

MIDTERM EXAMINATION #2 VERSION A
“Two-Variable Regression”
February 26, 2010

INSTRUCTIONS: This exam is closed-book, closed-notes. You may use a calculator on this exam, but not a graphing calculator or a calculator with alphabetical keys. Point values for each question are noted in brackets. A table of the t-distribution is attached.

NOTATION: In this exam, $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the least-squares estimators of the intercept and slope of the line $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, \hat{y}_i denotes a least-squares fitted value, $\hat{\varepsilon}_i$ denotes a least-squares residual, and the sample size is denoted n . The true or population value of the variance of the unobserved error term ε_i is denoted σ^2 . The (unbiased) least-squares estimate of σ^2 is denoted $\hat{\sigma}^2$. The sample means of x and y are denoted \bar{x} and \bar{y} respectively. The natural logarithm function is denoted $\ln(\cdot)$.

I. MULTIPLE CHOICE: Circle the one best answer to each question. Feel free to use margins for scratch work [2 pts each—10 pts total]

(1) Suppose we wish to fit the equation $y = \beta_1 + \beta_2 x$ to data by the method of least squares. This method minimizes which function of the data?

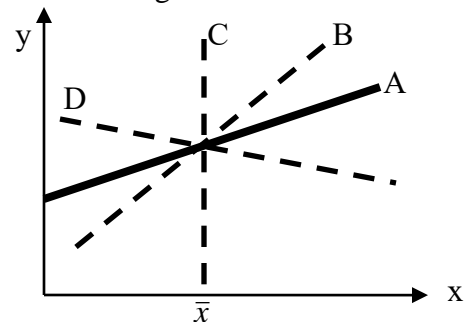
- a. $f(\beta_1, \beta_2) = \sum (\beta_1 + \beta_2 x_i)^2$.
- b. $f(\beta_1, \beta_2) = \sum (y_i - \beta_1 - \beta_2 x_i)$.
- c. $f(\beta_1, \beta_2) = \sum (y_i^2 - (\beta_1 + \beta_2 x_i)^2)$.
- d. $f(\beta_1, \beta_2) = \sum (y_i - \beta_1 - \beta_2 x_i)^2$.
- e. $f(\beta_1, \beta_2) = \sum |y_i - \beta_1 - \beta_2 x_i|$.

(2) In the model $y_i = 2.7 - 1.3 x_i + \varepsilon_i$, assuming $E(\varepsilon_i|x_i)=0$, the conditional mean of y (that is, $E(y_i|x_i)$) is

- a. $-1.3 x_i$.
- b. $2.7 - 1.3 x_i$.
- c. zero.
- d. 2.7.
- e. -1.3 .

(3) In the graph below, the solid line denoted "A" is the true population regression line. If the error term has mean zero but is *negatively correlated* with x , then the least-squares estimated line will tend to resemble

- a. line A.
- b. line B.
- c. line C.
- d. line D.
- e. cannot be determined from the information given.



- (4) In a time-series dataset, any two trended variables must be
- related through cause and effect.
 - correlated with each other.
 - uncorrelated with each other.
 - normally-distributed.
 - caused by a third variable.

- (5) In the equation $\ln(y) = 7 + 0.3 \ln(x)$, which of the following is correct?
- If x increases by 1 unit, then y increases by 0.3 units.
 - If x increases by 1%, then y increases by 0.3 units.
 - If x increases by 1 unit then y increases by 0.3%.
 - The elasticity of y with respect to x is 0.3.

II. SHORT ANSWER: Please write your answers in the boxes on this question sheet. Use margins for scratch work.

(1) [Algebraic properties: 3 pts] Which equations hold necessarily, regardless of the data or the model? Write "TRUE" or "FALSE" in the boxes below.

a. $\sum x_i \hat{\epsilon}_i = 0$

b. $\sum (x_i - \bar{x}) = 0$

b. $\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2 = \sum \hat{\epsilon}_i^2$

(2) [Algebraic properties: 2 pts] Suppose least-squares estimation of $y = \beta_1 + \beta_2 x$ yields a sum of squared residuals $\sum \hat{\epsilon}_i^2 = 14$ while the total sum of squares is $\sum (y_i - \bar{y})^2 = 20$.

a. Compute the explained or regression sum of squares $\sum (\hat{y}_i - \bar{y})^2$ for this regression equation.

b. Compute the value of r^2 for this regression equation.

(3) [Algebraic properties: 3 pts] Suppose the equation $y = \beta_1 + \beta_2 y$ were estimated by least-squares by mistake. Note that the regressor and the dependent variable are the same. Give numerical answers to the following questions.

a. What would be the least-squares estimate of the intercept, $\hat{\beta}_1$?

b. What would be the least-squares estimate of the slope, $\hat{\beta}_2$?

c. What would be the value of $(\hat{y}_i - y_i)$?

(4) [Properties: 5 pts] Which assumptions are required for the least-squares estimators to be *unbiased* estimators? Write "REQUIRED" or "NOT REQUIRED" in the boxes below.

- a. Homoskedasticity: $\text{Var}(\varepsilon_i) = \sigma^2$, a constant.
- b. Conditional mean of error term is zero: $E(\varepsilon_i|x_i) = 0$.
- c. No autocorrelation: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
- d. Variance of error term is zero: $\text{Var}(\varepsilon_i) = 0$.
- e. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$

(5) [Properties: 5 pts] Which assumptions are required for the least-squares estimators to be *best linear unbiased estimators* (BLUE)? Write "REQUIRED" or "NOT REQUIRED" in the boxes below.

- a. Homoskedasticity: $\text{Var}(\varepsilon_i) = \sigma^2$, a constant.
- b. Conditional mean of error term is zero: $E(\varepsilon_i|x_i) = 0$.
- c. No autocorrelation: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
- d. Variance of error term is zero: $\text{Var}(\varepsilon_i) = 0$.
- e. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$

(6) [Properties: 5 pts] Which assumptions are required for the least-squares estimators to be *maximum likelihood estimators*? Write "REQUIRED" or "NOT REQUIRED" in the boxes below.

- a. Homoskedasticity: $\text{Var}(\varepsilon_i) = \sigma^2$, a constant.
- b. Conditional mean of error term is zero: $E(\varepsilon_i|x_i) = 0$.
- c. No autocorrelation: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
- d. Variance of error term is zero: $\text{Var}(\varepsilon_i) = 0$.
- e. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$

(7) [Variance of LS estimators: 4 pts] Write "TRUE" or "FALSE" in the boxes below. The variance of the least-squares slope estimator $\hat{\beta}_2$ is *smaller*, and thus the true value of β_2 is estimated *more precisely*,

- a. the larger the variance of the error term σ^2 .
- b. the larger the sample variance of x : $\frac{1}{n} \sum (x_i - \bar{x})^2$.
- c. the larger the sample mean of x , that is, \bar{x} .
- d. the larger the sample size n .

(8) [Variance of LS predictor: 4 pts] Write "TRUE" or "FALSE" in the boxes below. Suppose we use the least-squares predictor ($\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$) to predict y_{n+1} . The variance of the prediction error ($y_{n+1} - \hat{y}_{n+1}$) is *larger*, and thus prediction is *less precise*,

- a. the smaller the variance of the error term σ^2 .
- b. the smaller the sample variance of x : $\frac{1}{n} \sum (x_i - \bar{x})^2$.
- c. the smaller the sample size n .
- d. the closer x_{n+1} is to \bar{x} .

(9) [Units of measure: 4 pts] Suppose the relationship between house size and family income is estimated by least squares as follows. The r^2 value is 0.36.

House size in square feet	=	207 (54.2)	+	0.18 (0.05)	Family income in dollars
------------------------------	---	---------------	---	----------------	-----------------------------

Now suppose the data on house size were converted from square feet to square yards, and the equation were re-estimated by least squares. Note that 1 square yard = 9 square feet. Compute the new values of the following items.

- a. New value of least-squares intercept ($\hat{\beta}_1$).
- b. New value of least-squares slope ($\hat{\beta}_2$).
- c. The new r^2 value.
- d. New t-statistic for slope ($\hat{\beta}_2$), for testing H_0 : true slope = 0.

III. PROBLEMS: Write your answers in the boxes on this question sheet. Show your work and circle your final answers.

(1) [LS confidence intervals, tests, elasticity: 24 pts] The relationship between weekly income and weekly spending on food has been estimated for a sample of $n=400$ families. For each family i , let y_i denote spending on food and x_i denote income. The model $y_i = \beta_1 + \beta_2 x_i$ is estimated with the following results. The numbers on top are the least-squares estimates of the intercept and slope, and the numbers at the bottom in parentheses are standard errors.

Spending on food	=	87.5 (20)	+	0.05 (0.02)	Family income
---------------------	---	--------------	---	----------------	------------------

a. [3 pts] Suppose a family has a weekly income of \$800. According to these results, how much would this family spend on food?

b. [3 pts] Suppose a family's income increased by \$200. By how much would its spending on food increase? That is, what is the predicted change Δy when $\Delta x = 200$?

c. [3 pts] Suppose the sample mean income is \$750 and the sample mean spending on food is \$125. (That is, $\bar{y} = 125$. and $\bar{x} = 750$.) Compute the estimated elasticity of spending on food with respect to family income at the sample means.

d. [6 pts] Compute a **95%** confidence interval for the **intercept, β_1** .

e. [9 pts] Test the hypothesis that income has a positive effect on spending on food, against the null hypothesis that income has no effect (a **one-tailed test**) at **5%** significance. Give the value of the test statistic, the critical point(s) from a table, and your conclusion (whether you can reject null hypothesis).

Value of test statistic = _____. Critical point(s) = _____.

Can you reject null hypothesis? _____.

(2) [LS confidence intervals, prediction: 24 pts] The effect of the price of high-speed internet access on penetration is measured using a sample of $n=11$ cities. For each city i , let y_i denote the penetration rate (which varies from zero to 100) and let x_i denote the monthly price of internet access in the same city. The model $y_i = \beta_1 + \beta_2 x_i$ is estimated with the following results. The numbers on top are the least-squares estimates of the intercept and slope, and the numbers at the bottom in parentheses are standard errors. Assume the error term is **normally distributed**.

$y_i = 86 - 0.52 x_i$ <p style="text-align: center;">$(16.5) \qquad (0.30)$</p>
--

a. [3 pts] What are the "degrees of freedom" for these estimates? Give a numerical answer.

b. [6 pts] Compute a **95%** confidence interval for the **slope, β_2** .

We wish to predict penetration rate (y_{n+1}) when the price (x_{n+1}) is \$50. So we first transform the data to simplify calculations.

c. [3 pts] Which variable (x_i , y_i , or both) should be transformed? How?

Suppose the following equation has been estimated on the *transformed data* with the following results. (Numbers on top are the least-squares intercept and slope. Numbers at the bottom in parentheses are standard errors.) The estimated variance of the error term is $\hat{\sigma}^2 = 7.0$.

$\text{new } y_i = 60.0 - 0.52 \text{ new } x_i$ <p style="text-align: center;">$(3.0) \qquad (0.30)$</p>
--

d. [3 pts] Predict the penetration rate (y_{n+1}) when the price (x_{n+1}) is \$50.

e. [3 pts] Compute the standard error of prediction error.

f. [6 pts] Compute a 95% prediction interval for the penetration rate (y_{n+1}) when the price (x_{n+1}) is \$50.

IV. CRITICAL THINKING: [7 pts] To investigate the possible effect of libraries on violent crime, the following regression was estimated using data on U.S. states¹:

$$y_i = -1091. + 11.85 x_i$$

(535) (1.30)

where y_i denotes the number of violent crimes committed in state i in 2007 and x_i denotes the number of libraries in state i in the same year. The estimate for the slope coefficient is significantly greater than zero at 0.1 percent significance. So the number of violent crimes is clearly *positively correlated* with the number of number of libraries across states. Is this evidence that libraries *cause* violent crime? If yes, explain why. If no, explain why not and suggest a better way to estimate the regression equation using these state-level data.

[end of exam]

¹ These are actual least-squares estimates using data from the *Statistical Abstract of the United States*, 2010. $n=50$.