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

February 23, 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

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

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

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

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

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

$\mu = E(X)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SECTION I: MULTIPLE CHOICE (60 points)

1. Suppose that all of our bivariate population assumptions are true. Increasing the sample size will:
   (a) Change the expected value of the slope coefficient.
   (b) Decrease the standard error of the slope coefficient.
   (c) Both (a) and (b).
   (d) Neither (a) nor (b).

2. Suppose that we regress Y on X and get estimates of the intercept b_1 and the slope coefficient b_2. The R^2 of the regression is .65. We then run the regression again but force the constant to be zero and get a new estimate for the slope coefficient, $\tilde{b_2}$. Which of the following statements will definitely be false?
   (a) $b_2 = \tilde{b_2}$.
   (b) $b_2 \neq \tilde{b_2}$.
   (c) The R^2 of the new regression is greater than .65.
   (d) The R^2 of the new regression is less than .65.

3. Suppose that population (P) grows exponentially over time (t). Which of the following equations would you use to model the relationship between populaton and time?
   (a) $ln(P) = \beta_1 + \beta_2 ln(t) + \varepsilon$.
   (b) $P = \beta_1 + \beta_2 ln(t) + \varepsilon$.
   (c) $ln(P) = \beta_1 + \beta_2 t + \varepsilon$.
   (d) $P = \beta_1 + \beta_2 t + \varepsilon$.

4. Which of the following would not decrease the standard error of b_2 (the estimated slope coefficient in a bivariate regression)?
   (a) Increasing the sample size.
   (b) Greater variation in x.
   (c) Larger magnitudes for the residuals.
   (d) None of the above would decrease the standard error.

5. If the covariance of two variables is equal to 400, we can say for certain that:
   (a) The correlation between the two variables is negative.
   (b) The correlation between the two variables is less than one but greater than zero.
   (c) The correlation between the two variables is positive.
   (d) The correlation between the two variables is equal to one.

6. Suppose that we regress life expectancy (in years) on the natural log of annual income where annual income is measured in dollars and we get a slope coefficient of 4. Which of the following is a correct interpretation of this coefficient?
   (a) An increase in annual income of one dollar is associated with a 4 percent increase in life expectancy.
   (b) An increase in annual income of one dollar is associated with a 400 percent increase in life expectancy.
   (c) A one percent increase in annual income is associated with a 4 year increase in life expectancy.
   (d) A one percent increase in annual income is associated with a .04 year increase in life expectancy.
7. Suppose we run a regression of Y on X. Which of the following would prove that a positive change in X *causes* a negative change in Y?
   (a) A positive, statistically significant slope coefficient.
   (b) A negative, statistically significant slope coefficient.
   (c) An R^2 value of 1.
   (d) None of the above.
8. If we knew the true population values of β_1 and β_2 and used these to calculate $\hat{y}_i$, the predicted value of y_i given x_i, then:
   (a) The predicted value of y_i will be equal to the actual value y_i.
   (b) The error term ε_i will be equal to zero.
   (c) The expected value of $\hat{y}_i$ will be zero.
   (d) The expected value of $\hat{y}_i$ will be equal to y_i.

The Excel output below gives the results of a regression with life expectancy in years as the dependent variable and the natural log of gross national product per capita as the independent variable. The data is a cross section of countries from the year 2006. Use the output to answer questions 9 through 13.

Regression Statistics	
Multiple R	0.810066205
R Square	0.656207257
Adjusted R Square	0.654253889
Standard Error	6.242631332
Observations	178

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9.052733817	3.211483423	2.81886363	0.005371621	2.714760977	15.39070666	2.714760977	15.39070666
ln(GNP per capita)	6.733180227	0.367359896	18.32856635	1.15581E-42	6.008182854	7.458177599	6.008182854	7.458177599

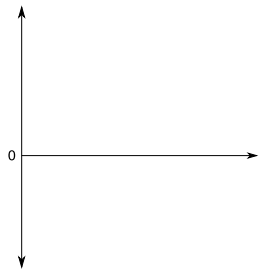
9. Which of the following is a correct interpretation of the intercept?
   (a) The predicted life expectancy for a country with zero GNP per capita would be 6.7 years.
   (b) The predicted life expectancy for a country with zero GNP per capita would be 9 years.
   (c) A best fit line on a scatterplot with life expectancy on the vertical axis and GNP per capita on the horizontal axis would intersect the vertical axis at 9 years.
   (d) A best fit line on a scatterplot with life expectancy on the vertical axis and the natural log of GNP per capita on the horizontal axis would intersect the vertical axis at 9 years.
10. We would reject the null hypothesis that there is no relationship between the natural log of GNP per capita and life expectancy at a:
   (a) 10% significance level.
   (b) 5% significance level.
   (c) 1% significance level.
   (d) All of the above.
11. The predicted average life expectancy for a country with a GNP per capita of $2000 is:
   (a) 55 years.
   (b) 57 years.
   (c) 60 years.
   (d) 62 years.
12. The 90% confidence interval for the slope coefficient:
   (a) Will contain the value 6.
   (b) Will not contain the value 6.
   (c) Will be centered at a value greater than the center of the 95% confidence interval.
   (d) Will be centered at a value less than the center of the 95% confidence interval.
13. If we ran the same regression but used the natural log of GNP as our independent variable (rather than the natural log of GNP per capita):
   (a) The new slope coefficient would be larger than the old one.
   (b) The new slope coefficient would be smaller than the old one.
   (c) The new slope coefficient would be the same as the old one.
   (d) Not enough information.
14. Suppose that test scores increase as hours of studying increase but that the increase in the test score from an additional hour of studying gets smaller and smaller as studying increases. Suppose we run a regression to estimate $S = b_1 + b_2 H + b_3 H^2$ where S is test score and H is hours of studying. We would expect:
   (a) b_1 and b_2 to be positive.
   (b) b_1 and b_2 to be negative.
   (c) b_1 to be positive and b_2 to be negative.
   (d) b_1 to be negative and b_2 to be positive.

15. If Y and X have a correlation of -1, a scatter plot with X on the horizontal axis and Y on the vertical axis will have:
    (a) All of the points either above and to the right of the point $(\bar{X}, \bar{Y})$ or below and to the left of $(\bar{X}, \bar{Y})$.
    (b) All of the points either above and to the left of the point $(\bar{X}, \bar{Y})$ or below and to the right of $(\bar{X}, \bar{Y})$.
    (c) Most but not all of the points either above and to the right of the point $(\bar{X}, \bar{Y})$ or below and to the left of $(\bar{X}, \bar{Y})$.
    (d) Most but not all of the points either above and to the left of the point $(\bar{X}, \bar{Y})$ or below and to the right of $(\bar{X}, \bar{Y})$.
16. Suppose we run a regression with height as the dependent variable and average daily caloric intake as the independent variable. If we switch from measuring height in inches to measuring height in meters (assume that the R^2 was originally less than one and greater than zero):
    (a) The standard error of the regression will increase but the R^2 will stay the same.
    (b) The standard error of the regression will decrease but the R^2 will stay the same.
    (c) The standard error of the regression and the R^2 will both increase.
    (d) The standard error of the regression and the R^2 will both decrease.
17. The total sum of squares for a regression is 100 and the R^2 is .25. What is the sum of the square of the residuals ($\sum (y_i - \hat{y}_i)^2$)?
    (a) 25.
    (b) 50.
    (c) 75.
    (d) Not enough information.
18. The predicted mean value of y when x is equal to 5 is 100. The predicted actual value for y when x is equal to 5:
    (a) Will also be equal to 100.
    (b) Will be greater than 100.
    (c) Will be less than 100.
    (d) (a), (b) and (c) are all possible.
19. Which of the following would lead to a larger slope coefficient when Y is regressed on X (assume the correlation between X and Y is positive):
    (a) A larger variance of X, holding the variance of Y and the correlation of X and Y constant.
    (b) A larger variance of Y, holding the variance of X and the correlation of X and Y constant.
    (c) A smaller correlation of X and Y, holding the variances of X and Y constant.
    (d) All of the above.
20. Which of the following is definitely true when regressing Y on X?
    (a) The regression line will pass through the point $(\bar{X}, \bar{Y})$.
    (b) The regression line will pass through the origin.
    (c) The regression line will pass through at least one data point in the sample.
    (d) The regression line will pass through all of the data points in the sample.

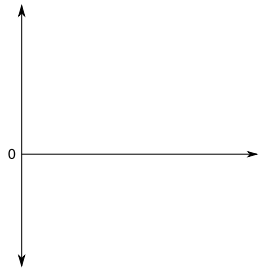
SECTION II: SHORT ANSWER (40 points)

1. (12 points) For each graph below draw a scatter plot with at least 10 data points that depicts the situation described. Next to each graph include a one sentence explanation of why the graph violates the assumption of interest. For all three questions, assume that all of the observed values of X are positive.

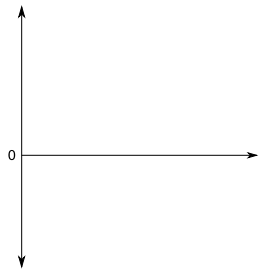
   (a) Suppose we regress Y on X and discover that our residuals violate the assumption that the error has mean zero. Draw a scatter plot with x_i on the horizontal axis and $y_i - \hat{y}_i$ on the vertical axis that depicts this situation.



   (b) Suppose we regress Y on X and discover that our residuals violate the assumption that the error is unrelated with the regressor. Draw a scatter plot with x_i on the horizontal axis and $y_i - \hat{y}_i$ on the vertical axis that depicts this situation.



   (c) Suppose we regress Y on X and discover that our residuals violate the assumption that the error has constant variance (in other words, σ_ε^2 depends on the value of x_i). Draw a scatter plot with x_i on the horizontal axis and $y_i - \hat{y}_i$ on the vertical axis that depicts this situation.



2. (18 points) The table below shows the first several rows of observations for a dataset contain income and personal characteristics for a random sample of Davis residents. The total sample size is 1000 observations. The variable GENDER can take on the values 'MALE' or 'FEMALE'. The variable RACE can take on the values 'WHITE', 'BLACK', 'ASIAN', or 'OTHER'. The variable EDU measures years of education. The variable INC measures annual income in dollars.

OBSERVATION	GENDER	RACE	EDU	INC
1	MALE	WHITE	12	45000
2	MALE	BLACK	12	47000
3	FEMALE	ASIAN	14	39000
4	MALE	WHITE	16	48000
5	FEMALE	WHITE	12	52000
6	FEMALE	WHITE	16	33000

Suppose you overhear someone make the following statement: White residents in Davis earn $5,000 more a year on average than non-white residents of Davis. You decide to test this statement by running a bivariate regression using the data described above.

(a) What, if any, data transformations would you need to do before running your regression? If you do need to use a data transformation, how would you do it in Excel?

(b) What is the regression equation you would use? Explain in words how you would interpret the values of the intercept and the slope coefficient in this equation.

(c) Write the null and alternative hypotheses you use to test the statement given in the beginning of the problem. Your hypotheses should be written in terms of the parameters of your regression equation in part (b).

(d) Assume you did the data transformations (if any were necessary) from part (a) and ran the regression you specified in part (b). Desribe the steps you would take to test the set of hypotheses given in part (c). Be certain to include your decision rule for rejecting the null hypothesis. Be as specific as possible.

3. (10 points) For each situation below, write down a regression equation that would satisfy our assumption of a linear relationship between the dependent variable and the independent variable(s).
   (a) We are interested in the effect of ocean temperature (T) on the population of fish (F). The population of fish decreases by a constant percentage everytime the ocean temperature increases by one degree.
   (b) We are interested in the effect of age (A) on the number of visits to the doctor (V) a person makes in a year. Trips to the doctor per year start out high at very young ages, then decrease as age increases up to early adulthood at which point they start increasing again.
   (c) We are interested in figuring out how the number of burritos purchases (B) is influenced by the price of a burrito (P). We want to estimate the elasticity of demand which is the percent change in quantity associated with a one percent increase in price.