Ecn 102 - Analysis of Economic Data
University of California - Davis May 19, 2009
Instructor: John Parman

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 for full credit.

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

(POTENTIALLY) USEFUL FORMULAS AND EXCEL OUTPUT

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

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

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

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

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

$\mu = E(X)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$E(b_2) = \beta_2$

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

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

$t^* = \frac{b_2 - \beta_2^*}{s_{b_2}}$

TINV(.005,50)=2.94	TINV(.02,50)=2.40	TINV(.01,49)=2.68
TINV(.01,50)=2.68	TINV(.005,49)=2.94	TINV(.02,49)=2.40

SECTION I: MULTIPLE CHOICE (60 points)

1. Which of the following statements is always true?

   (a) The correlation between two variables is greater than or equal to the covariance.
   (b) The covariance between two variables is greater than or equal to the correlation.
   (c) The correlation between x and y will have the same sign as the slope coefficient from a regression of y on x.
   (d) The correlation between x and y will have the opposite sign of the slope coefficient from a regression of y on x.

2. Which of the following would reduce the standard error of the slope coefficient?

   (a) Decreasing the number of observations.
   (b) Greater variation in the independent variable.
   (c) Having data points that are further from the regression line.
   (d) None of the above.

3. Suppose that you run a regression of y on x. Your 95% confidence interval for the slope cofficient turns out to be (2.8, 3.2). Which of the following statements are definitely true?

   (a) The slope coefficient is statistically significant at the 5% significance level.
   (b) The slope coefficient is economically significant at the 5% significance level.
   (c) The slope coefficient is economically significant at the 10% significance level.
   (d) (a) and (b).

4. Which of the following statements is true?

   (a) The regression line will always pass through at least one data point.
   (b) The regression line will pass through the point $(\bar{x}, \bar{y})$.
   (c) The regression line will always pass through the origin.
   (d) (a) and (b).

5. Suppose there is an exponential growth relationship between x and y (y is growing exponentially as x increases). which of the following regression equations would you use to estimate the relationship between x and y?

   (a) $y = b_1 + b_2 x$.
   (b) $y = b_1 + b_2 \ln x$.
   (c) $\ln y = b_1 + b_2 x$.
   (d) $\ln y = b_1 + b_2 \ln x$.

6. If two variables are perfectly correlated, when we regress one variable on the other, which of the following will definitely be true?

   (a) The estimated intercept will be equal to zero.
   (b) The estimated slope coefficient will be equal to one.
   (c) The R^2 will be equal to one.
   (d) The covariance will be equal to one.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	11.7	0.35	33.07	5.23E-83	10.98	12.37
HUMIDITY	-0.05	0.0063	-7.22	1.04E-11	-0.058	-0.033

Use the regression output above to answer questions (7) through (10). The output corresponds to a regression of visibility (measured in miles) on relative humidity (measured as a percentage). The observations are daily observations over a seven month period. A day with 10 miles of visibility and a relative humidity of 30% would be entered as $(10, 30)$ in the dataset.

7. Which of the following conclusions would you make based on the regression results?
   (a) There is a positive, statistically significant (at a 5% significance level) relationship between humidity and visibility.
   (b) There is a positive but statistically insignificant (at a 5% significance level) relationship between humidity and visibility.
   (c) There is a negative, statistically significant (at a 5% significance level) relationship between humidity and visibility.
   (d) There is a negative but statistically insignificant (at a 5% significance level) relationship between humidity and visibility.
8. What is the predicted visibility when relative humidity is 50%?
   (a) 585 miles.
   (b) 9.2 miles.
   (c) 5.8 miles.
   (d) 11.7 miles.
9. Suppose b_2^{upper} is the upper end of the 90% confidence interval for the slope coefficient. Which of the following is definitely true?
   (a) $b_2^{upper} > -.033$.
   (b) $b_2^{upper} < -.033$.
   (c) $b_2^{upper} > 12.37$.
   (d) Not enough information.
10. If the average humidity in the sample was 80%, what was the average visibility?
   (a) 11.7 miles.
   (b) 12.7 miles.
   (c) 15.7 miles.
   (d) 7.7 miles.

11. Suppose that we run a regression of heart rate on days of exercise per week and get a negative t-stat for the slope coefficient. If the magnitude of the t-stat is large enough to reject the null hypothesis that the coefficient is zero at a 5% significance level, which of the following statements may be false?
    (a) We would reject the null hypothesis that the coefficient is greater than or equal to zero at a 5% significance level.
    (b) We would reject the null hypothesis that the coefficient is greater than or equal to zero at a 2.5% significance level.
    (c) We would reject the null hypothesis that the coefficient is less than or equal to zero at a 5% significance level.
    (d) Both (b) and (c) may be false.
12. Which of the following would not lead to a wider confidence interval for the slope coefficient when regressing y on x?
    (a) Making the significance level (α) larger.
    (b) A larger standard error for b_2.
    (c) A smaller sample variance for x.
    (d) A smaller sample size.
13. The ordinary least squares estimator for β_2 (the one we have used in class and in Excel):
    (a) Minimizes the sum of $(y_i - \hat{y}_i)$.
    (b) Minimizes the sum of $(y_i - \bar{y})$.
    (c) Minimizes the sum of $(y_i - \hat{y}_i)^2$.
    (d) Minimizes the sum of $(y_i - \bar{y})^2$.

    Use the following information to answer questions (14) and (15). Suppose that we regress the number of races a marathon runner wins on the average time it takes the runner to finish a marathon in minutes and get the estimates for the intercept and slope coefficient, b_1 and b_2 respectively. Now we run the same regression but measure the average finishing time in hours instead of minutes and get a new set of estimated coefficients $\tilde{b_1}$ and $\tilde{b_2}$.
14. Which of the following will be true?
    (a) $b_2 = \tilde{b_2}$.
    (b) $|b_2| > |\tilde{b_2}|$.
    (c) $|b_2| < |\tilde{b_2}|$.
    (d) Not enough information.
15. Which of the following will be true?
    (a) $b_1 = \tilde{b_1}$.
    (b) $|b_1| > |\tilde{b_1}|$.
    (c) $|b_1| < |\tilde{b_1}|$.
    (d) Not enough information.

16. Suppose that for a given sample, the sample variance of y is 400, the sample variance of x is 100 and the slope coefficient from regressing y on x is 1. What is the correlation between x and y for this sample?

    (a) 1.
    (b) $\frac{2}{5}$.
    (c) $\frac{1}{2}$.
    (d) It depends on the estimated intercept from the regression.

17. Suppose that days of exercise increase early in life and then decrease later in life. If you had data on exercise and age and wanted to model the relationship between exercise and age and estimate it with ordinary least squares, what data transformations would you use?

    (a) Take the natural log of exercise.
    (b) Take the natural log of age.
    (c) Take the natural log of exercise and the natural log of age.
    (d) Use a polynomial in age.

18. The residual for a particular observation (x_i, y_i) is the:

    (a) The vertical distance between (x_i, y_i) and the regression line.
    (b) The vertical distance between (x_i, y_i) and $(\bar{x}, \bar{y})$.
    (c) The horizontal distance between (x_i, y_i) and the regression line.
    (d) The horizontal distance between (x_i, y_i) and $(\bar{x}, \bar{y})$.

19. Which of the following is not an assumption we made when doing bivariate statistical inference?

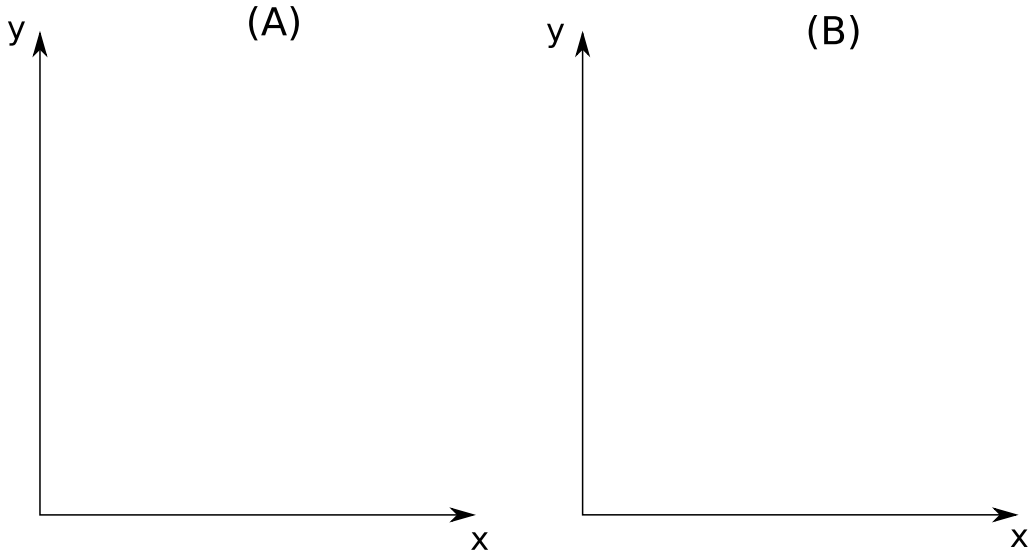
    (a) The population model relating y to x is a linear function.
    (b) The error has mean zero.
    (c) The variance of the error terms is zero if the sample size is large.
    (d) The errors for different observations are unrelated.

20. Dummy variables are used when:

    (a) The scale of the independent variables is very different for different observations.
    (b) There is an exponential relationship between x and y.
    (c) One of our variables of interest is a categorical variable.
    (d) We have a smooth but nonlinear relationship between x and y.

SECTION II: SHORT ANSWER (40 points)

1. (8 points) On the graphs below, sketch two scatter plots with regression lines (you do not need to include any numbers on your graphs). You should include at least ten data points on each scatter plot. The two graphs should represent data that would give the same estimated slope coefficient and estimated intercept, have the same sample range for x, and have a negative covariance between y and x. The one difference is that the regression line for graph A should have a lower R^2 than the regression line for graph B.



2. (14 points) The following descriptive statistics were calculated in Excel from a dataset of annual GDP growth rates and unemployment rates for the United States. The sample consists of annual observations from the year 1960 to the year 2007. The first two tables contain descriptive statistics for the unemployment rate and the growth rate of GDP, respectively. The third table contains correlations between the variables.

unemployment rate	
Mean	5.84
Standard Error	0.21
Median	5.6
Mode	5.5
Standard Deviation	1.42
Sample Variance	2.03
Kurtosis	0.59
Skewness	0.73
Range	6.2
Minimum	3.5
Maximum	9.7
Sum	280.4
Count	48

gdp growth rate	
Mean	3.31
Standard Error	0.29
Median	3.45
Mode	2.5
Standard Deviation	1.99
Sample Variance	3.98
Kurtosis	0.15
Skewness	-0.48
Range	9.1
Minimum	-1.9
Maximum	7.2
Sum	158.8
Count	48

	unemployment rate	*gdp growth rate*
unemployment rate	1	
gdp growth rate	-0.26	1

Suppose that you want to estimate the following relationship:

$$g_t = \beta_1 + \beta_2 urate_t + \varepsilon_t \tag{1}$$

where g_t is the growth rate of GDP in year t, $urate_t$ is the unemployment rate in year t and ε_t is an error term.

(a) List two properties that the error terms should have if we want to get unbiased, consistent estimates β_1 and β_2 using ordinary least squares.

(b) Calculate the value of the slope coefficient, b_2, you would get from regressing the growth rate of GDP on the unemployment rate. Explain in words what the meaning of this value is.

(c) Calculate the value of the intercept, b_1, you would get from regressing the growth rate of GDP on the unemployment rate. Explain in words what the meaning of this value is.

3. (18 points) Below is the regression output from regressing the number of prisoners per 100,000 people (*PRISON*) on the number of police officers per 100,000 people (*POLICE*). The unit of observation is a state (there is an observation for each of the 50 states and one observation for the District of Columbia).

Regression Statistics	
R Square	0.7041291
Standard Error	97.9046561
Observations	51

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-112.14	39.82	-2.82	0.00698	-192.15	-32.12
POLICE	1.40	0.13	10.8	1.48E-14	1.14	1.66

(a) Explain what the R^2 for this regression is telling us.

(b) Is there a statistically significant positive relationship between the number of police per 100,000 people and the number of prisoners per 100,000 people at a 5% significance level? Justify your answer.

(c) Use an upper one-tailed test and a 1% significance level to test whether the slope coefficient is greater than 1. Be certain to clearly state your null and alternative hypotheses, show your calculations, and clearly state your conclusions.

(d) Explain one reason that we may expect the direction of causality to be from the number of police to the number of prisoners (that is, an increase in the number of police would cause an increase in the number of prisoners).

(e) Explain one reason that we may expect the direction of causality to go in the other direction.