STAT 170 – Regression and Time Series
Drake University, Fall 2024
William M. Boal

Signature: ____________________

Printed name: ____________________

EXAMINATION 2 VERSION B
"Two-Variable Regression"
October 10, 2024

INSTRUCTIONS: This exam is closed-book, closed-notes. You may use a calculator on this exam, but not a graphing calculator, a calculator with alphabetical keys, or a phone. 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(.).

I. MULTIPLE CHOICE: Circle the one best answer to each question. Feel free to use margins for scratch work [1 pt each—9 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. $\sum|y_i - \beta_1 - \beta_2 x_i|$.

b. $\sum(\beta_1 + \beta_2 x_i)^2$.

c. $\sum(y_i - \beta_1 - \beta_2 x_i)$.

d. $\sum\left(y_i^2 - (\beta_1 + \beta_2 x_i)^2\right)$.

e. $\sum(y_i - \beta_1 - \beta_2 x_i)^2$.

(2) Suppose the equation $y = \beta_1 + \beta_2 x$ is fitted to n observations on x and y by the method of least squares. Then the equation

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

holds

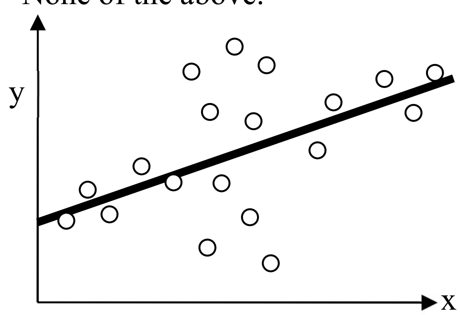
a. if there are enough observations.
b. always.
c. if the fit is good.
d. if the fit is poor.

(3) In a simple regression like $y = \beta_1 + \beta_2 x$, the R-square statistic measures

a. the probability that the regression is correct.
b. the fraction of observations that lie exactly on the fitted line.
c. whether the β_2 coefficient is statistically significant.
d. the squared correlation between the observed values of y and the fitted values $\hat{y}$.

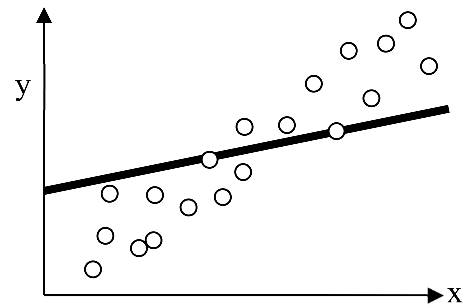
(4) In the graph below, the solid line is the true population regression line and the circles are observations in the sample. Which assumption appears to be violated in this sample?

a. $E(\varepsilon_i|x_i) = 0$.
b. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.
c. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j)=0$ for $i\neq j$.
d. All of the above.
e. None of the above.



(5) In the graph below, the solid line is the true population regression line and the circles are observations in the sample. Which assumption appears to be violated in this sample?

a. $E(\varepsilon_i|x_i) = 0$.
b. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.
c. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j)=0$ for $i\neq j$.
d. All of the above.
e. None of the above.



(6) The LS predictor of $\hat{y}_{n+1}$ given x_{n+1} differs from the actual value y_{n+1} because

a. the LS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ differ from the true parameter values β_1 and β_2 .
b. the actual value y_{n+1} depends in part on a new error term ε_{n+1}.
c. both of the above.
d. They do not differ. The LS predictor equals exactly the actual value y_{n+1} by definition.

(7) If the error term is normally-distributed, then as the sample size n increases, the 95% prediction error converges to

a. zero.
b. one.
c. $\pm$ 1.96.
d. $\pm$ 1.96 times the standard error of the residual.
e. $\pm$ 1.96 times the R-square value.

(8) Suppose we have the supply function

$$\ln(y) = 7 + 0.4 \ln(x),$$

where y denotes the quantity of corn supplied (in millions of tons) and x denotes the price of corn. Which of the following is correct?

a. If price increases by 1 dollar, then quantity supplied increases by 0.4 million tons.
b. If price increases by 1%, then quantity supplied increases by 0.4 million tons.
c. If price increases by 1 dollar then quantity supplied increases by 0.4%.
d. If price increases by 1%, then quantity supplied increases 0.4%.

(9) A *high leverage point* is an observation

a. whose y-value is quite different from the rest of the sample.
b. whose x-value is quite different from the rest of the sample.
c. that contains missing values for x and/or y.
d. that lies exactly on the true regression line.

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

(1) [Algebraic properties: 4 pts] Suppose the equation $y_i = \beta_1 + \beta_2 x_i, + \varepsilon_i$ is fitted by least squares. Which equations hold necessarily, regardless of the data? Write "TRUE" or "FALSE" in the boxes below.

a. $\sum y_i \hat{y}_i = 0$	
b. $\sum \hat{\varepsilon}_i = 0$	
c. $\sum x_i \hat{\varepsilon}_i = 0$	
d. $\sum (x_i - \bar{x}) = 0$	

(2) [Algebraic properties: 4 pts] Suppose least-squares estimation of $y = \beta_1 + \beta_2 x$ yields a sum of squared residuals $\sum \hat{\varepsilon}_i^2 = 12$ while the total sum of squares is $\sum (y_i - \bar{y})^2 = 30$.

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 R^2 ?

(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. Variance of error term is exactly one: $Var(\varepsilon_i) = 1$.

b. Conditional mean of error term is zero: $E(\varepsilon_i|x_i) = 0$.

c. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j) = 0$ for $i \neq j$.

d. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$

e. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.

(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. Variance of error term is exactly one: $Var(\varepsilon_i) = 1$.	
d. Conditional mean of error term is zero: $E(\varepsilon_i \mid x_i) = 0$.	
e. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j) = 0$ for $i \neq j$.	
b. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$	
c. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.	

(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. Variance of error term is exactly one: $Var(\varepsilon_i) = 1$.	
d. Conditional mean of error term is zero: $E(\varepsilon_i \mid x_i) = 0$.	
e. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j) = 0$ for $i \neq j$.	
b. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$	
c. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.	

(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 *larger*, and thus the true value of β_2 is estimated *less precisely* ...

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

(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 *smaller*, and thus prediction is *more precise* ...

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

(9) [Units of measure: 8 pts] Suppose the relationship between electricity price and electricity usage is estimated by least squares as follows. Numbers in parentheses are standard errors. The R^2 value is 0.25 .

Avg usage (kilowatt-hours per month)	=	1392 (71.8)	-	37.0 (5.0)	Price (cents per kilowatt-hour)

Now suppose the data on usage were converted from monthly to annual values—in other words, the usage data were multiplied by 12—and the equation were re-estimated by least squares. 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: 21 pts] Suppose we estimate the effect of temperature on ice cream sales using a sample of $n=300$ days. Let y_i denote daily ice cream sales and x_i denote temperature. The model $y_i = \beta_1 + \beta_2 x_i$ is estimated with the following results. Numbers in parentheses are standard errors.

Ice cream sales	=	65.0 (20.0)	+	22.0 (5.0)	Temperature

a. [3 pts] Suppose the temperature is 80 degrees. According to these results, what are predicted sales?

b. [3 pts] Suppose the temperature increases by 5 degrees. By how much would ice cream sales increase? That is, what is the predicted change Δy when $\Delta x = 5$?

c. [6 pts] Compute a **90%** confidence interval for the **intercept, β_1** .

d. [9 pts] Test the hypothesis that temperature has a **positive** effect on ice cream sales, against the null hypothesis that temperature has no effect (a **one-tailed test**) at **10%** 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] Suppose the relationship between weekly income (x) and weekly food expenditures (y) is estimated using a sample of **17** families as follows. The numbers in parentheses are standard errors. Assume the error term is normally distributed.

y	=	12.9	+	0.18	x
		(29.0)		(0.05)	

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**.

We wish to predict food expenditures (y_{n+1}) when income (x_{n+1}) is **$600**. 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 in parentheses are standard errors. The estimated variance of the error term is $\hat{\sigma}^2 = 682$.

new y_i	=	122.3	+	0.18	new x_i
		(6.5)		(0.05)	

d. [3 pts] Predict food expenditures (y_{n+1}) when income (x_{n+1}) is $600.

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

f. [6 pts] Compute a 95% prediction interval for food expenditures (y_{n+1}) when income (x_{n+1}) is $600.

IV. CRITICAL THINKING: Answer just *one* of the questions below (your choice). [4 pts]

(1) Why should we ever care whether the error term is distributed as normal, since we can compute asymptotic t-statistics and confidence intervals without that assumption? Give two reasons.

(2) Consider the following statement: "Ordinary least squares is limited to fitting linear relationships between variables." Do you agree or disagree? Justify your answer with at least two examples.

[end of exam]