

**MIDTERM EXAMINATION #2 VERSION B**  
**“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  $\varepsilon_i$  is denoted  $\sigma^2$ . The (unbiased) least-squares estimate of  $\sigma^2$  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?

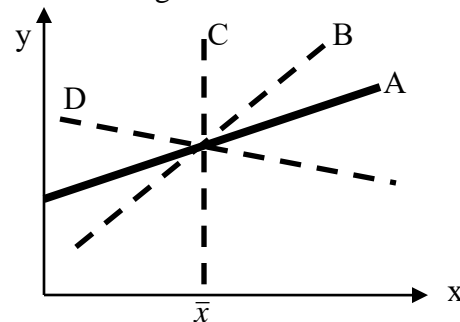
- a.  $f(\beta_1, \beta_2) = \sum |y_i - \beta_1 - \beta_2 x_i|$ .
- b.  $f(\beta_1, \beta_2) = \sum (\beta_1 + \beta_2 x_i)^2$ .
- c.  $f(\beta_1, \beta_2) = \sum (y_i - \beta_1 - \beta_2 x_i)$ .
- d.  $f(\beta_1, \beta_2) = \sum (y_i^2 - (\beta_1 + \beta_2 x_i)^2)$ .
- e.  $f(\beta_1, \beta_2) = \sum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

(2) In the model  $y_i = 4.5 + 2.5 x_i + \varepsilon_i$ , assuming  $E(\varepsilon_i|x_i)=0$ , the conditional mean of  $y$  (that is,  $E(y_i|x_i)$ ) is

- a. 4.5 .
- b. 2.5 .
- c.  $2.5 x_i$  .
- d.  $4.5 + 2.5 x_i$  .
- e. zero.

(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 *positively 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
- normally-distributed.
  - caused by a third variable.
  - related through cause and effect.
  - correlated with each other.
  - uncorrelated with each other.

- (5) In the equation  $y = 5 + 2 \ln(x)$ , which of the following is correct?
- If  $x$  increases by 1 unit, then  $y$  increases by 2 units.
  - If  $x$  increases by 1%, then  $y$  increases by 0.02 units.
  - If  $x$  increases by 1 unit then  $y$  increases by 2%.
  - The elasticity of  $y$  with respect to  $x$  is 2.

**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.  $\sum y_i \hat{\epsilon}_i = 0$

b.  $\sum \hat{\epsilon}_i = 0$

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


(2) [Algebraic properties: 2 pts] Suppose least-squares estimation of  $y = \beta_1 + \beta_2 x$  yields a sum of squared residuals  $\sum \hat{\epsilon}_i^2 = 12$  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  $x = \beta_1 + \beta_2 x$  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 slope,  $\hat{\beta}_2$  ?

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

c. What would be the value of  $r^2$  ?


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

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


(5) [Properties: 5 pts] Which assumptions are required for the usual standard error formulas to be correct for *large samples*? Write "REQUIRED" or "NOT REQUIRED" in the boxes below.

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


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

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


(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  $\beta_2$  is estimated *more precisely*,

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


(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 0.48.

House size in square feet	=	245 (54.2)	+	0.14 (0.05)	Family income in dollars
------------------------------	---	---------------	---	----------------	-----------------------------

Now suppose the data on family income were converted from dollars to thousands of dollars, and the equation were re-estimated by least squares. For example, an observation that was formerly  $x = \$85,000$  would now be converted to  $x = \$85$ . 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	=	120.0 (20)	+	0.04 (0.03)	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  $\Delta y$  when  $\Delta x = 200$ ?

- c. [3 pts] Suppose the sample mean income is \$750 and the sample mean spending on food is \$150. (That is,  $\bar{y} = 150$  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,  $\beta_1$** .

- 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  $n=12$  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**.

$y_i$	=	82	-	0.45	$x_i$
		(16.5)		(0.30)	

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,  $\beta_2$** .

We wish to predict penetration rate ( $y_{n+1}$ ) when the price ( $x_{n+1}$ ) is \$60. 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  $\hat{\sigma}^2 = 9.0$ .

new $y_i$	=	55.0	-	0.45	new $x_i$
		(4.0)		(0.30)	

d. [3 pts] Predict the penetration rate ( $y_{n+1}$ ) when the price ( $x_{n+1}$ ) is \$60.

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 \$60.

**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. states<sup>1</sup>:

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

(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]

---

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