

MIDTERM EXAMINATION #2 VERSION A
“Two-Variable Regression”
March 4, 2008

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.

I. MULTIPLE CHOICE: Circle the one best answer to each question. Feel free to use margins for scratch work [2 pts each—12 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. $f(\beta_1, \beta_2) = \sum (y_i^2 - (\beta_1 + \beta_2 x_i)^2)$.
- c. $f(\beta_1, \beta_2) = \sum (y_i - \beta_1 - \beta_2 x_i)^2$.
- d. $f(\beta_1, \beta_2) = \sum |y_i - \beta_1 - \beta_2 x_i|$.
- e. $f(\beta_1, \beta_2) = \sum (\beta_1 + \beta_2 x_i)^2$.

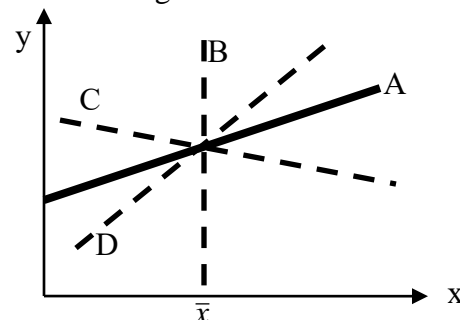
(2) In the model $y_i = 5.2 + 0.7 x_i + \varepsilon_i$, assuming $E(\varepsilon_i|x_i)=0$, the conditional mean of y (that is, $E(y_i|x_i)$) is

- a. zero.
- b. 5.2.
- c. 0.7.
- d. $0.7 x_i$.
- e. $5.2 + 0.7 x_i$.

(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) If the equation $y = \beta_1 + \beta_2 y$ were estimated by mistake, then the least-squares sum of squared residuals would equal

- a. -1.
- b. 0.
- c. 1.
- d. n.
- e. $\sum y_i$.

- a. zero.
- b. one.
- c. σ^2 .
- d. $\text{Var}(\hat{\beta}_1)$.
- e. $\text{Var}(\hat{\beta}_2)$.

(5) The variance of the prediction error tends to decrease as the sample size n used for estimation increases, finally approaching

- a. $y = 4.5 + 0.05 x$.
- b. $y = 4.5 + 0.05 (1/x)$.
- c. $y = 4.5 + 0.05 \ln(x)$.
- d. $\ln(y) = 4.5 + 0.05 x$.
- e. $\ln(y) = 4.5 + 0.05 \ln(x)$.

II. MULTIPLE ANSWER: The questions below may have more than one correct answer. Write “YES” next to all correct answers and “NO” next to all incorrect answers.

(1) [5 pts] Which equations hold necessarily, regardless of the data or the model?

- a. $\sum x_i \hat{\epsilon}_i = 0$
- b. $\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{\epsilon}_i^2$
- c. $\sum \hat{y}_i = 0$
- d. $\sum \hat{\epsilon}_i = 0$
- e. $\sum \hat{\epsilon}_i \hat{y}_i = 0$

(2) [5 pts] Which assumptions are required for the least-squares estimators to be unbiased estimators?

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

(3) [5 pts] Which assumptions are required for the least-squares estimators to be method-of-moments estimators?

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

(4) [5 pts] Which assumptions are required for the least-squares estimators to have the lowest variance of all unbiased estimators ("BUE")?

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

(5) [4 pts] 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 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 smaller the sample mean of x , that is, \bar{x} .

(6) [5 pts] 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 smaller the sample mean of x , that is, \bar{x} .
- e. the closer x_{n+1} is to \bar{x} .

(7) [4 pts] Suppose a demand function for gasoline of the form $y_i = \beta_1 + \beta_2 x_i$, is estimated by least squares. Here, y_i denotes quantity demanded in gallons and x_i denotes the price of gasoline in dollars. Now suppose the quantity data are converted to liters (there are about 3.8 liters in a gallon). How will the least-squares estimates change?

- a. $\hat{\beta}_1$ will increase by a factor of 3.8.
- b. $\hat{\beta}_2$ will increase by a factor of 3.8.
- c. The r^2 value will increase by a factor of 3.8.
- d. The t-statistic for $\hat{\beta}_2$ will increase by a factor of 3.8.

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 average test score and spending per pupil is estimated for a sample of $n=200$ school districts. For each district i , let y_i denote average test score and x_i denote spending per pupil (in thousands of dollars). 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.

Average test score	=	52.56	+	2.40	Spending per pupil in thousands
		(3.20)		(0.80)	

a. [3 pts] Suppose a district spends \$10 thousand per pupil (that is, $x = 10$). According to these results, what would be this district's predicted average test score?

b. [3 pts] Suppose a district increased its spending by \$2 thousand. By how much would its average test score increase? That is, what is the predicted change Δy when $\Delta x = 2$?

c. [3 pts] Suppose the sample mean spending per pupil in this dataset is \$8.1 thousand and the sample mean test score is 72. (That is, $\bar{y} = 72$. and $\bar{x} = 8.1$.) Compute the estimated elasticity of average test score with respect to spending per pupil at the sample means.

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

e. [9 pts] Test the hypothesis that spending has a positive effect on test scores, against the null hypothesis that spending 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) = _____ .

Reject null hypothesis? _____ .

(2) [LS confidence intervals, prediction: 24 pts] The effect of the price of water on daily water consumption per capita is measured using a sample of $n=14$ cities. For each city i , let y_i denote its water consumption per capita and let x_i denote its price of water. 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	=	129.1	-	205	x_i
		(5.8)		(30.0)	

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

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

We wish to predict water consumption per capita (y_{n+1}) when the price (x_{n+1}) is 0.02. So we first transform the data.

- c. [3 pts] Which variable (x_i or y_i) 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 $\hat{\sigma}^2 = 10.56$.

new y_i	=	125.0	-	205	new x_i
		(3.8)		(30.0)	

- d. [3 pts] Predict water consumption per capita (y_{n+1}) when the price (x_{n+1}) is 0.02.

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

- f. [6 pts] Compute a 95% prediction interval for water consumption when the price (x_{n+1}) is 0.02.

IV. CRITICAL THINKING: [7 pts] A researcher investigates the causes of house fires by running the following regression using data on U.S. states: $y_i = \beta_1 + \beta_2 x_i$, where y_i denotes the number of house fires reported in 2007 in state i , and x_i denotes the number of video games sold in state i in the same year. Using least-squares, the researcher finds a positive estimate for β_2 , significantly different from zero at 1 percent. So the number of house fires is clearly *positively correlated* with the number of video games sold across states. Is this evidence that video games *cause* house fires? 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]