Ecn 102 - Analysis of Economic Data
University of California - Davis
Instructor: John Parman

February 24, 2010

Midterm 2

You have until 10:20am 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SECTION I: MULTIPLE CHOICE (60 points)

1. Suppose that in our sample the correlation between X and Y is negative. If we run a regression with X as the independent variable and Y as the dependent variable, we can say for certain that:
   (a) The intercept will be negative.
   (b) The R^2 will be negative.
   (c) The slope coefficient will be negative.
   (d) None of the above.
2. Which of the following would decrease the standard error of the estimated slope coefficient?
   (a) Decreasing the sample size.
   (b) Having greater variation in the independent variable.
   (c) Having a larger error sum of squares.
   (d) All of the above.

SUMMARY OUTPUT: Dependent variable is minutes per day spent caring for household members

Regression Statistics	
Multiple R	0.414
R Square	0.171
Adjusted R Square	0.171
Standard Error	74.767
Observations	7216

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	9.86	1.12	8.82	1.37E-18	7.67	12.05
Number of children	31.58	0.82	38.63	6.91E-297	29.98	33.18

Use the regression results above to answer questions 3 through 6. The independent variable is the number of children a person has. The dependent variable is the number of minutes a person spends each day caring for household members.

3. What is the predicted amount of time a person with three children would spend caring for household members each day?
   (a) 9.86 minutes.
   (b) 41.44 minutes.
   (c) 94.74 minutes.
   (d) 104.60 minutes.

4. Suppose that the amount of additional time required to care for family members increases with family size up to five children but then decreases with each child after the fifth (at this point, some of the children can help out around the house). Which of the following statements is correct?
   (a) The regression model used above is the best choice.
   (b) A log-linear model should be used instead.
   (c) A linear-log model should be used instead.
   (d) A polynomial in number of children should be used instead.
5. Which of the following null hypotheses would we use to test the claim that individuals with no children spend no time caring for household member?
   (a) H_0: $\beta_1 = 0$.
   (b) H_0: $\beta_2 = 0$.
   (c) H_0: $\beta_2 \geq 0$.
   (d) H_0: $\beta_2 \leq 0$.
6. Given the regression results, which of the following statements is definitely true?
   (a) The 90% confidence interval for the slope coefficient will be wider than the 95% confidence interval for the slope coefficient.
   (b) The 90% confidence interval for the intercept will be wider than the 90% confidence interval for the slope coefficient.
   (c) Both (a) and (b) will be true.
   (d) Neither (a) nor (b) will necessarily be true.
7. Which of the following scenarios would make us more likely to make a Type I error when testing the null hypothesis that β_2 is greater than or equal to zero?
   (a) Switching from a five percent significance level to a ten percent significance level.
   (b) Increasing the sample size.
   (c) Both (a) and (b) would increase the probability of a Type I error.
   (d) Neither (a) nor (b) would increase the probability of a Type I error.
8. If the correlation between X and Y is -1, then the data points on a scatter plot of X and Y would:
   (a) Fall along a straight line with a negative slope.
   (b) Fall along a straight line with a positive slope.
   (c) Tend to follow a downward sloping line but some points would not be exactly on the line.
   (d) Tend to follow an upward sloping line but some points would not be exactly on the line.

9. For every data point in our sample, the values of X and Y are positive. Suppose we run two regressions in which Y is the dependent variable and X is the independent variable. In the first regression, we estimate both an intercept, b_1, and a slope coefficient, b_2. In the second regression, we force the intercept to be zero and estimate only a slope coefficient, $\widetilde{b_2}$. Which of the following statements is definitely true?

   (a) $b_2 > \widetilde{b_2}$.
   (b) $b_2 < \widetilde{b_2}$.
   (c) $b_2 > \widetilde{b_2}$ if the correlation between X and Y in the sample is negative.
   (d) $b_2 < \widetilde{b_2}$ if the correlation between X and Y in the sample is negative.

10. Suppose that the true relationship between X and Y in the population is given by:

$$Y = \beta_1 + \beta_2 X^2 + \varepsilon$$

    where ε satisfies all of our assumptions. If we were to regress Y on X to get an estimated slope coefficient b_2, which of the following would definitely be true?

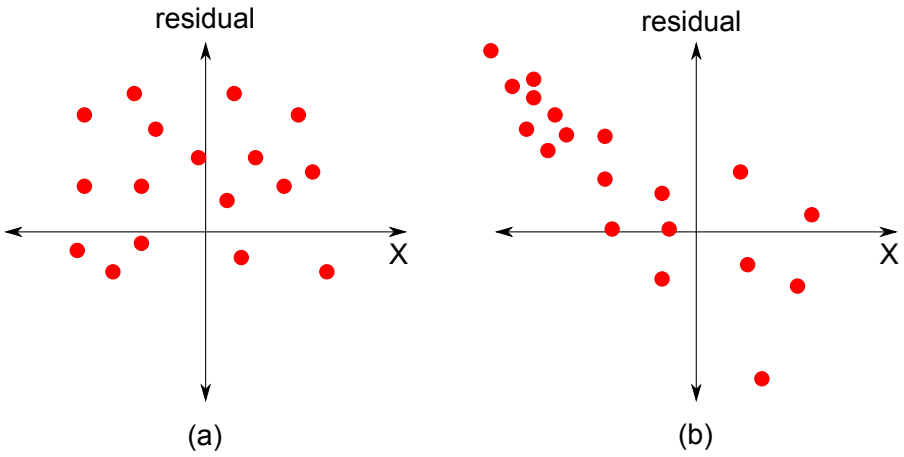
   (a) $b_2 = \beta_2$.
   (b) $E(b_2) = \beta_2$.
   (c) Both (a) and (b) would be true.
   (d) None of the above.

11. Suppose that the correlation between X and Y in our sample is negative. Which of the following is not true?

   (a) The covariance between X and Y in the sample will be less than zero.
   (b) If we regress Y on X, the error sum of squares will be greater than or equal to zero.
   (c) If we regress Y on X, the R^2 will be positive.
   (d) If we regress Y on X, the error sum of squares will be greater than or equal to the total sum of squares.

12. When the natural log of calories consumed per day is regressed on the natural log of weight, the slope coefficient on the natural log of weight is 0.10. Which of the following is a correct interpretation of this slope coefficient?

   (a) A one percent increase in weight causes a ten percent increase in calories consumed per day.
   (b) A ten percent increase in weight causes a one percent increase in calories consumed per day.
   (c) A one percent increase in weight is associated with a ten percent increase in calories consumed per day.
   (d) A ten percent increase in weight is associated with a one percent increase in calories consumed per day.



Use the graphs above and the following statements about the residuals to answer questions 13 and 14.

i. The residuals are uncorrelated with the regressor.
ii. The mean of the residuals is zero.
iii. The residuals have a constant variance that is independent of the regressor.

13. Given the plot of the residuals from regression (a), which of the above statements appear to be false?

(a) (i) only.
(b) (ii) only.
(c) (i) and (ii).
(d) (i), (ii) and (iii).

14. Given the plot of the residuals from regression (b), which of the above statements appear to be false?

(a) (i) only.
(b) (iii) only.
(c) (ii) and (iii).
(d) (i) and (iii).

15. The 95% confidence interval for the slope coefficient will be:

(a) Centered at b_2, the estimated slope coefficient.
(b) Centered at β_2, the population slope coefficient.
(c) Wider than the 99% confidence interval for the slope coefficient.
(d) Narrower than the 90% confidence interval for the slope coefficient.

16. Which of the following would change the estimated value of the intercept?

(a) Changing the units we use to measure the independent variable.
(b) Changing the units we use to measure the dependent variable.
(c) Both (a) and (b) would change the estimated value of the intercept.
(d) Neither (a) nor (b) would change the estimated value of the intercept.

17. We want to estimate the relationship between annual snowfall in the mountains and the average water level in Folsom Lake. We can choose either snowfall as recorded at Lake Tahoe or snowfall recorded at Donner Lake as our independent variable. In either case, our dependent variable would be the average water level in Folsom Lake. Which of the following statements is definitely true?
    (a) Either snowfall measure would give us the same total sum of squares for the regression.
    (b) Either snowfall measure would give us the same estimated slope coefficient.
    (c) Both (a) and (b) are definitely true.
    (d) Neither (a) nor (b) is necessarily true.

18. Suppose that you want to test the null hypothesis that the intercept is less than or equal to zero using a 10% significance level. Your sample has 200 observations and you are using a single independent variable in your regression. Which of the following critical values would you use?
    (a) $t_{0.05,198}$.
    (b) $t_{0.10,198}$.
    (c) $t_{0.05,199}$.
    (d) $t_{0.10,199}$.

19. Which of the following has a larger standard error?
    (a) The predicted actual value of y_i given x_i.
    (b) The predicted expected value of y_i given x_i.
    (c) (a) and (b) would have exactly the same standard errors.
    (d) Not enough information.

20. The estimated regression line:
    (a) Passes through the point $(\bar{y}, \bar{x})$ where $\bar{y}$ is the sample mean of y and $\bar{x}$ is the sample mean of x.
    (b) Passes through the point (μ_y, μ_x) where μ_y is the population mean of y and μ_x is the population mean of x.
    (c) Both (a) and (b) are true.
    (d) Neither (a) nor (b) are true.

SECTION II: SHORT ANSWER (40 points)

1. (12 points) The results of a regression of fuel economy measured in miles per gallon (MPG) regressed on air conditioner use (AC) are reported below. AC is a dummy variable equal to zero if the air conditioner is turned off and equal to one if the air conditioner is turned on. The numbers reported in parentheses are p-values. The number of observations and the R^2 of the regression are also reported.

MPG	=	23.4	-	0.02	x AC	N=150
		(0.001)		(0.005)		R^2=0.04

   (a) Is the sign of the slope coefficient what you would expect? Explain your answer in no more than two sentences.
   (b) Suppose that a researcher makes the following statement: "Based on the regression results, there is a statistically significant relationship between air conditioner use and fuel economy." Do you agree or disagree with the statement? Be certain to fully justify your answer.
   (c) Suppose that a researcher makes the following statement: "Based on the regression results, there is an economically significant relationship between air conditioner use and fuel economy." Do you agree or disagree with the statement? Be certain to fully justify your answer.

2. (12 points) For each scenario below, write down the equation for the regression you would use to estimate the described relationship in Excel (using the regression command from the data analysis toolpack). If you use any data transformations, make that clear in the way you write out the equation.

   (a) An ice cream store wants to know the effect of the outside temperature (T) on the number of ice cream cones sold (C). A one degree increase in temperature always leads to a constant percentage increase in cones sold.

   (b) You want to estimate the relationship between the number of cars on the highway (N) and the average speed of cars on the highway (S). You think that adding more cars to the highway decreases the average speed and that the decrease in speed from adding an additional car gets bigger as the total number of cars gets bigger.

   (c) A computer scientist wants to estimate the number of computers (C) infected by a virus as a function of time (T). The number of infected computers grows exponentially with time.

3. (16 points) You are given a dataset containing information on the number of house sales and the average sale price of houses for every month from October of 1999 to September of 2010. The format of the dataset is shown below. You can assume that there are no missing months of data.

Year	Month	Number of houses sold	Average sale price
1999	October	100	$200,000
1999	November	80	$150,000
1999	December	95	$110,000
2000	January	70	$240,000
2000	February	80	$235,000
2000	March	85	$180,000
…	…	…	…
…	…	…	…
…	…	…	…
2010	September	70	$150,000

Suppose that your realtor tells you to sell your house in the spring (the months of March, April and May) because average sale prices are 10% higher in the spring than in all other seasons. You would rather sell your house in the summer when you are not busy with classes. You decide that if average sale prices are over 10% larger in the spring than in the other seasons, you will sell in the spring. Otherwise, you will sell in the summer. Explain how you would use the dataset to determine whether or not to sell in the spring. Be certain that you cover all of the following details in your explanation:

- Any new variables you would need to create, including any transformations of existing variables you would need to do
- The exact regression equation you would use
- The null and alternative hypotheses you would test (make certain that you write these so that they are consistent with your regression equation)
- How you would calculate your test statistic
- How you would reach a decision based on your test statistic (assume that you want to use a 5% significance level)