecast of y_7 = _______.
- Best forecast of y_{17} = _______.
- Best forecast of y_{107} = _______.

Forecast interval for simple random walk

- Variance of forecast error = $Var(y_{t+h}|y_t)$
  $= 0 + Var(\varepsilon_{t+1}) + ... + Var(\varepsilon_{t+h})$
  $= 0 + h\,\sigma_\varepsilon^2$.
- This can be estimated as $(h\,\hat{\sigma}_\varepsilon^2)$.
- Example: suppose at time $t = 6$, $y_6 = 22$, and $\hat{\sigma}_\varepsilon^2 = 4$,
- Then the 95% forecast interval at time t+h is $22 \pm 1.96\sqrt{h\,4}$.

Example: forecasting a simple random walk



RANDOM WALK

How can we forecast a random walk with drift?

- Suppose at time t we want to forecast y_{t+1} .
- Now, $y_{t+1} = \beta_1 + y_t + \varepsilon_{t+1}$.
- At time t we know y_t but not ε_{t+1} .
- Best forecast is conditional mean $E(y_{t+1}|y_t) = \beta_1 + y_t + E(\varepsilon_{t+1})$
  $= \beta_1 + y_t + 0 = y_t + \beta_1$.

How can we forecast a random walk with drift? (cont'd)

- Suppose at time t we want to forecast y_{t+h} .
- Now, $y_{t+2} = \beta_1 + (\beta_1 + y_t + \varepsilon_{t+1}) + \varepsilon_{t+2}$
  $= y_t + 2\beta_1 + \varepsilon_{t+1} + \varepsilon_{t+2}$.
- Similarly, $y_{t+h} = y_t + h\beta_1 + \varepsilon_{t+1} + ... + \varepsilon_{t+h}$.
- Best forecast is conditional mean $E(y_{t+h}|y_t) = y_t + h\beta_1 + E(\varepsilon_{t+1}) + ... + E(\varepsilon_{t+h})$
  $= y_t + h\beta_1 + 0 + ... + 0 = y_t + h\beta_1$
- Estimate as $y_t + h\hat{\beta}_1$, a line.

Example: forecasting a random walk with drift

- For example, suppose we have a random walk with drift: $y_t = \beta_1 + y_{t-1} + \varepsilon_t$.
- Suppose at time $t = 6$, $y_6 = 22$, $\hat{\beta}_1 = 2$.
- Best forecast of $y_7 = 22 + 2 =$ _______.
- Best forecast of $y_{16} = 22 + 2(10) =$ _______.

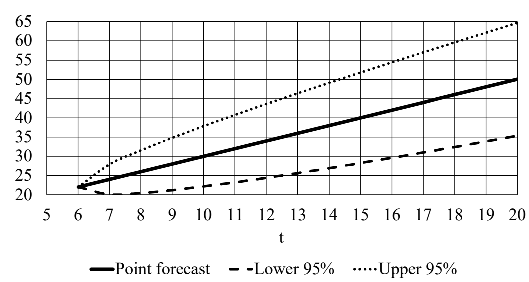
Forecast interval for random walk with drift

- Variance of forecast error = $Var(y_{t+h}|y_t)$
  $= 0 + Var(h\hat{\beta}_1) + Var(\varepsilon_{t+1}) + ... + Var(\varepsilon_{t+h})$
  $= 0 + Var(h\hat{\beta}_1) + h\sigma_\varepsilon^2$.
- $Var(h\hat{\beta}_1)$ is typically small and ignored in practice.
- So variance of forecast error estimated as $(h\hat{\sigma}_\varepsilon^2)$.

Example: forecast interval for random walk with drift

- Example: suppose at time $t = 6$, $y_6 = 22$, $\hat{\beta}_1 = 2$, and $\hat{\sigma}_\varepsilon^2 = 4$.
- Then the 95% forecast interval at time t+h is $(22 + 2h) \pm 1.96\sqrt{h\,4}$.
- Forecast interval does _______ converge, unlike ARMA(p,q).

Example: forecasting a random walk with drift



RANDOM WALK

Conclusions

- Random walks can be distinguished from stationary processes using the autocorrelation function plot or a Dickey–Fuller test.
- The intercept β_1 and the error-term variance σ^2 are estimated after differencing the data.
- Point forecasts are simple to compute.
- But forecast intervals do not converge, so long-term forecasts are not reliable.

ARIMA(p,d,q) PROCESS

ARIMA(p,d,q) PROCESS

- What is an ARIMA(p,d,q) process?
- How can we determine p, d, and q?

Definition of ARIMA(p, d, q) process

- A time series y_t follows an ARIMA(p,d,q) process if when y_t is differenced d times, it follows an ARMA(p,q) process.
- Flexible framework for modeling stationary or nonstationary time series (depending on d), with or without serial correlation (depending on p and q).

Differencing notation

- First differences: $\Delta y_t = y_t - y_{t-1}$.
- Second differences:
  $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$
  $= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$
  $= y_t - 2\, y_{t-1} + y_{t-2}$.
- Third differencing is possible in theory but never necessary in practice.

ARIMA(p,d,q) process: examples

- ARIMA(___________):
  $\Delta y_t = \beta_1 + \phi_1\, \Delta y_{t-1} + \phi_2\, \Delta y_{t-2} + \varepsilon_t + \alpha_1\, \varepsilon_{t-1}$.
- ARIMA(____________):
  $y_t = \beta_1 + \phi_1\, y_{t-1} + \phi_2\, y_{t-2} + \varepsilon_t + \alpha_1\, \varepsilon_{t-1} + \alpha_2\, \varepsilon_{t-2}$.
- ARIMA(____________):
  $\Delta^2 y_t = \beta_1 + \phi_1\, \Delta^2 y_{t-1} + \varepsilon_t + \alpha_1\, \varepsilon_{t-1} + \alpha_2\, \varepsilon_{t-2}$.

Box-Jenkins method

(1) *"Identification":* Determine degree of differencing needed to achieve stationarity. Then determine p and q for differenced series.

(2) *Estimation:* Estimate β_1 , ϕs, and αs.

(3) *Forecasting:* Use $\hat{\beta}_1$, $\hat{\varphi}$s, and $\hat{\alpha}$s to compute point forecasts and intervals.

G.E.P. Box and G.M. Jenkins, *Time Series Analysis, Forecasting, and Control,* Holden-Day, 1976, pp. 173-186.

(1) Identification: determine degree of differencing

- Plot autocorrelation (AC) function. If series is nonstationary, AC will be very high and decrease slowly.
- Formal test: Augmented Dickey-Fuller test.*
- If series seems nonstationary, compute first differences Δy_t and repeat.

* "Augmented" with extra lags to accommodate possible serial correlation.

ARIMA(p,d,q) PROCESS

(1) Identification: determine p and q for differenced series

- Plot autocorrelation (AC) and partial autocorrelation (PAC) functions.
- If AC drops off abruptly after n lags, then p = 0 and q = n.
- If PAC drops off abruptly after n lags, then p = n and q = 0.
- If neither of the above, then p>0 and q>0.

(1) Identification: determine p and q for differenced series

- Alternatively, estimate all reasonable combinations of p and q. Choose model with lowest AIC.
- In practice, no reason for p or q greater than 3.*
- If two models seem to fit the data equally well, choose the simpler model.

*Except for seasonal effects, not covered here.

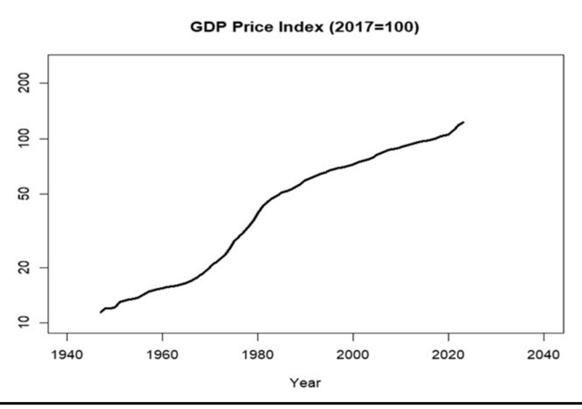
(2) Estimation: $\hat{\beta}_1$, $\hat{\varphi}$s, and $\hat{\alpha}$s

- Complicated, but statistical software handles this.
- If d = 0 (stationary), same as ARMA process.
- If d > 0 (nonstationary), then β_1 is often assumed to be ___________.

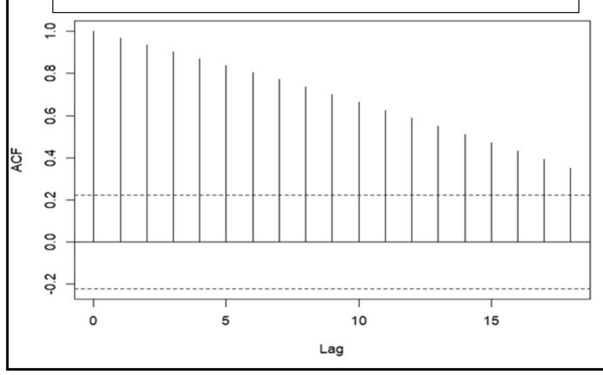
(3) Forecasting: y_{T+1}, y_{T+2}, etc.

- If d = 0 (stationary, ARMA) then
  - Point forecasts converge to the sample mean.
  - Forecast intervals are bounded.
- If d > 0 (nonstationary) then
  - Point forecasts converge to a line.
  - Forecast intervals do not converge.

Example: raw data

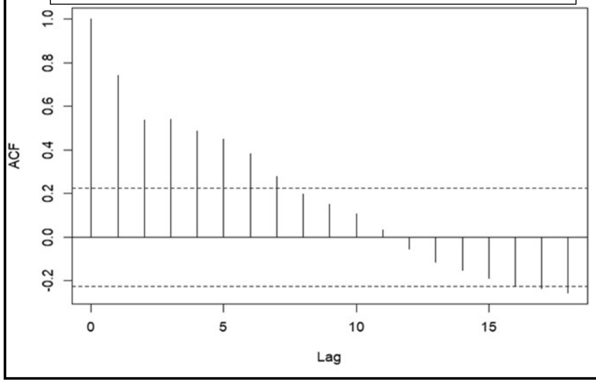


Example: AC plot (after taking logs)

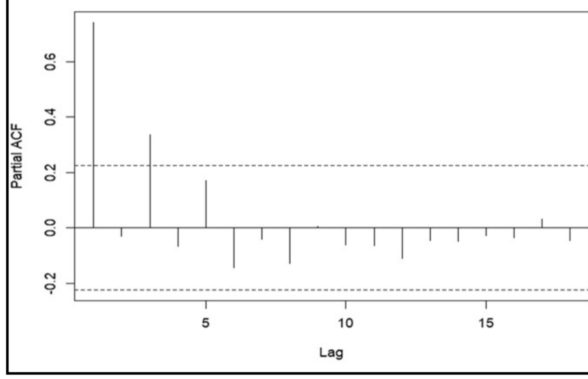


ARIMA(p,d,q) PROCESS

Example: AC plot
(after taking logs and differencing)



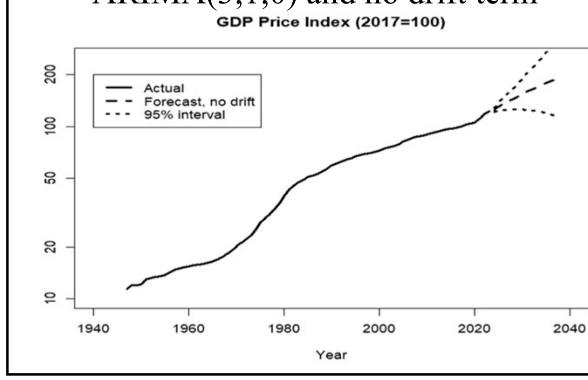
Example: PAC plot
(after taking logs and differencing)



Some models with d=1 and no drift terms

	(1)	(2)	(3)	(4)
ar1	0.924***	0.918***	0.960***	0.851***
	(0.042)	(0.129)	(0.114)	(0.074)
ar2		0.006	-0.542***	
		(0.132)	(0.164)	
ar3			0.542***	
			(0.114)	
ma1				0.434**
				(0.200)
Observations	76	76	76	76
Log Likelihood	210.677	210.678	220.149	211.071
sigma2	0.0002	0.0002	0.0002	0.0002
Akaike Inf. Crit.	**-417.354**	**-415.357**	**-432.299**	**-416.142**

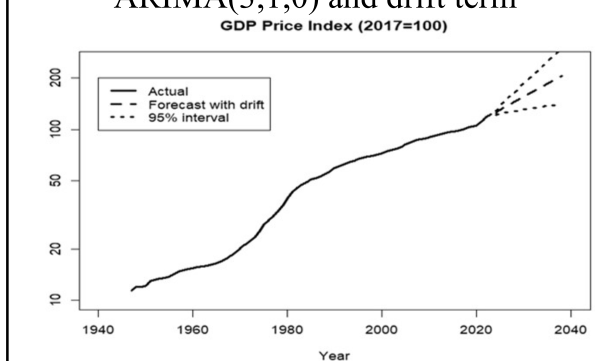
Example: forecast with
ARIMA(3,1,0) and no drift term



Best model with d=1 and drift term

	(1)
ar1	0.895***
	(0.116)
ar2	-0.5426***
	(0.1600)
ar3	0.4744***
	(0.1175)
Drift	0.0325***
	(0.0079)
Observations	76
Log Likelihood	222.68
sigma2	0.0002
Akaike Inf. Crit.	**-435.36**

Example: forecast with
ARIMA(3,1,0) and drift term



ARIMA(p,d,q) PROCESS

Conclusions

- If y_t is nonstationary, and follows an ARMA(p,q) process only after differencing d times, then y_t follows an ARIMA(p,d,q) process.
- Parameter d can be identified from the AC plot and/or an augmented Dicky-Fuller test.
- Parameters p and q can be identified from AC and PAC plots and/or from comparing AIC values.
- If d > 0 (nonstationary) then forecasts converge to a line and forecast intervals do not converge.

SPURIOUS REGRESSION

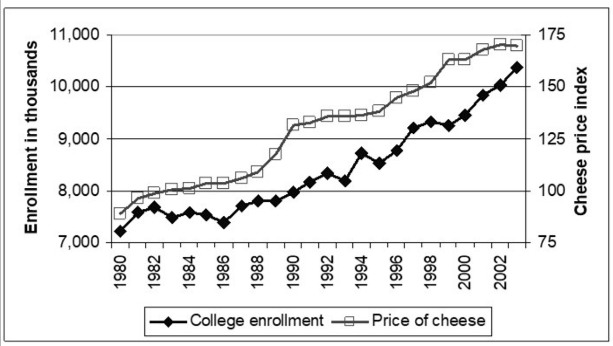
SPURIOUS REGRESSION

- Can we trust regression results from trended series?

Trends in time series

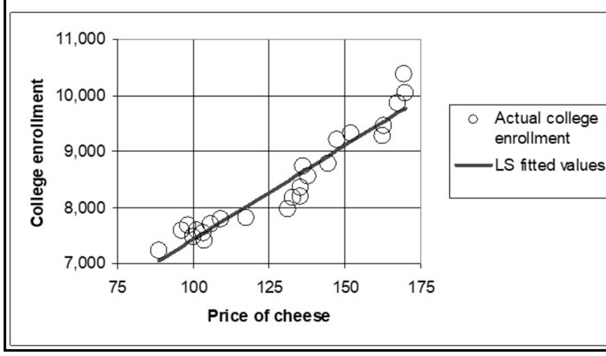
- Many time series show clear upward or downward trends.
- Two trended variables will appear correlated even if they are unrelated in any way: so-called "spurious regression."

Example: two unrelated trended data series



SOURCES: Enrollment: National Center for Educational Statistics, *Indicator 22*. Cheese price index: USDA Economic Research Service, *Consumer Price Index: Cheese* (from Bureau of Labor Statistics).

Example: actual values and least-squares fitted values for enrollment$_t$ = $\beta_1 + \beta_2$ price of cheese$_t$



Example: least-squares estimates for enrollment$_t$ = $\beta_1 + \beta_2$ price of cheese$_t$

- $R^2 = 0.91955$.
- Adjusted $R^2 = 0.91589$.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**4075.26**	**279.03**	**14.60**	**8.4E-13**
Price of cheese	**33.48**	**2.11**	**15.86**	**1.6E-13**

Avoiding spurious regression by controlling for trends

- To investigate whether time series are truly related, we must control for trends.
- This can be done by including a trend as an additional regressor, or by "detrending" the data.

SPURIOUS REGRESSION

"Detrending" the data

- One way to avoid spurious correlation of trended variables is to "detrend" variables before using them in regressions.
- For example, suppose we regress $x_t = \alpha_1 + \alpha_2 t + \varepsilon_t$.
- Residuals from this regression are called "detrended x" because the time trend has been removed from x.

Should variables be "detrended" before use in regression equations?

- Compare the following three regressions.

(1) $y_t = \beta_1 + \beta_2 x_t + \beta_3 t + \varepsilon_t$.

(2) $y_t = \beta_1 + \beta_2 dtx_t + \varepsilon_t$, where dtx_t = detrended x_t.

(3) $dty_t = \beta_1 + \beta_2 dtx_t + \varepsilon_t$, where dtx_t = detrended x_t, and dty_t = detrended y_t .

Should variables be "detrended" before use in regression equations? (cont'd)

- It can be proved that all 3 regressions yield the same estimates of β_1 and β_2 and the same standard errors!
- So including a time trend is equivalent to detrending regression variables.

Example: avoiding spurious regression by including a time trend

- $\text{enrollment}_t = \beta_1 + \beta_2 \text{ price of cheese}_t + \beta_3 \text{ trend}_t$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**5197.245**	**1149.049**	**4.523**	**0.000**
Price of cheese	**19.860**	**13.698**	**1.450**	**0.162**
Trend	**51.481**	**51.146**	**1.007**	**0.326**

R^2 in time-series regressions

- R^2 and adjusted R^2 values are often *very high* in time series regressions, for two reasons.

(1) Often the dependent variable is less "noisy" than in cross section data. Economy-wide averages or totals (typical of time-series data) are often easier to explain than individual firms and consumers (typical of cross-section data).

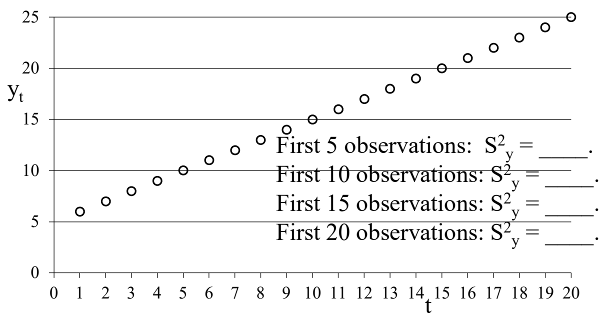
(2) The dependent variable is often ____________.

Why trends raise R^2 and adjusted R^2

- If y_t is trended, R^2 (and adjusted R^2) tend to increase with sample size (denoted T).
- To see this, assume y_t is trended and the variance of the error term is constant.
- Then $\left(\frac{1}{T-K}\right)\sum \hat{\varepsilon}^2 \xrightarrow{P} \sigma^2$, a constant.
- But $\left(\frac{1}{T-1}\right)\sum \left(y_i - \bar{y}\right)^2$ grows without bound.

SPURIOUS REGRESSION

Sample variance (S^2_y) of a trended variable grows without bound: example



Why trends raise R^2 (cont'd)

- Now consider the second term of Theil's adjusted R^2.

$$\overline{R}^2 = 1 - \frac{\left(\frac{1}{T-K}\right)\sum \hat{\varepsilon}^2}{\left(\frac{1}{T-1}\right)\sum \left(y_i - \overline{y}\right)^2}$$

- Clearly the second term must approach zero, so the adjusted R^2 must approach one.
- Wooldridge proposes that a better, more honest R^2 be computed from a regression on detrended variables, but this is rarely done.

Conclusions

- Many time series show clear linear or exponential time trends.
- Unrelated series may appear correlated if both have trends, causing ___________ regression.
- ___________ regression can be avoided if a time trend is included as a regressor.
  - This is equivalent to detrending the variables.
- R^2 and adjusted R^2 are often very ________ if the dependent variable is trended.