Ecn 102 - Analysis of Economic Data
University of California - Davis June 8, 2009
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 for full credit.

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

(POTENTIALLY) USEFUL FORMULAS

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

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

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

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

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

$\mu = E(X)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\hat{y}_i = b_1 + b_2 x$

$e_i = y_i - \hat{y}_i$

$b_2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

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

$b_2 = r_{xy}\sqrt{\frac{s_{yy}}{s_{xx}}}$

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

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

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

$E(b_j) = \beta_j$

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

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

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

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

$F^* = \frac{n-k}{k-1}\frac{R^2}{1-R^2}$ (for testing all regressors)

$F^* = \frac{n-k}{k-g}\frac{ESS_r - ESS_u}{ESS_u} = \frac{n-k}{k-g}\frac{R_u^2 - R_r^2}{1-R_u^2}$ (for testing subset of regressors)

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

SECTION I: MULTIPLE CHOICE (60 points)

1. If an additional explanatory variable is added to a regression:
   (a) The R^2 for the regression may decrease.
   (b) The adjusted R^2 for the regression will either stay the same or increase.
   (c) The R^2 for the regression will either stay the same or increase.
   (d) Both (b) and (c).
2. We would expect the coefficient of variation for height measured in inches:
   (a) To be larger than the coefficient of variation for height measured in meters.
   (b) To be smaller than the coefficient of variation for height measured in meters.
   (c) To be equal to the the coefficient of variation for height measured in meters.
   (d) (a), (b) and (c) could all be true.
3. Suppose we ran a regression of student's high school GPA on parents' income. Which of the following would not lead to a biased coefficient for parents' income?
   (a) Measurement error in parents' income.
   (b) The fact that student ability is correlated with parents' income and student's GPA.
   (c) Measurement error in student's GPA.
   (d) Selecting students only from honors classes.
4. A histogram is useful for showing:
   (a) The distribution of the observed values for a single variable.
   (b) The relationship between a single dependent variable and a single regressor.
   (c) The correlation between two variables.
   (d) The precision of an estimated slope coefficient.
5. Inclusion of an irrelevant variable in a regression:
   (a) Will tend to bias the coefficients for the other variables downward and increase their standard errors.
   (b) Will tend to bias the coefficients for the other variables upward and increase their standard errors.
   (c) Will not bias the coefficients for the other variables but will increase their standard errors.
   (d) Will not bias the coefficients for the other variables but will decrease their standard errors.
6. If we have four different types of cars in our dataset and would like to include dummy variables for car type in a regression, how many dummy variables will we use?
   (a) Four.
   (b) Three.
   (c) Two.
   (d) One.

Use the following information to answer questions (7) and (8). Suppose we run a regression of annual medical expenses (MED) measured in dollars on age (AGE) measured in years, hours of exercise per day ($HOURS$, note that this can be no larger than 24 and no smaller than 0), and age interacted with hours of exercise and come up with the following coefficient estimates:

$$MED = 10000 + 50 \cdot AGE - 500 \cdot HOURS - 1 \cdot AGE \cdot HOURS$$

7. Based on these coefficients, if predicted medical expenses were graphed as a function of hours of exercise:
   (a) The slope of the line would be steeper for a 50-year-old person compared to a 40-year-old person.
   (b) The slope of the line would be flatter for a 50-year-old person compared to a 40-year-old person.
   (c) The slope of the line would be the same for a 50-year-old person and a 40-year-old person but the intercepts would be different.
   (d) The slope of the line and the value of the intercept would be the same for a 50-year-old person and a 40-year-old person.
8. The predicted change in medical expenses associated with an increase in age of one year will be:
   (a) Positive regardless of a person's level of exercise.
   (b) Negative regardless of a person's level of exercise.
   (c) Could be postive or negative depending on a person's level of exercise.
   (d) Will be independent of a person's level of exercise.
9. The sample mean:
   (a) Will be equal to the population mean on average.
   (b) Will always be equal to the population mean.
   (c) Is a constant.
   (d) Has a distribution centered at zero.
10. The error sum of squares:
   (a) Will be larger the better our model fit is.
   (b) Will be larger than the total sum of squares when we have a misspecified model.
   (c) Will always be less than or equal to the total sum of squares.
   (d) (a) and (c).
11. Which of the following is always true?
   (a) The correlation between two variables has the same sign as the covariance.
   (b) The magnitude of the correlation between two variables will be larger than the magnitude of the covariance.
   (c) The magnitude of the correlation between two variables will be smaller than the magnitude of the covariance.
   (d) The correlation between x and y will have the same sign as the R^2 for a regression of y on x.

12. Suppose that the true relationship between happiness ($HAPPY$, measured on a scale of 0 to 100) and wealth ($WEALTH$, measured in dollars) is given by:

$$HAPPY = \beta_1 + \beta_2 WEALTH + \varepsilon$$

where ε is an error term that satisfies all of our assumptions. If our measure of wealth contains some random measurement error, the magnitude of the coefficient we estimate for wealth:

(a) Will be smaller than $|\beta_2|$ on average.
(b) Will be larger than $|\beta_2|$ on average.
(c) Will be equal to $|\beta_2|$ on average.
(d) The answer depends on the sign of β_2.

13. We are more likely to reject the null hypothesis that a particular coefficient β_j from a multivariate regression is equal to zero when:

(a) The standard error for the coefficient is very small.
(b) The R^2 for the regression is close to zero.
(c) The estimated coefficient is very close to zero.
(d) The estimated coefficient divided by its standard error is very small.

14. When regressing y on x, an estimated coefficient for x of 5 that is statistically significant tells us:

(a) That a one unit increase in x is associated with a five unit increase in y.
(b) That a one unit increase in x causes a five unit increase in y.
(c) That either an increase in x causes an increase in y or an increase in y causes an increase in x.
(d) (a) and (c).

15. Suppose that we regress y on x_2 but don't include x_3 leading to an omitted variable bias for the coefficient for x_2. The sign of the bias will depend on:

(a) The sign of the correlation between x_2 and x_3.
(b) The sign of the correlation between x_3 and y.
(c) Both (a) and (b).
(d) Neither (a) nor (b).

16. Suppose we have data on food expenditures and income in the year 2007 for 200 different households. This is an example of:

(a) Univariate data.
(b) Cross-section data.
(c) Time series data.
(d) None of the above.

17. Suppose that a person's health increases with an increase in the amount of calories consumed at low levels of calories but decreases with an increase in the amount of calories consumed at high levels of calories. Which equation would best model this relationship?
    (a) $HEALTH = \beta_1 + \beta_2 CALORIES$
    (b) $\ln(HEALTH) = \beta_1 + \beta_2 CALORIES$
    (c) $\ln(HEALTH) = \beta_1 + \beta_2 \ln(CALORIES)$
    (d) $HEALTH = \beta_1 + \beta_2 CALORIES + \beta_3 CALORIES^2$
18. Increasing the sample size used for a multivariate regression will tend to:
    (a) Make the standard errors for the individual coefficients larger.
    (b) Make the confidence intervals for the individual coefficients wider.
    (c) Increase the size of the standard errors relative to the magnitude of the coefficients.
    (d) Decrease the standard errors of the individuals coefficients.
19. Which of the following are random variables:
    (a) The estimated coefficients in a multivariate regression.
    (b) The sample mean.
    (c) The standard error of a coefficient in a bivariate regression.
    (d) All of the above.
20. The _____ the difference between the R^2 for the unrestricted model and the R^2 for the restricted model, the _____ we are to reject the null hypothesis that the coefficients for a subset of regressors are all equal to zero.
    (a) Larger, less likely.
    (b) Smaller, more likely.
    (c) Larger, more likely.
    (d) The R^2 values are irrelevant.
21. If $\ln y$ is regressed on $\ln x$, the coefficient on $\ln x$ tells us:
    (a) The predicted change in y with a one unit change in x.
    (b) The predicted percentage change in y with a one percent change in x.
    (c) The predicted change in y with a one percent change in x.
    (d) The predected percentage change in y with a one unit change in x.
22. Suppose that we regress height on a dummy variable equal to one if a person is female and equal to zero if a person is male. The coefficient for the dummy variable gives us an estimate of:
    (a) The average height of females.
    (b) The average height of males.
    (c) The difference between the average height of females and the average height of males.
    (d) The difference in the variance of female heights and the variance of male heights.

23. Which of the following will not decrease the standard error of a coefficient in a bivariate regression of y on x?
    (a) Greater variation in the observed x values.
    (b) A larger number of observations.
    (c) A smaller error sum of squares.
    (d) Larger residuals.
24. Which of the following would be the best type of graph to display changes in the unemployment rate over time?
    (a) A line graph.
    (b) A bar chart.
    (c) A histogram.
    (d) A pie chart.
25. Which of the following scenarios violates our assumptions for multivariate hypothesis testing?
    (a) Two of the regressors have a correlation of .80.
    (b) Two of the regressors have a correlation of 0.
    (c) The correlation between one of the regressors and the error term is .55.
    (d) (a) and (c).

SECTION II: SHORT ANSWER (40 points)

1. The following page contains the results from three different regressions using the same sample of standardized test score data for California high schools. Each observation corresponds to a single high school. The variables are the following (all percentages are measured from 0 to 100, not from 0 to 1):
   - $SCORE$ - average test score for the high school (higher scores mean students performed better)
   - PCT_SD - percentage of students who are socioeconomically disadvantaged (meaning the students come from poor households)
   - $PCT\_NONHSGRAD$ - percentage of students who have parents that did not graduate high school
   - $PCT\_LEARNERS$ - percentage of students who are learning English as a second language
   - PCT_CRED - percentage of teachers who are fully credentialed

   (a) Explain why the coefficient on PCT_SD differs between Regression A and Regression B. Be certain that your explanation specifically addresses why the coefficient is *greater* (more positive) in Regression B than it is in Regression A. (4 points)

   (b) In Regression B, is the coefficient for $PCT\_NONHSGRAD$ significantly different than zero at a 5% significance level? Note, even if you do not need to do any calculations, you must still explain how you arrived at your answer. (3 points)

   (c) For Regression B, would you reject the null hypothesis that the coefficients for PCT_SD and $PCT\_NONHSGRAD$ are both zero at a 5% significance level? Be certain to justify your answer even if you do not need to do any calculations. (3 points)

   (d) Test whether the addition of $PCT\_LEARNER$ and PCT_CRED improved the fit of the regression at a 5% significance level. Clearly state your null and alternative hypotheses, show any calculations you may need to make, and be certain to clearly state your conclusions. (5 points)

Regression A Output: Dependent variable is SCORE

Regression Statistics	
R Square	0.295305163
Adjusted R Square	0.294941731
Standard Error	102.1472355
Observations	1941

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8478148.006	8478148	812.5456	1.4437E-149
Residual	1939	20231637.93	10434.06		
Total	1940	28709785.94			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	775.9137107	4.831396195	160.5982	0
PCT_SD	-2.340641476	0.082112829	-28.50519	1.4E-149

Regression B Output: Dependent variable is SCORE

Regression Statistics	
R Square	0.301084321
Adjusted R Square	0.300363045
Standard Error	101.7537638
Observations	1941

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	8644066.395	4322033	417.4333	1.7685E-151
Residual	1938	20065719.54	10353.83		
Total	1940	28709785.94			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	774.4214522	4.827200746	160.4287	0
PCT_SD	-1.947588476	0.127794267	-15.24003	1.28E-49
PCT_NONHSGRAD	-0.816419724	0.203946713	-4.003103	6.49E-05

Regression C Output: Dependent variable is SCORE

Regression Statistics	
R Square	0.301377117
Adjusted R Square	0.299933681
Standard Error	101.7849819
Observations	1941

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	8652472.518	2163118	208.7915	4.9375E-149
Residual	1936	20057313.42	10360.18		
Total	1940	28709785.94			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	769.9690284	19.18397495	40.13605	7.9E-257
PCT_SD	-1.926887117	0.130714519	-14.74119	1.08E-46
PCT_NONHSGRAD	-0.717102472	0.233177276	-3.075353	0.002132
PCT_LEARNERS	-0.190856233	0.220795236	-0.864404	0.387473
PCT_CRED	0.045892944	0.189469762	0.242218	0.808637

2. Below is the regression output from running a regression of educational attainment (measured as years of secondary and post-secondary education) on a dummy variable for whether the individual was the oldest child in his or her family (*oldest*, equal to one if individual was the oldest child, zero otherwise) and a dummy variable for whether the individual was the youngest child in his or her family (*youngest*, equal to one if the individual was the youngest child, zero otherwise). The sample includes only individuals with at least two other siblings.

SUMMARY OUTPUT: Dependent variable is years of education

Regression Statistics	
R Square	0.018739908
Adjusted R Square	0.008131691
Standard Error	2.005199409
Observations	188

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.205128205	0.16054444	13.73531	4.71E-30
youngest	0.044871795	0.476251772	0.094219	0.925037
oldest	1.128205128	0.600702289	1.878144	0.061935

(a) Could you run the same regression for a sample of individuals with a single sibling? If not, how would you need to modify the regression. (3 points)

(b) Explain in words how we should interpret the estimated coefficient for *youngest*. (3 points)

(c) Test the hypothesis that the oldest child receives more education than a middle child at a 5% significance level. Be certain to clearly state your null and alternative hypotheses, show any necessary calculations, and state your conclusions clearly. (4 points)

(d) Suppose that a researcher claims, "Birth order accounts for the majority of the observed variation in educational attainment." Based on these regression results, would you agree or disagree with the statement? Be certain to cite any relevant regression output that supports your answer. (3 points)

3. For each scenario below, explain whether the estimated coefficient for education would be biased and if so, what the direction of the bias would be. Be certain to justify your answers.
   (a) You run a regression of income on education. You have mistakenly omitted ability from the regression which is positively correlated with both income and education. (4 points)
   (b) You run a regression of a measure of job stress on education. The coefficient on education turns out to be negative and significant but because of survey difficulties, the education variable has problems of random measurement error. (4 points)
   (c) You run a regression of annual income on education but it turns out that your sample consists of many older people who have either retired or cut back to working part time. (4 points)