Ecn 102 - Analysis of Economic Data
University of California - Davis March 19, 2010
Instructor: John Parman

Final Exam

You have until 5:30pm 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

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

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

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

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

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

$\mu = E(X)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$F = \frac{ESS_r - ESS_u}{ESS_u}\frac{n-k}{k-g}$

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

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

SECTION I: MULTIPLE CHOICE (60 points)

1. Suppose we regress SAT score on parent's education and parent's income. If we run the regression again but also include the student's GPA as an additional regressor:
   (a) The R^2 for the regression will either stay the same or increase.
   (b) The adjusted R^2 for the regression will either stay the same or increase.
   (c) Both (a) and (b) are true.
   (d) Neither (a) nor (b) is true.

2. Suppose we have a sample of the heights of Davis students and want to use the sample mean to get a confidence interval for the mean height in the population. Which of the following would increase the width of this confidence interval?
   (a) Switching from a 95% confidence interval to a 90% confidence interval.
   (b) Increasing the sample size used to calculate the sample mean.
   (c) Switching from a 95% confidence interval to a 99% confidence interval.
   (d) All of the above.

3. Suppose we can reject the null hypothesis that $\beta_2 \geq 0$ at a 5% significance level where β_2 is the slope coefficient from a bivariate regression. Which of the following is definitely true?
   (a) Our test statistic was negative.
   (b) We can reject the null hypothesis that $\beta_2 = 0$ at a 5% significance level.
   (c) We can reject the null hypothesis that $\beta_2 \geq 0$ at a 2.5% significance level.
   (d) We can reject the null hypothesis that $\beta_2 < 0$ at a 5% significance level.

4. Suppose we regress y on x_2. Which of the following would lead to a biased coefficient for x_2?
   (a) There is a variable x_3 that is correlated with y but not with x_2
   (b) There is a variable x_3 that is correlated with x_2 but not with y.
   (c) y is measured with some random error.
   (d) x is measured with some random error.

5. When testing the significance of a subset of regressors, the R^2 of the unrestricted model will always be:
   (a) Greater than or equal to the R^2 of the restricted model.
   (b) Less than or equal to the R^2 of the restricted model.
   (c) Equal to the R^2 of the restricted model.
   (d) It could be greater than, less than or equal to the R^2 of the restricted model.

6. Suppose that the true population model of the relationship between weight (W) and hours of exercise per day (H) is:
$$W = 200 - 10H + \varepsilon$$
where ε meets all of our assumptions. If hours of exercise is measured with some random, mean zero error, which of the following statements about the estimated slope coefficient $\tilde{b_2}$ is true?
   (a) $E(\tilde{b_2}) = -10$.
   (b) $E(\tilde{b_2}) = 10$.
   (c) $E(\tilde{b_2}) > -10$.
   (d) $E(\tilde{b_2}) < -10$.

7. Doubling the value of the largest observation in a sample of incomes will:
   (a) Increase the median.
   (b) Increase the mode.
   (c) Increase the mean.
   (d) All of the above.

8. Suppose that the risk of catching the flu is high for young children and the elderly but low for teenagers and younger adults. Which of the following equations would be the best choice for modeling the relationship between flu risk (R) and age (A)?
   (a) $R = \beta_1 + \beta_2 A + \varepsilon$.
   (b) $ln(R) = \beta_1 + \beta_2 ln(A) + \varepsilon$.
   (c) $R = \beta_1 + \beta_2 A + \beta_2 A^2 + \varepsilon$.
   (d) $R = \beta_1 + \beta_2 ln(A) + \varepsilon$.

9. Suppose that snow depth measured in feet is included as a regressor in a multivariate regression and the magnitude of the estimated coefficient for snowdepth is 5. If we rerun the regression using snow depth measured in inches, the new estimated coefficient on snowfall will be:
   (a) Larger than 5.
   (b) Smaller than 5.
   (c) Still equal to 5.
   (d) Not enough information.

10. When regressing anuual work hours on income, a researcher finds that the variance of the residuals increases as work hours increases. This will affect:
   (a) The expected value of the slope cofficient for income.
   (b) The magnitude of the standard error for the slope coefficient for income.
   (c) Both (a) and (b).
   (d) Neither (a) nor (b).

11. The histogram for hours of study per week based on a sample of 400 Davis students is symmetric and centered at 15 hours. Which of the following statements is true?
    (a) The sample median is 15 hours.
    (b) The sample mean is 15 hours.
    (c) The skewness for the sample is zero.
    (d) All of the above.

12. When running a bivariate regression, which of the following is not possible?
    (a) The error sum of squares is larger than the total sum of squares.
    (b) The error sum of squares is equal to the total sum of squares.
    (c) The error sum of squares is zero.
    (d) The error sum of squares is positive.

13. Which of the following would definitely not lead to the error term being correlated with a regressor x?
    (a) Random measurement error in x.
    (b) An omitted variable correlated with x.
    (c) Choosing an incorrect functional form for the regression equation.
    (d) Random measurement error in y.

14. Adding an irrelevant variable to a regression will:
    (a) Have no effect on the regression results.
    (b) Tend to bias the coefficients for the other regressors.
    (c) Lower the R^2.
    (d) None of the above.

15. Suppose we run a regression with GPA as the dependent variable and SAT score as the independent variable. Which of the following statements is definitely true?
    (a) The sign of the estimated slope coefficient will be the same as the sign of the correlation between GPA and SAT score.
    (b) The sign of the estimated slope coefficient could be different than the sign of the correlation between GPA and SAT score if there are omitted variables.
    (c) The magnitude of the slope coefficient will be equal to the magnitude of the correlation between GPA and SAT score.
    (d) The slope coefficient will be statistically significant.

16. Which of the following is not a measure of central tendency?
    (a) The mean.
    (b) The mode.
    (c) The sample range.
    (d) The median.

17. Suppose we use an F test after running a multivariate regression to test the null hypothesis that $\beta_3 = \beta_4 = 0$ and get an F statistic that is larger than the critical value for a 5% significance level. We would conclude that:
    (a) $\beta_3 \neq \beta_4$.
    (b) $\beta_3 > 0$ or $\beta_4 > 0$.
    (c) $\beta_3 \neq 0$ and $\beta_4 \neq 0$.
    (d) None of the above.

18. Which of the following would make you more likely to reject the hypothesis that an individual slope coefficient is equal to zero?
    (a) A larger standard error for that slope coefficient.
    (b) A smaller t statistic for that slope coefficient.
    (c) A smaller F statistic for the regression.
    (d) A larger value for the ratio of the coefficient to its standard error.

19. Suppose that we include a dummy variable for male and a dummy variable for female in a regression. This will create a:
    (a) Omitted variable bias problem.
    (b) Heteroskedasticity problem.
    (c) Multicollinearity problem.
    (d) Homoskedasticity problem.

20. The distribution of the sample mean will:
    (a) Be centered at zero.
    (b) Have a smaller variance for smaller sample sizes.
    (c) Centered at the population mean.
    (d) (b) and (c).

21. Suppose the R^2 for a bivariate regression is equal to 1. This tells us that:
    (a) The correlation between the dependent and independent variables is equal to 1.
    (b) The slope coefficient is equal to 1 or -1.
    (c) The error sum of squares is equal to the total sum of squares.
    (d) The dependent and independent variables are perfectly correlated.

22. Suppose we ran a regression of Y on X 1000 times using 1000 different samples and made a histogram of the resulting slope coefficient values. Which of the following is true about the distribution shown on the histogram?
    (a) It would be centered at zero.
    (b) It would look like a normal distribution.
    (c) All of the observations would be located at the true value of the slope coefficient.
    (d) It would be right skewed.

23. Which of the following depends on the units variables are measured in?

    (a) Correlation.
    (b) Coefficient of variation.
    (c) Estimated slope coefficient.
    (d) t statistic.

24. Suppose the size of your social network grows exponentially over time. Which of the following equations would be the most appropriate for modeling social network size (S) as a function of time (T):

    (a) $S = \beta_1 + \beta_2 T + \varepsilon$.
    (b) $ln(S) = \beta_1 + \beta_2 ln(T) + \varepsilon$.
    (c) $S = \beta_1 + \beta_2 ln(T) + \varepsilon$.
    (d) $ln(S) = \beta_1 + \beta_2 T + \varepsilon$.

25. Suppose that we want to test whether eye color influences the likelihood of being hired for a job. Our dataset includes five different values for the eye color variable. If we want to regress the probability of being hired on eye color, we will:

    (a) Convert each eye color to a number and include this new variable in for eye color number in the regression.
    (b) Create dummy variables for each eye color and include all of the dummy variables as regressors.
    (c) Create dummy variables for each eye color and include four of the dummy variables as regressors.
    (d) Create dummy variables for each eye color and include three of the dummy variables as regressors.

SECTION II: SHORT ANSWER (40 points)

1. (14 points) Suppose that the number of traffic accidents (N) is a function of the number of cars on the road (C) and the average speed of cars on the road (S). The true population relationship between accidents, cars and average speed is given by:

$$N = 1000 + 50C + 25S + \varepsilon \tag{1}$$

where ε is a random error that meets all of our standard assumptions. The number of cars on the road is negatively correlated with average speed due to the increased congestion associated with additional cars. The true population relationship between the number of cars and average speed is given by:

$$S = 80 - 4C + \nu \tag{2}$$

where ν is a random error that meets all of our standard assumptions.

(a) If you ran a regression with N as the dependent variable and C and S as the independent variables, what would the expected value of the estimated slope coefficient for C be? Assume that you include a constant term in your regression.

(b) If you ran a regression with N as the dependent variable and C as the only independent variable, what would the expected value of the estimated slope coefficient for C be? Assume that you include a constant term in your regression.

(c) Suppose that you ran a regression with average speed as the dependent variable and number of cars on the road as the independent variable but you forced the intercept to be zero (in other words, you do not include a constant term). Will the the expected value of the estimated slope coefficient be greater than, equal to or less than the true population value of the slope coefficient? Include a written explanation and a scatter plot showing speed as a function of number of cars to illustrate your answer. Assume that we always observe positive numbers of cars and positive average speeds.

SUMMARY OUTPUT: height as dependent variable

Regression Statistics	
Multiple R	0.747320875
R Square	0.558488491
Adjusted R Square	0.533259262
Standard Error	0.357993439
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	11.34802735	2.837007	22.13657	7.68637E-12
Residual	70	8.971151182	0.128159		
Total	74	20.31917853			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	68.41515323	0.082797499	826.2949	2.3E-141
northeast	-0.731809115	0.102140602	-7.16472	6.25E-10
south	-0.260178112	0.137428521	-1.89319	0.062467
west	0.256275028	0.141697586	1.808605	0.074807
typhoiddeaths	0.02171783	0.009977496	2.176682	0.03288

Omitted region dummy variable is midwest.

SUMMARY OUTPUT: height as dependent variable

Regression Statistics	
Multiple R	0.260580275
R Square	0.06790208
Adjusted R Square	0.055133615
Standard Error	0.509357157
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.379714485	1.379714	5.317952	0.023948764
Residual	73	18.93946405	0.259445		
Total	74	20.31917853			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	68.09642269	0.07802453	872.7566	2E-148
typhoiddeaths	0.026346173	0.011424714	2.306068	0.023949

2. (14 points) For this problem, use the regression output shown on the previous page. Both regressions use the same data set. The dataset is a sample of 75 cities. *height* is a variable giving the average height in inches of adult males from the city. *tyhpoiddeaths* is a variable giving the number of typhoid deaths per 1,000 people in the city. The variables *northeast*, *south* and *west* are all dummy variables that are equal to one if the city is in that region and zero otherwise. All cities are located either in the Northeast, the South, the West or the Midwest.

   (a) Based on the regression results, what is the difference in the average male height between a city in the South and a city in the Northeast.

   (b) What is the average male height for a city in the West with no typhoid deaths?

   (c) Based on the first set of regression results, can you reject the null hypothesis that the coefficient for typhoid deaths is less than or equal to zero at a 5% significance level? Be certain to justify your answer.

   (d) Calculate the test statistic you would use to test the following set of hypotheses:

$$H_o\text{: } \beta_{ne} = \beta_s = \beta_w = 0$$

$$H_a\text{: at least one of } \beta_{ne},\ \beta_s \text{ and } \beta_w \text{ is different than zero}$$

   (e) Explain how you would use your test statistic from part (d) to decide whether or not to reject the null hypothesis. Be as specific as possible.

3. (12 points) Suppose that we are interested in the relationship between hours of weekly exercise and resting heart rate. The more a person exercises on average, the lower his or her resting heart rate is. For individuals who don't exercise at all, males have a lower resting heart rate on average than females. The decrease in resting heart rate from an additional hour of exercise per week is bigger for males than females.

   (a) Write down the regression model you would use to estimate the relationship between resting heart rate, gender and weekly exercise. Resting heart rate should be your dependent variable. Provide clear definitions of all variables you include in your model.

   (b) Based on the information given above, what are the expected signs for each of your coefficients in the regression model you specified in part (a)?

   (c) Suppose people tend to make random mistakes when measuring their heart rate. What effects will this have on the estimation results when you run the regression model specified in part (a)?