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

February 24, 2010

Midterm 2 - Solutions

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.

   (c) The sign of the slope coefficient will be the same as the sign of the correlation between the two variables. Knowing the sign of the correlation does not give you enough information to determine the sign of the intercept.

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.

   (b) In general, if our independent variable is more spread out, we will be able to getter a better estimate of the slope. Decreasing the sample size or having a larger error sum of squares (and therefore a larger standard error for the regression) would both increase the standard error of the slope coefficient.

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.

(d) The predicted amount of time can be calculated by plugging three in for the number of children in the regression equation:

$$time = 9.86 + 31.58 \cdot 3$$

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.

(d) To account for the relationship between time and children being increasing at first and then decreasing, we would need to include an additional term for number of children squared. A linear, log-linear or linear-log model would all correspond to situations where the sign of the relationship does not change.

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: $\beta_2 \leq 0$.

(a) The intercept is telling us the time spent caring for household members for a person with zero children. Therefore we would need to test whether this intercept is different than zero.

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.

(b) Notice that the 95% confidence interval for the intercept is wider than the 95% confidence interval for the slope coefficient. This means that the 90% confidence interval for the intercept will be wider than the 90% confidence interval for the slope coefficient (switching from 95% to 90% scales down both intervals by the same factor).

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.

   (a) The chosen significance level determines the probability of a Type I error (the probability is exactly equal to the significance level). So increasing the significance level will increase the probability of a Type I error. Changing the sample size will not affect the probability of a Type I error since it would not change α.

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.

   (a) The two variables are perfectly correlated so they will lie along a straight line. The slope of that line will have the same sign as the correlation.

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, $\tilde{b_2}$. 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.

   (d) Since all of the data points have positive values for both X and Y, any line we fit through the data points that passes through the origin will have a positive slope. So $\widetilde{b_2}$ must be positive. If the correlation between X and Y is negative, b_2 will be negative. So if the correlation is negative, b_2 will be less than $\widetilde{b_2}$.

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.

(d) Note that are basically trying to fit a straight line through a parabola. The slope of that straight line will almost certainly be different than the coefficient in front of X^2 that defines the parabola.

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

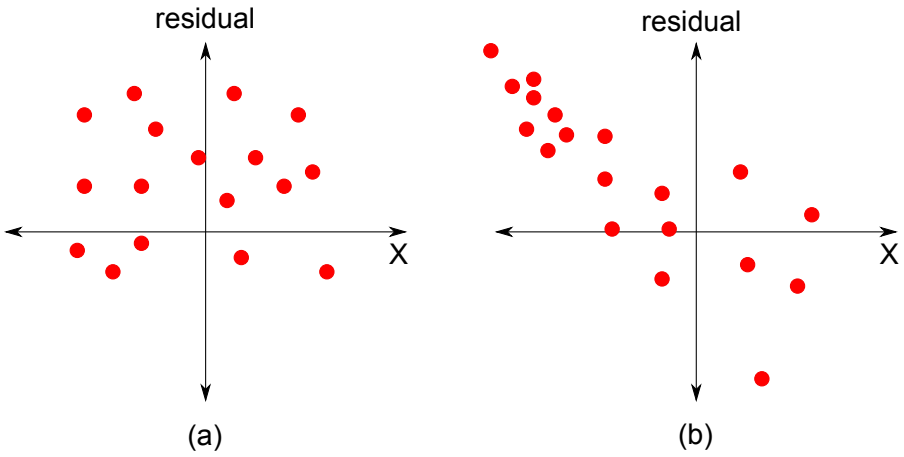
(a) The covariance between X and Y in the sample will be less than zero.
(b) If we regress Y on X, the error sum of squares will be greater than or equal to zero.
(c) If we regress Y on X, the R^2 will be positive.
(d) If we regress Y on X, the error sum of squares will be greater than or equal to the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

(d) The correlation and the covariance will always have the same sign. Because the correlation is not zero, the R^2 will not be zero, meaning that the error sum of squares must be less than 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.

(d) The coefficient is telling us that a one unit change in the log of weight is associated with a 0.10 unit change in the log of calories. A one unit change in the log of weight is equivalent to a one hundred percent change in weight. A 0.10 unit change in the log of calories is equivalent to a ten percent change in calories. So a one hundred percent change in weight is associated with a ten percent change in calories, or more simply a ten percent change in weight is associated with a one percent change in calories.



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).

(b) The residuals do not appear to be correlated with X. They do not systematically increase or decrease as the value of X changes. Additionally, the variation in the residuals does not appear to change as the value of X changes. The one problem with the residuals is that the mean of the residuals is greater than zero.

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).

(d) Notice that the residuals are decreasing on average as X increases. This means that the residuals are not uncorrelated with X. Also, the residuals are getting more spread out as X increases, suggesting that the variance of the residuals is correlated with X.

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.

(a) The confidence interval will be centered at the estimated slope coefficient (we don't know the true population slope coefficient). A 95% confidence interval will be wider than the 90% confidence interval and narrower than the 99% confidence interval.

16. Which of the following would definitely change the estimated value of the intercept (assume the intercept is not equal to zero)?

(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.

(b) The intercept is in the units of the dependent variable. If we change the units for the dependent variable, this will change the value of the intercept. Changing the units for the independent variable will affect the slope coefficient but not 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.

(a) The two different measures could lead to different slope coefficients and a different error sum of squares since we would be using different datasets for our X. However, the total sum of squares would stay the same as long as we are using the same data for our dependent variable ($\sum(y_i - \bar{y})^2$ will not change since all of the y_i values would be same).

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) $t_{0.10,199}$.

(b) Since we are using a one-tailed test with $n-2$ degrees of freedom (we are estimating two parameters), we would want to use $t_{\alpha,n-2}$ as the critical value.

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.

    (a) Trying to predict the actual value of y given x is much harder than trying to predict the expected value of y given x. The standard error for the predicted actual value will always be larger than the standard error for the predicted expected value.

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.

    (a) The value of the intercept is chosen such that the regression line passes through $(\bar{y}, \bar{x})$. The regression line may pass through the point (μ_y, μ_x) but most times it will not.

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.

   The slope coefficient is telling us that using air conditioning is associated with getting 0.02 fewer miles per gallon relative to not using air conditioning. It seems reasonable that this is a negative relationship: using the air conditioning requires additional energy, requiring more gas and reducing the miles per gallon that you get.

   (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.

   At any reasonable significance level, we would agree with the researcher. The p-value for the air conditioner coefficient is equal to 0.005, meaning that we would reject the null hypothesis that the air conditioner has no effect on fuel economy at any significance level greater than 0.5%. So there is a statistically significant relationship between air conditioner use and fuel economy.

   (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.

   Most people would disagree with this statement. Notice that using the air conditioner is associated with a 0.02 mile per gallon reduction in fuel economy. To gauge how whether this is a meaningful reduction, we can consider miles per gallon without using the air conditioner and with using the air conditioner. Without using the air conditioner, a car would get 23.4 miles per gallon (the intercept from the regression results). With the air conditioner on, a car would get 23.38 miles per gallon (the intercept plus the slope coefficient). It is clear from these numbers that the change in fuel economy from air conditioner use is incredibly small relative to the overall levels of fuel economy we are considering. No reasonable person would be concerned about getting 23.38 miles per gallon instead of 23.4 miles per gallon, so the magnitude of the slope coefficient is not economically significant.

Note that given the slope coefficient's p-value, the confidence interval for the slope coefficient is very small. So either end of the confidence interval would lead us to the same conclusion that the coefficient is not economically significant. If the confidence interval were large, we would want to consider whether the values at either end of the interval could be considered economically significant. In this case we don't need to bother since all of the values would be very small.

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.

   $$\ln C = \beta_1 + \beta_2 T + \varepsilon$$

   Since we think that ice cream cones increasing by a constant percentage for every one unit increase in temperature, we should think of the natural log of ice cream cones as a linear function of temperature. This way the slope coefficient would give us the change in the log of ice cream cones with a one degree increase in temperature which would tell us the percent change in ice cream cones with a one degree increase in temperature.

   (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.

   $$S = \beta_1 + \beta_2 N + \beta_3 N^2 + \varepsilon$$

   A polynomial in N allows us to capture the fact that the magnitude of the change in the average speed with a change in the number of cars is dependent on current number of cars (β_3 will capture the change in the slope as the number of cars increases).

   (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.

   $$\ln C = \beta_1 + \beta_2 T + \varepsilon$$

   This is a case of exponential growth. We could write C as a function of $e^{\beta_2 T}$ but we wouldn't be able to use the regression function in Excel to handle this. By taking the log of both sides, we can transform the equation into a linear model relating $\ln C$ to T that can be estimated with the regression function.

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)

  We are trying to test whether or not the average sale prices in the months of March, April and May are 10% higher than the sale prices in other months. To test this, we could follow the following procedure:

  - First, we need to transform our data so that we have the appropriate variables to estimate the relationship of interest. First, we need a new variable that tells us whether a particular observation corresponds to spring or a different season. We can do this by creating a dummy variable called $SPRING$ that is defined as follows:

$$SPRING = 1 \text{ if Month} \in (\text{March, April, May})$$

$$SPRING = 0 \text{ otherwise}$$

We also need to transform sale price so that we can look at percent changes in sale prices. We can do this by calculating a new variable that is equal to the natural log of average sale price:

$$LNPRICE = ln(\text{Average sale price})$$

- Now we can write out the population relationship we are trying to estimate in terms of these new variables:

$$LNPRICE = \beta_1 + \beta_2 SPRING + \varepsilon$$

- Given this population relationship we can write out the hypothesis we are trying to test in terms of the parameters of the above equation:

$$H_0\text{: } \beta_2 \leq 0.10$$

$$H_a\text{: } \beta_2 > 0.10$$

Notice that we are trying to determine whether the slope coefficient is greater than 0.10. This is the slope coefficient that would correspond to spring prices being 10% higher than prices in other seasons. It would be incorrect to write the null hypothesis as $\beta_2 \leq 10$ (this would be testing whether sale prices in the spring are 1000% higher than in other seasons).

- Now we have everything in place to run our regression in Excel. We would use the regression option from the data analysis toolpack and set $LNPRICE$ as our dependent variable and $SPRING$ as the independent variable. Excel would then provide us with the regression results.
- To perform our hypothesis test, we would first use the estimated slope coefficient (b_2) and its standard error (s_{b_2}) reported in the regression results to calculate the following test statistic:

$$t^* = \frac{b_2 - 0.10}{s_{b_2}}$$

- Next we would need to calculate the p-value for this test statistic. We are doing an upper one-tailed test so our p-value would be calculated in Excel as:

$$p = TDIST(t^*, 130, 1) \text{ if } t^* > 0$$

$$p = 1 - TDIST(|t^*|, 130, 1) \text{ if } t^* < 0$$

Notice that we are using 130 for the degrees of freedom (11 years of monthly data gives us 132 observations so $n - 2$ is 130) and one tail (since we are doing a one-tailed test).

- If this p-value is less than our chosen signficance level, we would reject the null hypothesis that the sale prices in spring are less than or equal to 10% higher than prices in other seasons in favor of the alternative hypothesis that sale prices in spring are over 10% higher in spring than in other seasons. So

if the p-value is less than our chosen significance level, we will sell in spring. Otherwise, we will sell in the summer.