Ecn 102 - Analysis of Economic Data
University of California - Davis March 17, 2010
Instructor: John Parman

Final Exam

You have until 12:30pm to complete this exam. Please remember to put your name, section and ID number on both your scantron sheet and the exam. Fill in test form A on the scantron sheet. Answer all multiple choice questions on your scantron sheet. Choose the single best answer for each multiple choice question. Answer the long answer questions directly on the exam. Keep your answers complete but concise. For the long answer questions, you must show your work where appropriate for full credit.

Name: **ID Number:** **Section:**

(POTENTIALLY) USEFUL FORMULAS

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$CV = \frac{s}{\bar{x}}$

$skew = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s})^3$

$kurt = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$

$\mu = E(X)$

$z^* = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

$t^* = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

$t^* = \frac{b_j-\beta_j}{s_{b_j}}$

$Pr[T_{n-k} > t_{\alpha,n-k}] = \alpha$

$Pr[|T_{n-k}| > t_{\frac{\alpha}{2},n-k}] = \alpha$

$\sum_{i=1}^{n} a = na$

$\sum_{i=1}^{n}(ax_i) = a\sum_{i=1}^{n} x_i$

$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

$s^2 = \bar{x}(1-\bar{x})$ for proportions data

$t_{\alpha,n-k} = TINV(2\alpha, n-k)$

$Pr(|T_{n-k}| \geq |t^*|) = TDIST(|t^*|, n-k, 2)$

$Pr(T_{n-k} > t^*) = TDIST(t^*, n-k, 1)$

$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$

$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}}$

$b_2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = r_{xy}\sqrt{\frac{s_{yy}}{s_{xx}}}$

$b_1 = \bar{y} - b_2\bar{x}$

$\hat{y}_i = b_1 + \sum_{j=2}^{k} b_j x_{j,i}$

$s_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

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

$R^2 = 1 - \frac{ESS}{TSS}$

$s_{b_2} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

$\bar{R}^2 = 1 - \frac{n-1}{n-k}\frac{ESS}{TSS}$

$F^* = \frac{n-k}{k-1}\frac{R^2}{1-R^2}$

$F^* = \frac{n-k}{k-g}\frac{ESS_r - ESS_u}{ESS_u} = \frac{n-k}{k-g}\frac{R_u^2 - R_r^2}{1-R_u^2}$

$Pr(F_{k-g,n-k} > F^*) = FDIST(F^*, k-g, n-k)$

SECTION I: MULTIPLE CHOICE (60 points)

1. Suppose that we regress Y on X. In which of the following scenarios would the estimated slope coefficient not be biased?
   (a) There is random measurement error in X.
   (b) There is random measurement error in Y.
   (c) There is an omitted variable Z that is correlated with both X and Y.
   (d) Both (b) and (c).

2. Which of the following would be the best graph for showing the distribution of heights in a sample of one thousand students?
   (a) Scatter plot.
   (b) Line chart.
   (c) Histogram.
   (d) Bubble chart.

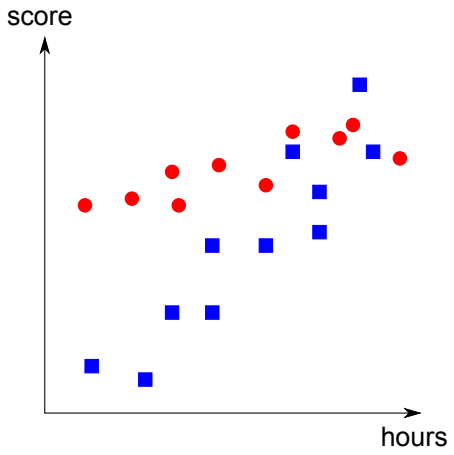
3. Suppose $\tilde{b_j}$ is an estimator for the slope coefficient β_j. The expected value of $\tilde{b_j}$ is equal to β_j plus ten. As the sample size gets larger and larger, the standard error of $\tilde{b_j}$ gets closer and closer to zero while the expected value of $\tilde{b_j}$ remains the same. Which of the following statements is true?
   (a) $\tilde{b_j}$ is an unbiased, consistent estimator of β_j.
   (b) $\tilde{b_j}$ is a consistent but biased estimator of β_j.
   (c) $\tilde{b_j}$ is an unbiased estimator fo β_j but is not consistent.
   (d) b_j is a biased estimator of β_j and is not consistent.

4. Suppose that on average, an extra inch in height is associated with an extra five pounds in weight. We have a dataset containing height and weight information in which height is rounded to the nearest inch and weight is rounded to the nearest pound. If we regress weight on height, the expected value of the estimated height coefficient would be:
   (a) Equal to 5.
   (b) Equal to $\frac{1}{5}$.
   (c) Less than 5.
   (d) Greater than 5.

5. Suppose that we run a regression of Y on X_2 and get an R^2 of 0.60. If we run a regression of Y on X_2 and X_3, we would expect:
   (a) The new R^2 to be greater than or equal to 0.60.
   (b) The new R^2 to be greater than or equal to 0.60 if and only if the correlation between X_2 and X_3 is positive.
   (c) The new R^2 to be less than or equal to 0.60.
   (d) The new R^2 to be less than or equal to 0.60 if and only if the correlation between X_2 and X_3 is negative.

6. Suppose we have a sample of commute times for 1,000 people. If we measure commute time in minutes, the variance will be _____ if we measured commute time in hours and the mean will be _____ if we measured commute time in hours.

   (a) Smaller than, the same as.
   (b) Larger than, smaller than.
   (c) Smaller than, smaller than.
   (d) None of the above.



Use the figure above to answers questions 7 through 10. The figure is a scatter plot with hours of study on the horizontal axis and final exam score on the vertical axis for 21 students in an ECN 100 class. The round data points correspond to economics majors. The square data points correspond to nonmajors. Suppose we use this data to estimate the following model:

$$SCORE = \beta_1 + \beta_2 MAJOR + \beta_3 HOURS + \beta_4 MAJOR \cdot HOURS + \varepsilon$$

where $SCORE$ is a student's final exam score, $MAJOR$ is a dummy variable equal to one if the student is an economics major and zero otherwise, $HOURS$ is the number of hours the student studies for the final and ε is a random error term that satisfies all of our assumptions.

7. Based on the scatterplot, we would expect our estimated value of β_2 to be:

   (a) Positive.
   (b) Negative.
   (c) Larger for economics majors than nonmajors.
   (d) Larger for nonmajors than economics majors.

8. Based on the scatterplot, we would expect our estimated value of β_4 to be:

   (a) Positive.
   (b) Negative.
   (c) Larger for economics majors than nonmajors.
   (d) Larger for nonmajors than economics majors.

9. The predicted score for an economics major who studies ten hours will be:
   (a) Greater than the predicted score for a nonmajor who studies ten hours.
   (b) Less than the predicted score for a nonmajor who studies ten hours.
   (c) Equal to the predicted score for a nonmajor who studies ten hours.
   (d) Not enough information.
10. The predicted increase in score for an economics major associated with one extra hour of studying will be:
   (a) Equal to b_3, where b_3 is our estimated value of β_3.
   (b) Equal to $b_3 + b_4$, where b_3 and b_4 are our estimated values of β_3 and β_4, respectively.
   (c) Equal to $b_3 + b_4 \cdot HOURS$, where b_3 and b_4 are our estimated values of β_3 and β_4, respectively.
   (d) Equal to b_4, where b_4 is our estimated value of β_4.
11. Which of the following would lead to a smaller standard error for the slope coefficient in a bivariate regression?
   (a) A smaller variance of the independent variable.
   (b) A smaller average size of the residuals.
   (c) A larger error sum of squares.
   (d) A smaller sample size.
12. Which of the following would make us less likely to reject the null hypothesis that the true population mean is 150?
   (a) Getting a sample mean that is farther from 150.
   (b) Getting a larger t statistic.
   (c) Switching to a larger value for the significance level α.
   (d) None of the above.
13. Which of the following statements is definitely not true?
   (a) Adding an additional regressor can change the estimated slope coefficients of the other regressors.
   (b) Adding an additional regressor can decrease the R^2 of the regression.
   (c) Adding an additional regressor can lower the error sum of squares of the regression.
   (d) Adding an additional regressor can increase the standard errors of the estimated slope coefficients of the other regressors.
14. Suppose we want to estimate the effect of coffee on hours of sleep. We believe that every extra cup of coffee a person drinks decreases her hours of sleep by a constant percentage. To estimate the relationship of interest, we would:
   (a) Run a regression with minutes of sleep as the independent variable and the log of cups of coffee as the dependent variable.
   (b) Run a regression with minutes of sleep as the dependent variable and the log of cups of coffee as the independent variable.
   (c) Run a regression with the log of minutes of sleep as the independent variable and cups of coffee as the dependent variable.
   (d) Run a regression with the log of minutes of sleep as the dependent variable and cups of coffee as the independent variable.

15. The 95% confidence interval for the population mean of X:
    (a) Will be wider than the 90% confidence interval for the population mean.
    (b) Will tend to get wider as we use more observations.
    (c) Will not depend on the variance of X.
    (d) None of the above.
16. Which of the following is not an assumption we make when doing bivariate statistical inference?
    (a) The expected value of the error term is equal to zero.
    (b) The errors are uncorrelated with the regressor.
    (c) The errors are uncorrelated with the dependent variable.
    (d) We assume all of the above.
17. Suppose we are testing whether high school GPA and college GPA are jointly significant in a regression with log wage as the dependent variable. Which of the following statements is true?
    (a) We will reject the null hypothesis that both coefficients are equal to zero at a 5% significance level if and only if both of the variables are individually significant at the 5% significance level.
    (b) We will reject the null hypothesis that both coefficients are equal to zero at a 5% significance level if and only if at least one of the two variables is individually significant at the 5% significance level.
    (c) We will reject the null hypothesis that both coefficients are equal to zero at a 5% significance level if one of the two variables is individually significant at the 5% significance level.
    (d) None of the above.
18. Suppose that we run a regression with the log of pounds of rice purchased from a store as the dependent variable and log of the price of a pound of rice as the independent variable. A slope coefficient of 0.3 would be interpreted as:
    (a) A one dollar increase in the price of a pound of rice is associated with a 0.3 pound increase in the amount of rice purchased.
    (b) A one percent increase in the price of a pound of rice is associated with a 30 percent increase in the amount of rice purchased.
    (c) A one percent increase in the price of a pound of rice is associated with a 0.3 percent increase in the amount of rice purchased.
    (d) A one dollar increase in the price of a pound of rice is associated with a 30 percent increase in the amount of rice purchased.
19. Suppose that the slope coefficient for a particular regressor X_j has a p-value of 0.02. We would conclude that the coefficient is:
    (a) Economically significant at a 5% significance level.
    (b) Economically significant at a 10% significance level.
    (c) Statistically significant at a 10% significance level.
    (d) All of the above.

20. Adding an irrelevant variable to a regression will tend to lower:
    (a) The R^2.
    (b) The adjusted R^2.
    (c) The standard errors of the coefficients for the other variables.
    (d) The total sum of squares.

21. If X and Y are perfectly correlated, when we regress Y on X we would get:
    (a) A slope coefficient equal to one.
    (b) A slope coefficient equal to either one or negative one.
    (c) An R^2 equal to one.
    (d) An intercept equal to zero.

22. The distribution of the sample mean of X will:
    (a) Be centered at zero.
    (b) Be centered at the population mean of X.
    (c) Have a variance equal the variance of X.
    (d) Both (b) and (c).

23. Suppose that the true relationship between Y and X is given by:

$$Y = \beta_1 + \beta_2 X + \varepsilon$$

    where ε is a random error term that meets all of our assumptions. Which of the following is not a random variable?
    (a) b_2, the estimated value of the slope coefficient.
    (b) β_2.
    (c) The sample mean of Y.
    (d) The error sum of squares from a regression of Y on X.

24. Suppose we have a categorical variable that can take on eight different values. We want to control for this variable in a wage regression. We would need to include:
    (a) One continuous variables in our regression to do this.
    (b) Seven different continuous variables in our regression to do this.
    (c) Seven different dummy variables in our regression to do this.
    (d) Eight different dummy variables in our regression to do this.

25. Suppose that we can reject the null hypothesis that the population mean is less than or equal to 10 at a 5% significance level. Which of the following statements is definitely true?
    (a) We can reject the null hypothesis that the population mean is equal to 10 at a 5% significance level.
    (b) We can reject the null hypothesis that the population mean is greater than or equal to 10 at a 5% significance level.
    (c) We can reject the null hypothesis that the population mean is equal to 10 at a 1% significance level.
    (d) None of the above.

SECTION II: SHORT ANSWER (40 points)

1. (18 points) For each scenario below, a researcher is attempting to estimate a slope coefficient of particular interest. Explain whether the researcher will get an unbiased estimate of the slope coefficient and, if not, what direction the bias will be in. Note that there may be multiple correct answers for each question. If you can properly justify your answer, you will receive full credit.

   (a) A researcher wants to estimate the effect of winter on happiness for the population of the United States. To do this, a researcher asks 1,000 people from Southern California how happy they are on a scale of 0 to 100 (with 100 being the happiest). The researcher creates a dummy variable that equals one if a person was asked this question in the winter months and equal to zero otherwise. To estimate the effect of winter on happiness, the researcher regresses the happiness number on this dummy variable for winter.

   (b) A high school wants to know if assigning more reading leads to higher test scores. To determine this, the school takes a random sample of one hundred students and regresses their test scores on the number of pages they were assigned to read. Teachers who assign more reading also tend to spend more time preparing their lectures and answering their students' questions.

   (c) A student wants to know the effect of hours of exercise per week on her weekly quiz scores. To do this, the student looked up all of her quiz scores for the past year and tried to remember how many hours she exercised each week for the past year. She knows the exact quiz scores but can only remember roughly (not exactly) how many hours she exercised each week. To determine the effect of exercise on quiz performance, she regresses quiz score on hours of exercise per week. She uses 52 data points for this regression (one for every week in the past year).

2. (12 points) A researcher is interested in how outside temperature influences electricity usage. The researcher thinks that at very low temperatures, people use a lot of electricity to run their electric space heaters. As outside temperature rises, people use their space heaters less frequently and electricity usage goes down. However, as temperatures get very hot, electricity usage begins to rise again as people use their air conditioners more and more. To estimate this relationship between outside temperature and electricity usage, the researcher decides to use the following model:

$$E = \beta_1 + \beta_2 T_c + \beta_3 T_f + \varepsilon \tag{1}$$

where E is total electricity usage over a one week period measured in kilowatt-hours, T_c is the average temperature over that week measured in degrees Celcius, and T_f is the average temperature over that week measured in degrees Fahrenheit. The data the researcher uses come from a cross-section of US households, each observation represents one household.

   (a) Explain two ways in which the researcher has misspecified the model. Explain what changes you would make to the model to deal with with these problems.
   (b) Suppose that high income households use more electricity at any given temperature than low income households. The average difference in electricity usage between high and low income households is the same at every temperature level. Given this information, what is the correct popluation model (assume that the dataset also contains a variable I measuring household income)?
   (c) Suppose you ran a regression to estimate all of the parameters in your model from part (b). Based on the information given throughout the problem, what would you predict the signs to be for each parameter?

SUMMARY OUTPUT: API score as dependent variable

Regression Statistics	
Multiple R	0.379
R Square	0.144
Adjusted R Square	0.143
Standard Error	76.927
Observations	6426

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	688.49	5.93	116.09	0	676.86	700.11
CLASSSIZE	3.91	0.21	18.77	1.41E-76	3.50	4.32
YEARROUND	-21.82	4.90	-4.46	8.5E-06	-31.43	-12.22
NONHSGRAD	-88.76	3.56	-24.94	4.1E-131	-95.73	-81.78

SUMMARY OUTPUT: API score as dependent variable

Regression Statistics	
Multiple R	0.234
R Square	0.055
Adjusted R Square	0.055
Standard Error	80.801
Observations	6426

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	671.78	6.19	108.53	0	659.65	683.91
CLASSSIZE	4.22	0.22	19.33	5.42E-81	3.80	4.65

SUMMARY OUTPUT: API score as dependent variable

Regression Statistics	
Multiple R	0.311
R Square	0.097
Adjusted R Square	0.096
Standard Error	79.003
Observations	6426

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	798.17	1.04	765.51	0	796.13	800.22
YEARROUND	-23.39	5.03	-4.65	3.38E-06	-33.25	-13.53
NONHSGRAD	-92.40	3.65	-25.32	7.2E-135	-99.55	-85.24

Summary Statistics for the Regression Sample

Variable	Mean	Minimum	Maximum	Standard Deviation
API	789.854	310	998	83.112
CLASSSIZE	27.949	1	50	4.613
YEARROUND	0.040	0	1	0.197
NONHSGRAD	0.080	0	1	0.271

3. (10 points) Use the regression results on the previous page to answer this problem. Below the regression results is a table of summary statistics for the regression sample. The regression results are from a cross-section of California school districts. Each observation represents one school district. The variables are defined as follows:
   - API score - academic performance index score. This is a measure of the overall academic performance of students in the school district. The score can range from 200 to 1000, with 1000 being the highest possible academic performance.
   - CLASSSIZE - class size. This is the average number of students per classroom in the school district.
   - YEARROUND - dummy variable for year round schools. This variable is equal to one if schools in a district are open year round and equal to zero if schools close for the summer.
   - NONHSGRAD - dummy variable indicating districts in which many parents are not high school graduates. This variable is equal to one if over half of the parents in the district did not graduate from high school. It is equal to zero if over half of the parents in the district did graduate from high school.

   (a) Suppose that school district A has an average class size of 30, it holds classes year round, and 80% of the parents in the district are high school graduates. School district B has an average class size of 20, it does not hold classes in the summer, and 40% of the parents in the district are high school graduates. How much higher (or lower) would district A's predicted API score be compared to district B? Show all of your calculations.

   (b) Suppose you wanted to use an F test to determine whether the $YEARROUND$ and $HSGRAD$ variables are jointly significant in the regression that includes $CLASSSIZE$, $YEARROUND$ and $HSGRAD$. In other words, you want to test the following set of hypotheses:

   $$H_0\text{: } \beta_{YEARROUND} = \beta_{HSGRAD} = 0$$

   $$H_a\text{: at least one of } \beta_{YEARROUND} \text{ and} \beta_{HSGRAD} \text{ is not equal to zero}$$

   Calculate the F statistic you would use to do this test. You should calculate an exact numerical value for the F statistic.