Midterm Exam

July 27^{th} 2011

Student name:.....

Instructions:

- Write your name in the space provided above.
- You have 2 hours to complete the exam.
- Please read all question and possible answers carefully before choosing your answer.
- In the multiple choice questions, indicate your answer by circling the letter that corresponds to the correct answer.
- In the very short answers and short answers questions write your answer in the space provided.
- You can earn a total of 100 points:
  - 60 points total for the multiple choice questions section
  - 12 points total for the very short answers questions
  - 28 points total for the short answers questions part

Good luck!

Multiple choice questions (3 points each)

Please circle the answer which you think is correct. There are no negative points, so do not leave any question without an answer.

1. Consider the least square approach (minimizing the sum of squared residuals) to deriving the OLS estimator for the multiple linear regression model. Which assumption is required to obtain the estimator $\hat{\beta} = (X'X)^{-1}X'y$?
(a) Linearity in parameters
(b) Zero conditional mean
(c) Homoskedasticity
(d) No assumptions are required
(e) None of the above

2. Which of the following are algebraic properties of the OLS estimator for the multiple linear regression model?
(a) $\sum_i u_i = 0$ and $\sum_i x_{ij}u_i = 0$ for $j = 1, 2, ..., k$
(b) $E(\beta) = \beta$ and $Var(\beta) = \sigma^2(X'X)^{-1}$
(c) $\sum_i \hat{u}_i = 0$ and $\sum_i x_{ij}\hat{u}_i = 0$ for $j = 1, 2, ..., k$
(d) $E(\hat{\epsilon}_i) = 0$ and $E(x_j, \hat{\epsilon}_i) = 0$ for $j = 1, 2, ..., k$
(e) None of the above

The next two question refer to the Stata output provided in the end of the exam.

3. Compare the two specification. What do the results suggest about the relationship between the size of the house and number of bedrooms.
(a) They are positively correlated.
(b) They are negatively correlated.
(c) It is impossible to separate the effect of house size and number of bedrooms.
(d) Number of bedrooms has a much larger effect than house size on the house price.
(e) None of the above.

4. In the second specification, what is the interpretation of the R^2 goodness-of-fit measure?
(a) Number of bedrooms explain approximately 0.9 percent of the variation in house prices.
(b) Number of bedrooms, house size and lot size jointly explain 67 percent of the variation in house prices.
(c) Number of bedrooms, house size and lot size jointly explain 6.7 percent of the variation in house prices.
(d) The sample size is too small too draw any conclusions about R^2.
(e) None of the above.

5. Consider a linear regression model where x_1 is the variable of interest; that is, you wish to obtain unbiased estimates of the marginal effect of x_1 on y. You believe that the following covariance relationships exist (for the purpose of this question you can think of covariances as correlations):

$$Cov(x_1, x_2) > 0, \quad Cov(x_1, x_3) > 0, \quad Cov(x_1, x_4) = 0, \quad Cov(x_2, x_5) \neq 0,$$
$$Cov(y, x_2) \neq 0, \quad Cov(y, x_3) = 0, \quad Cov(y, x_4) < 0, \quad Cov(y, x_5) \neq 0,$$

Which variables (aside from x_1) should be included in the regression?
(a) x_2, x_3
(b) x_2, x_4
(c) x_3, x_5
(d) x_2, x_3, x_5
(e) x_2, x_5

6. Consider the following multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

Suppose that you do not observe x_2 and that $Cov(x_1, x_2) = 0$ and $Cov(x_3, x_2) \neq 0$ and $Cov(x_1, x_3) \neq 0$. Which coefficients will be biased if x_2 is left out of the estimated equation?
(a) $\hat{\beta}_1$
(b) $\hat{\beta}_2$
(c) $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$
(d) $\hat{\beta}_3$
(e) $\hat{\beta}_3, \hat{\beta}_1, \hat{\beta}_0$

Next two questions refer to the following model of education outcomes. Assume the true model is:

$$TS_i = \beta_0 + \beta_1 EL_i + \beta_2 FI_i + u_i, \quad \text{where } i = 1, 2, ..., n$$

and where TS_i measures average test score, EL_i the proportion of students who are not native English speakers, and FL_i measures the average family income.

7. What is the unit of observation in this model?
(a) individual
(b) school
(c) school district
(d) household
(e) (b) or (c)

8. Now suppose that you know that all of the variables are measured at the level of a school district. There is an evidence that, on average, districts with lower average income have, on average, a higher proportion of English learners and lower education attainment.
However, your data set does not have information on income, so you are forced to estimate:

$$TS_i = \beta_0 + \beta_1 EL_i + u_i$$

Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ denote the OLS estimators from this regression. What can you tell about the bias of the estimator $\tilde{\beta}_1$?
(a) $\tilde{\beta}_1$ is unbiased for β_1.
(b) $\tilde{\beta}_1$ is an upward biased estimator for β_1.
(c) $\tilde{\beta}_1$ is an downward biased estimator for β_1.
(d) $\tilde{\beta}_1$ is an biased estimator for β_1, but there is insufficient information to determine the sign of the bias.
(e) None of the above.

The next two question refer to the following model of wage determination:

$$ln(w_i) = \beta_0 + \beta_1 U_i + \beta_2 s_i + \beta_3 s_i U_i + \epsilon_i$$

where w_i measures hourly wage and s_i years of education, U_i is a dummy variable such that $U_i = 1$ if the individual resides in an urban areas. (In this data set all individuals are recorded as residing in either and urban area or rural area.) The OLS estimates are:

$$\hat{\beta}_1 = 0.76 \qquad \hat{\beta}_2 = 0.73 \qquad \hat{\beta}_3 = 0.09$$

9. What is the expected difference in log wages between an individual residing in an urban area and an individual residing in a rural area, everything else equal?
(a) 0.03
(b) 0.09
(c) 0.76
(d) It depends on years of education.
(e) None of the above.

10. What is the average return to schooling (impact of schooling on wage) for an individual residing in an urban areas?
(a) 0.73
(b) 0.09
(c) 0.82
(d) Cannot be determined
(e) None of the above

11. Consider the following multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

In which of the following situations will you be unable to obtain estimates?
(a) $x_{i3} = 2.46$ for $i = 1, 2, ..., n$
(b) $x_{i2} = 0.7x_{i1} + 0.3x_{i3}$
(c) $x_{i1} = 0.5 + 0.3x_{i2}^2$
(d) (a) and (b)
(e) all of the above.

12. Consider the following model in which years of education are regressed on gender ($F_i = 1$ for female), family income (fi_i), and father's years of education (fs_i).

$$s_i = \beta_0 + \beta_1 F_i + \beta_2 fi_i + \beta_3 fs_i + u_i$$

What is the interpretation of the intercept term?
(a) Average years of education for males.
(b) Average years of education for females.
(c) Predicted years of education when all regressors are at their mean.
(d) It has no meaningful interpretation since father's years of education is unlikely to be equal to zero.
(e) None of the above.

The next two questions refer to the variance of the estimator $\hat{\beta}_j$ in the multiple regression model with $k = 6$ regressors. Assume assumptions A1-A5 are satisfied in the underlying model.

13. Which factors may contribute to a high variance of the estimator $\hat{\beta}_4$, i.e. $j = 4$ in the above expression?
(a) The sample size is large.
(b) There is a high degree of correlation between x_1, x_2 and x_3.
(c) There is a high degree of correlation between x_4, x_5 and x_6.
(d) (a) and (b)
(e) (a) and (c)

14. Which of the following is an unbiased estimator for the variance of the error term, σ^2?
(a) $\hat{\sigma}^2 = \frac{1}{n-7} \sum_i \hat{u}_i^2$
(b) $\hat{\sigma}^2 = \frac{1}{n} \sum_i \hat{u}_i^2$
(c) $\hat{\sigma}^2 = \frac{1}{n} \sum_i u_i^2$
(d) $\hat{\sigma}^2 = \frac{1}{n-7} \sum_i u_i^2$
(e) None of the above.

The next two questions refer to the five assumptions we have seen in class:

- A1: Linearity in parameters
- A2: Random sampling
- A3: Zero Conditional Mean
- A4: Variation in the $x's$
- A5: Homoskedasticity

15. Which statement is the most accurate expression of the result obtained from the Gauss-Markov Theorem:
(a) Under A1-A5, $\hat{\beta}_{OLS}$, has the smallest variance of all unbiased estimators for β.
(b) Under A1-A5, $\hat{\beta}_{OLS}$, has the smallest variance of all linear unbiased estimators for β.
(c) Under A1-A5, $\hat{\beta}_{OLS}$, has the smallest variance of all linear estimators for β.
(d) Under A1-A5, $\hat{\beta}_{OLS}$ is an unbiased estimators for β.
(e) None of the above.

16. Compare the following two expressions for the expected value of the OLS estimator for the multiple linear regression model:

$$E(\hat{\beta}) = E[(X'X)^{-1}X'y] \quad (1)$$
$$E[\hat{\beta}] = \beta \quad (2)$$

Which assumption(s) is crucial to obtain expression (2) from expression (1)?
(a) No assumption is required.
(b) A1 and A2
(c) A3
(d) A5
(e) A4

17. Consider a linear regression model estimating the relationship between two variables, x and y. You are interested in knowing how the change in x affects the change in y in percentage terms. How would you set up the model to get the estimate of the percentage change?
(a) $y_i = \beta_0 + \beta_1 ln(x_i) + u_i$
(b) $ln(y_i) = \beta_0 + \beta_1 ln(x_i) + u_i$
(c) $ln(y_i) = \beta_0 + \beta_1 x_i + u_i$
(d) $y_i = \beta_0 + \beta_1 x_i + u_i$
(e) Cannot obtain estimates of percentage terms directly.

18. Variation in the dependent variable, y, can be decomposed into two components: one is the variation explained by the regression model, and the other is the unexplained variation. The latter is captured by the residuals. How can this decomposition be expressed?
(a) $y_i = \hat{y}_i + \hat{u}_i$
(b) $\sum_i (y_i - \bar{y}_i) = \sum_i (\hat{y}_i - \bar{y}_i) + \sum_i \hat{u}_i$
(c) $\sum_i (y_i - \bar{y}_i) = \sum_i (\hat{y}_i - \bar{y}_i) + \sum_i u_i$
(d) $\sum_i (y_i - \bar{y}_i)^2 = \sum_i (\hat{y}_i - \bar{y}_i)^2 + \sum_i \hat{u}_i^2$
(e) None of the above.

19. You run a regression model in which you regress y on x and obtain a $R^2 = 0.24$. What conclusions can you draw based on this result?
(a) 0.24 percent of variation in y is explained with the variation in x.
(b) 24 percent of variation in y is explained with the variation in x.
(c) The model is bad.
(d) 24 percent of variation in y is explained with the variation in x, therefore too little variation is explained and the model is bad.
(e) Changing x by one unit will, on average, change y by 0.24.

20. Assume that a true model is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

All relevant assumptions are satisfied for this model. However, you do not have information on x_2 and so estimate:

$$y_i = \beta_0 + \beta_1 x_{i1} + v_i$$

You get an estimate, $\hat{\beta}_1 = 1.2$ and believe that $Cov(x_1, x_2) > 0$ and $Cov(y, x_2) < 0$. Can you use this estimate to draw any useful conclusions about the relationship between x_1 and y?
(a) Since $\hat{\beta}_1$ is upward biased we can infer that the marginal effect of x_1 is at least 1.2.
(b) Since $\hat{\beta}_1$ is downward biased we can infer that the marginal effect of x_1 is at least 1.2.
(c) Since $\hat{\beta}_1$ is unbiased we can infer that the marginal effect of x_1 is at least 1.2.
(d) Since $\hat{\beta}_1$ is downward biased we can infer that the marginal effect of x_1 is less than 1.2.
(e) None of the above.

Very short answers questions (4 points each)

1. Suppose the SLR model applies to $y = \beta_0 + \beta_1 x + \epsilon$. The slope coefficient $\hat{\beta}_1$ from regressing x on y is just the inverse of the slope coefficient, $\hat{\alpha}_1$, from regressing y on x. True, false or uncertain? Explain.

2. Suppose that:

$$income = \beta_0 + \beta_1 \text{ experience } + \beta_2 \text{ education } + \beta_3 \text{ sex } + \beta_4 \text{ age } + \epsilon$$

What would you speculate the direction of the bias of the estimate of β_1 to be if *age* were omitted from the regression? Explain your reasoning.

3. Explain in what sense "dropping" a variable can be a solution to multicollinearity?

Short answers questions

1. (8 points) Consider the OLS estimator for β in the SIMPLE regression model (so only one regressor and a constant). Assume that the following assumptions are satisfied:

- A1: Linearity in parameters
- A2: Random sampling
- A3: Zero Conditional Mean
- A4: Variation in the $x's$
- A5: Homoskedasticity

Derive the expression for $Var(\hat{\beta}_1|x)$, indicating where the different assumptions are required for this proof.

2. (10 points) Imagine that you are carrying out a research in the field of education policy. You are investigating whether student attainment could be improved by providing assistance with housing costs, enabling students to live closer to campus. In particular, you want to examine whether the distance from campus effects GPA, and whether this varies between urban and rural areas.

Your dataset contains information on these variables and also gender and measures of previous educational attainment (high school GPA, SAT scores). Discuss how you will set up the model, and how you will interpret the parameter estimates.

3. (10 points) Consider a linear regression model, regressing wage on years of education (s_i) and years of experience (e_i) in the work force. The dataset has a sample size of 500, drawn at random from a large population. You estimate the following regression:

$$ln(w_i) = \beta_0 + \beta_1 s_i + \beta_2 e_i + \beta_3 e_i^2 + u_i$$

and you obtain the given estimates:

$$\hat{\beta}_2 = 0.03 \qquad \hat{\beta}_3 = -0.005$$

(a) What is your interpretation of the relationship between years of experience and wages?

(b) Your dataset also includes information on tenure (years of experience with the current employer). Should this be included in the regression? Why?

(c) Do you believe that this regression, either inducing or excluding tenure as decided in (b), will provide best linear unbiased estimator, i.e. does Gauss-Markov Theorem hold in this estimation scenario? Discuss with reference to the five assumptions we have seen in class.