Ecn 102 - Analysis of Economic Data
University of California - Davis March 17, 2010
Instructor: John Parman

Final Exam - Solutions

You have until 12:30pm 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

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

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

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

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

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

$\mu = E(X)$

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

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

$t^* = \frac{b_j-\beta_j}{s_{b_j}}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\hat{y}_i = b_1 + \sum_{j=2}^{k} b_j x_{j,i}$

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

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

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

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

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

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

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

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

$Pr(F_{k-g,n-k} > F^*) = FDIST(F^*, k-g, n-k)$

SECTION I: MULTIPLE CHOICE (60 points)

1. Suppose that we regress Y on X. In which of the following scenarios would the estimated slope coefficient not be biased?

   (a) There is random measurement error in X.
   (b) There is random measurement error in Y.
   (c) There is an omitted variable Z that is correlated with both X and Y.
   (d) Both (b) and (c).

   (b) Measurement error in X would bias the slope coefficient toward zero. Omitting the variable Z would bias the coefficient in a direction that depends on the signs of the correlations between Z and X and between Z and Y. Measurement error in Y would not bias the coefficient, it would just lead to a larger standard error for the coefficient.

2. Which of the following would be the best graph for showing the distribution of heights in a sample of one thousand students?

   (a) Scatter plot.
   (b) Line chart.
   (c) Histogram.
   (d) Bubble chart.

   (c) A histogram shows the distribution of a single variable.

3. Suppose $\tilde{b_j}$ is an estimator for the slope coefficient β_j. The expected value of $\tilde{b_j}$ is equal to β_j plus ten. As the sample size gets larger and larger, the standard error of $\tilde{b_j}$ gets closer and closer to zero while the expected value of $\tilde{b_j}$ remains the same. Which of the following statements is true?

   (a) $\tilde{b_j}$ is an unbiased, consistent estimator of β_j.
   (b) $\tilde{b_j}$ is a consistent but biased estimator of β_j.
   (c) $\tilde{b_j}$ is an unbiased estimator fo β_j but is not consistent.
   (d) $\tilde{b_j}$ is a biased estimator of β_j and is not consistent.

   (d) The estimator is biased because its expected value is not equal to β_j. It is not consistent because as the sample size increases, even though the standard error of the estimator approaches zero the value of the estimator does not approach β_j.

4. Suppose that on average, an extra inch in height is associated with an extra five pounds in weight. We have a dataset containing height and weight information in which height is rounded to the nearest inch and weight is rounded to the nearest pound. If we regress weight on height, the expected value of the estimated height coefficient would be:

   (a) Equal to 5.
   (b) Equal to $\frac{1}{5}$.
   (c) Less than 5.
   (d) Greater than 5.

(c) If we had an exact measure of height, the expected value of the estimated slope coefficient would be equal to five, the true value of the slope coefficient. However, the measure of height is rounded to the nearest inch, meaning that there will be random measurement error (any observation for which height is not an exact integer will be mismeasured). This measurement error will bias the slope coefficient toward zero.

5. Suppose that we run a regression of Y on X_2 and get an R^2 of 0.60. If we run a regression of Y on X_2 and X_3, we would expect:

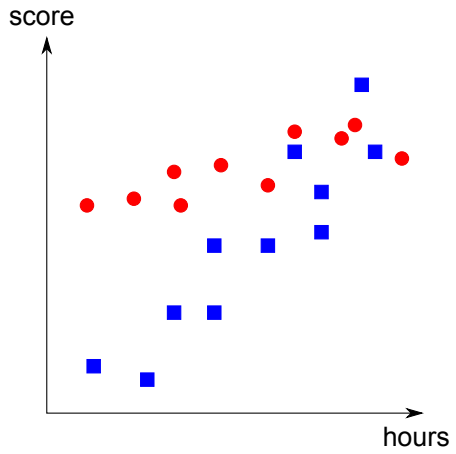
   (a) The new R^2 to be greater than or equal to 0.60.
   (b) The new R^2 to be greater than or equal to 0.60 if and only if the correlation between X_2 and X_3 is positive.
   (c) The new R^2 to be less than or equal to 0.60.
   (d) The new R^2 to be less than or equal to 0.60 if and only if the correlation between X_2 and X_3 is negative.

   (a) Adding an additional regressor will increase the R^2 if that regressor helps explain variation in Y. It cannot lower the R^2. If the variable does not help explain variation in Y, it would just get a coefficient of zero and the R^2 would remain unchanged (an additional variable should never make the fit of the model any worse than it was without that variable).

6. Suppose we have a sample of commute times for 1,000 people. If we measure commute time in minutes, the variance will be _____ if we measured commute time in hours and the mean will be _____ if we measured commute time in hours.

   (a) Smaller than, the same as.
   (b) Larger than, smaller than.
   (c) Smaller than, smaller than.
   (d) None of the above.

   (d) When switching from hours to minutes, all of our observations become 60 times as large as they were before. This means that the sample mean will now be 60 times larger and the sample variance will be 60^2 times larger than before.



Use the figure above to answers questions 7 through 10. The figure is a scatter plot with hours of study on the horizontal axis and final exam score on the vertical axis for 21 students in an ECN 100 class. The round data points correspond to economics majors. The square data points correspond to nonmajors. Suppose we use this data to estimate the following model:

$$SCORE = \beta_1 + \beta_2 MAJOR + \beta_3 HOURS + \beta_4 MAJOR \cdot HOURS + \varepsilon$$

where $SCORE$ is a student's final exam score, $MAJOR$ is a dummy variable equal to one if the student is an economics major and zero otherwise, $HOURS$ is the number of hours the student studies for the final and ε is a random error term that satisfies all of our assumptions.

7. Based on the scatterplot, we would expect our estimated value of β_2 to be:
   (a) Positive.
   (b) Negative.
   (c) Larger for economics majors than nonmajors.
   (d) Larger for nonmajors than economics majors.

   (a) β_2 will be the difference in exam scores between economics majors and nonmajors when hours of study are zero. On the graph, this is the difference between the vertical intercept for majors and the vertical intercept for nonmajors. From the graph, it is clear that the vertical intercept for majors is much larger than the vertical intercept for nonmajors, so we would expect β_2 to be positive.

8. Based on the scatterplot, we would expect our estimated value of β_4 to be:
   (a) Positive.
   (b) Negative.
   (c) Larger for economics majors than nonmajors.
   (d) Larger for nonmajors than economics majors.

   (b) β_4 is the difference between majors and nonmajors in the change in exam score from an extra hour of studying. On the graph, this is the difference in the slopes of the lines passing through the majors data points and the nonmajors data points.

From the graph, it is clear that slope is flatter for the majors than for the nonmajors. So we would expect β_4 to be negative.

9. The predicted score for an economics major who studies ten hours will be:
   (a) Greater than the predicted score for a nonmajor who studies ten hours.
   (b) Less than the predicted score for a nonmajor who studies ten hours.
   (c) Equal to the predicted score for a nonmajor who studies ten hours.
   (d) Not enough information.

   (d) Notice that if you were to draw a line through the major data points and a line through the nonmajor data points, the lines would intersect at a positive number of hours of studying. To the left of this point, the predicted score for majors would be greater than the predicted score for nonmajors. To the right of this point, the predicted score for majors would be less than the predicted score for nonmajors. We would need to know whether ten hours is to the left or the right of this point to be able to answer the question.

10. The predicted increase in score for an economics major associated with one extra hour of studying will be:
    (a) Equal to b_3, where b_3 is our estimated value of β_3.
    (b) Equal to $b_3 + b_4$, where b_3 and b_4 are our estimated values of β_3 and β_4, respectively.
    (c) Equal to $b_3 + b_4 \cdot HOURS$, where b_3 and b_4 are our estimated values of β_3 and β_4, respectively.
    (d) Equal to b_4, where b_4 is our estimated value of β_4.

    (b) For every extra hour of studying, the score for an economics major goes up by b_3 (the component of the return to studying common to both majors and nonmajors) and by b_4 (the additional return to an hour of studying for an economics major).

11. Which of the following would lead to a smaller standard error for the slope coefficient in a bivariate regression?
    (a) A smaller variance of the independent variable.
    (b) A smaller average size of the residuals.
    (c) A larger error sum of squares.
    (d) A smaller sample size.

    (b) A smaller average size of the residuals would mean a smaller standard error of the regression which also implies a smaller standard error for the slope coefficient (the standard error of the slope coefficient is proportional to the standard error of the regression).

12. Which of the following would make us less likely to reject the null hypothesis that the true population mean is 150?
    (a) Getting a sample mean that is farther from 150.
    (b) Getting a larger t statistic.
    (c) Switching to a larger value for the significance level α.
    (d) None of the above.

(d) A sample mean farther from 150 would lead to a larger t statistic making it more like that the t statistic is greater than our critical value (and we reject the null when t is greater than the critical value). Increasing the value for α would decrease the critical value, which would also make it more likely that the t statistic is greater than the critical value.

13. Which of the following statements is definitely not true?
    (a) Adding an additional regressor can change the estimated slope coefficients of the other regressors.
    (b) Adding an additional regressor can decrease the R^2 of the regression.
    (c) Adding an additional regressor can lower the error sum of squares of the regression.
    (d) Adding an additional regressor can increase the standard errors of the estimated slope coefficients of the other regressors.

    (b) If the additional regressor is correlated with the other regressors and with the dependent variable, leaving it out of the regression would have led to an omitted variable bias on the other regressors. If the additional regressor helps explain variation in Y, adding it to the regression would lower the error sum of squares. If the additional regressor does not help explain variation in Y, it won't reduce the R^2 but it may increase the standard errors of the other coefficients.

14. Suppose we want to estimate the effect of coffee on hours of sleep. We believe that every extra cup of coffee a person drinks decreases her hours of sleep by a constant percentage. To estimate the relationship of interest, we would:
    (a) Run a regression with minutes of sleep as the independent variable and the log of cups of coffee as the dependent variable.
    (b) Run a regression with minutes of sleep as the dependent variable and the log of cups of coffee as the independent variable.
    (c) Run a regression with the log of minutes of sleep as the independent variable and cups of coffee as the dependent variable.
    (d) Run a regression with the log of minutes of sleep as the dependent variable and cups of coffee as the independent variable.

    (d) Since we think that coffee causes the change in sleep, we should have coffee as the independent variable and sleep as the dependent variable. Since we think that a one unit change in coffee leads to a constant percentage change in sleep, we should use a log-linear model.

15. The 95% confidence interval for the population mean of X:
    (a) Will be wider than the 90% confidence interval for the population mean.
    (b) Will tend to get wider as we use more observations.
    (c) Will not depend on the variance of X.
    (d) None of the above.

    (a) A 95% confidence interval will always be wider than a 90% confidence interval for a variable. Increasing the sample size will narrow the confidence interval. The larger the variance of X, the wider the confidence interval for the mean of X.

16. Which of the following is not an assumption we make when doing bivariate statistical inference?
    (a) The expected value of the error term is equal to zero.
    (b) The errors are uncorrelated with the regressor.
    (c) The errors are uncorrelated with the dependent variable.
    (d) We assume all of the above.

    (c) The errors will always be correlated with the dependent variable (if the error term is larger by one unit, Y will be larger by one unit).

17. Suppose we are testing whether high school GPA and college GPA are jointly significant in a regression with log wage as the dependent variable. Which of the following statements is true?
    (a) We will reject the null hypothesis that both coefficients are equal to zero at a 5% significance level if and only if both of the variables are individually significant at the 5% significance level.
    (b) We will reject the null hypothesis that both coefficients are equal to zero at a 5% significance level if and only if at least one of the two variables is individually significant at the 5% significance level.
    (c) We will reject the null hypothesis that both coefficients are equal to zero at a 5% significance level if one of the two variables is individually significant at the 5% significance level.
    (d) None of the above.

    (c) If we can show that one of the variables is significant at a 5% significance level, then we can clearly say that at least one of the two variables has a coefficient different than zero (in other words, the two variables are jointly significant). However, it is not necessary that at least one of the variables is individually significant for the variables to be jointly significant. Particularly in the case of highly correlated variables (such as high school and college GPA), it is possible that the two variables could be jointly significant while neither variable is individually significant.

18. Suppose that we run a regression with the log of pounds of rice purchased from a store as the dependent variable and log of the price of a pound of rice as the independent variable. A slope coefficient of 0.3 would be interpreted as:
    (a) A one dollar increase in the price of a pound of rice is associated with a 0.3 pound increase in the amount of rice purchased.
    (b) A one percent increase in the price of a pound of rice is associated with a 30 percent increase in the amount of rice purchased.
    (c) A one percent increase in the price of a pound of rice is associated with a 0.3 percent increase in the amount of rice purchased.
    (d) A one dollar increase in the price of a pound of rice is associated with a 30 percent increase in the amount of rice purchased.

    (c) Since both variables are in logs, the slope coefficient can be interpreted as the percent change in the dependent variable with a one percent change in the independent variable.

19. Suppose that the slope coefficient for a particular regressor X_j has a p-value of 0.02. We would conclude that the coefficient is:

    (a) Economically significant at a 5% significance level.
    (b) Economically significant at a 10% significance level.
    (c) Statistically significant at a 10% significance level.
    (d) All of the above.

    (c) Since the p-value is smaller than 0.10, we would conclude that the coefficient is statistically significant at a 10% significance level. Without knowing the magnitude of the coefficient and what the coefficient is measuring, we have no way of saying whether the coefficient is economically significant.

20. Adding an irrelevant variable to a regression will tend to lower:

    (a) The R^2.
    (b) The adjusted R^2.
    (c) The standard errors of the coefficients for the other variables.
    (d) The total sum of squares.

    (b) The R^2 will not decrease (it will likely stay the same since the new variable will not explain any of the variation in the dependent variable). The adjusted R^2 will tend to decrease since it takes into account not only the R^2 but also the number of variables used. The standard errors of the other coefficients will tend to increase when you add an irrelevant variable to the regression.

21. If X and Y are perfectly correlated, when we regress Y on X we would get:

    (a) A slope coefficient equal to one.
    (b) A slope coefficient equal to either one or negative one.
    (c) An R^2 equal to one.
    (d) An intercept equal to zero.

    (c) Knowing that X and Y are perfectly correlated tells us the the data points will all lie along a straight line, giving us an R^2 value of one. However, it does not tell us what the slope of that line is or whether it has a nonzero intercept.

22. The distribution of the sample mean of X will:

    (a) Be centered at zero.
    (b) Be centered at the population mean of X.
    (c) Have a variance equal the variance of X.
    (d) Both (b) and (c).

    (b) The distribution of the sample mean will always be centered at the true population mean of X. The variance of the distribution of the sample mean will depend on both the variance of X and on the sample size.

23. Suppose that the true relationship between Y and X is given by:

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

where ε is a random error term that meets all of our assumptions. Which of the following is not a random variable?

(a) b_2, the estimated value of the slope coefficient.
(b) β_2.
(c) The sample mean of Y.
(d) The error sum of squares from a regression of Y on X.

(b) β_2, the true value of the slope coefficient, is a constant. b_2, the estimate of the value, will be a random variable that will depend on the sample we happen to draw. The same thing is true of the sample mean of Y and the error sum of squares, both of these values will vary from sample to sample.

24. Suppose we have a categorical variable that can take on eight different values. We want to control for this variable in a wage regression. We would need to include:

(a) One continuous variables in our regression to do this.
(b) Seven different continuous variables in our regression to do this.
(c) Seven different dummy variables in our regression to do this.
(d) Eight different dummy variables in our regression to do this.

(c) To account for differences between all eight different groups, we would need to create a dummy variable for each group. If we include dummy variables for all eight groups, we run into a multicollinearity problem. We have to drop one of the dummy variables when running the regression, leaving us with seven different dummy variables.

25. Suppose that we can reject the null hypothesis that the population mean is less than or equal to 10 at a 5% significance level. Which of the following statements is definitely true?

(a) We can reject the null hypothesis that the population mean is equal to 10 at a 5% significance level.
(b) We can reject the null hypothesis that the population mean is greater than or equal to 10 at a 5% significance level.
(c) We can reject the null hypothesis that the population mean is equal to 10 at a 1% significance level.
(d) None of the above.

(d) If we can reject the null hypothesis that the population mean is less than or equal to 10 at a 5% significance level we must have gotten a sample mean larger than 10 that led to a t-stat greater than $t_{0.05,n-1}$. This means that we could also reject the null hypothesis that the population mean is equal to 10 at a 10% significance level (since the t-stat would be greater than $t_{\frac{0.1}{2},n-1}$). We cannot say whether we could reject this null hypothesis for an α less than 0.10.

SECTION II: SHORT ANSWER (40 points)

1. (18 points) For each scenario below, a researcher is attempting to estimate a slope coefficient of particular interest. Explain whether the researcher will get an unbiased estimate of the slope coefficient and, if not, what direction the bias will be in. Note that there may be multiple correct answers for each question. If you can properly justify your answer, you will receive full credit.

   (a) A researcher wants to estimate the effect of winter on happiness for the population of the United States. To do this, a researcher asks 1,000 people from Southern California how happy they are on a scale of 0 to 100 (with 100 being the happiest). The researcher creates a dummy variable that equals one if a person was asked this question in the winter months and equal to zero otherwise. To estimate the effect of winter on happiness, the researcher regresses the happiness number on this dummy variable for winter.

   There is a potential problem with sample selection bias in this scenario. The researcher wants to study the relationship between winter and happiness for the entire US population. The relationship between winter and happiness for Southern Californians is likely very different than the relationship for people in other parts of the country, particularly parts of the country that have more severe winter weather. If you think that the cold and snow associated with winter decreases people's happiness, you would expect a negative coefficient on the winter dummy and you expect that the magnitude of the effect will be larger for people from colder parts of the country. So using a sample of Southern Californians who experience mild winters will lead to an underestimate of the effect of winter on happiness. In this case of a negative true coefficient, underestimating the coefficient implies an upward bias.

   (b) A high school wants to know if assigning more reading leads to higher test scores. To determine this, the school takes a random sample of one hundred students and regresses their test scores on the number of pages they were assigned to read. Teachers who assign more reading also tend to spend more time preparing their lectures and answering their students' questions.

   There is an omitted variable bias resulting from not controlling for teacher quality. The coefficient on reading will pick up the direct effect of extra reading on test scores but also the indirect effect of more reading being associated with better teachers and better teachers being associated with higher test scores. Since the sign of the correlation between assigned reading and teacher quality is positive and the sign of the correlation between teacher quality and test scores is likely to be positive, the overall sign of the bias will be positive. So we have an upward bias on the assigned reading coefficient.

   (c) A student wants to know the effect of hours of exercise per week on her weekly quiz scores. To do this, the student looked up all of her quiz scores for the past year and tried to remember how many hours she exercised each week for the past year. She knows the exact quiz scores but can only remember roughly (not exactly) how many hours she exercised each week. To determine the effect of exercise on quiz performance, she regresses quiz score on hours of exercise per week. She uses 52 data points for this regression (one for every week in the past year).

Here the problem is the measurement of hours of exercise. Since the student does not remember the exact value for each week, this variable will be measured with some random error. Random measurement error in the independent variable will bias the coefficient on that variable toward zero. If the effect of exercise on quiz score is positive, this would mean there is a downward bias. If the effect of exercise on quiz score is negative, it would mean there is an upward bias.

2. (12 points) A researcher is interested in how outside temperature influences electricity usage. The researcher thinks that at very low temperatures, people use a lot of electricity to run their electric space heaters. As outside temperature rises, people use their space heaters less frequently and electricity usage goes down. However, as temperatures get very hot, electricity usage begins to rise again as people use their air conditioners more and more. To estimate this relationship between outside temperature and electricity usage, the researcher decides to use the following model:

$$E = \beta_1 + \beta_2 T_c + \beta_3 T_f + \varepsilon \tag{1}$$

where E is total electricity usage over a one week period measured in kilowatt-hours, T_c is the average temperature over that week measured in degrees Celcius, and T_f is the average temperature over that week measured in degrees Fahrenheit. The data the researcher uses come from a cross-section of US households, each observation represents one household.

(a) Explain two ways in which the researcher has misspecified the model. Explain what changes you would make to the model to deal with with these problems.

There are two major problems with the way the researcher has specified the model. First, the two different measures of temperature will be perfectly correlated. One of the two measures needs to be dropped from the equation. The second problem is that the way the researcher thinks temperature affects electricity usage is nonlinear. If you think of a graph with electricity usage on the vertical axis and temperature on the horizontal axis, the description above implies a U-shaped curve for the relationship, with electricity usage first falling and then rising as temperature increases. To model this, the researcher needs to include a polynomial in temperature. This leads us to the following specification:

$$E = \beta_1 + \beta_2 T_c + \beta_3 T_c^2 + \varepsilon$$

(b) Suppose that high income households use more electricity at any given temperature than low income households. The average difference in electricity usage between high and low income households is the same at every temperature level. Given this information, what is the correct population model (assume that the dataset also contains a variable I measuring household income)?

This information suggests that the curve relating electricity usage to temperature shifts up as income increases. To account for this we would need to include an income term in our equation. Since the shift is that same at all temperatures, there is no need to include an interaction term between income and temperature. Including income changes our model to:

$$E = \beta_1 + \beta_2 T_c + \beta_3 T_c^2 + \beta_4 I + \varepsilon$$

(c) Suppose you ran a regression to estimate all of the parameters in your model from part (b). Based on the information given throughout the problem, what would you predict the signs to be for each parameter?

Note that no points were deducted if you could not figure out the sign of β_2. As you will see in the explanation, determining the sign of β_2 is a bit trickier than determining the signs of the other coefficients.

β_1 is the electricity used by a household with no income ($I = 0$) at a temperature of zero degrees Celcius ($T_c = 0$). This certainly will not be a negative number (the household cannot use negative amounts of electricity). Most likely, it will be a positive number since even if a family doesn't have a current source of income they will likely still need to use some electricity.

The signs of β_2 and β_3 need to give us the U-shape originally described in the model. Since the U-shape faces upward β_3 must be positive. Another way to see this is that the slope of the curve would be equal to $\beta_2 + 2\beta_3 T_c$ and the slope gets more positive as T_c gets larger, so β_3 must be positive.

Note that the minimum of this curve will likely occur at a temperature well above zero degrees since it starts rising again because of air conditioner usage. The minimum of the curve is where the slope is zero, implying that $\beta_2 = -2\beta_3 T_c$. Since the minimum occurs at a positive T_c and we just determined that β_3 is positive, this implies that β_2 should be negative.

Finally, we are told in part (b) that higher income families use more electricity, so β_4 should be positive.

SUMMARY OUTPUT: API score as dependent variable

Regression Statistics	
Multiple R	0.379
R Square	0.144
Adjusted R Square	0.143
Standard Error	76.927
Observations	6426

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	688.49	5.93	116.09	0	676.86	700.11
CLASSSIZE	3.91	0.21	18.77	1.41E-76	3.50	4.32
YEARROUND	-21.82	4.90	-4.46	8.5E-06	-31.43	-12.22
NONHSGRAD	-88.76	3.56	-24.94	4.1E-131	-95.73	-81.78

SUMMARY OUTPUT: API score as dependent variable

Regression Statistics	
Multiple R	0.234
R Square	0.055
Adjusted R Square	0.055
Standard Error	80.801
Observations	6426

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	671.78	6.19	108.53	0	659.65	683.91
CLASSSIZE	4.22	0.22	19.33	5.42E-81	3.80	4.65

SUMMARY OUTPUT: API score as dependent variable

Regression Statistics	
Multiple R	0.311
R Square	0.097
Adjusted R Square	0.096
Standard Error	79.003
Observations	6426

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	798.17	1.04	765.51	0	796.13	800.22
YEARROUND	-23.39	5.03	-4.65	3.38E-06	-33.25	-13.53
NONHSGRAD	-92.40	3.65	-25.32	7.2E-135	-99.55	-85.24

Summary Statistics for the Regression Sample

Variable	Mean	Minimum	Maximum	Standard Deviation
API	789.854	310	998	83.112
CLASSSIZE	27.949	1	50	4.613
YEARROUND	0.040	0	1	0.197
NONHSGRAD	0.080	0	1	0.271

3. (10 points) Use the regression results on the previous page to answer this problem. Below the regression results is a table of summary statistics for the regression sample. The regression results are from a cross-section of California school districts. Each observation represents one school district. The variables are defined as follows:

   - API score - academic performance index score. This is a measure of the overall academic performance of students in the school district. The score can range from 200 to 1000, with 1000 being the highest possible academic performance.
   - CLASSSIZE - class size. This is the average number of students per classroom in the school district.
   - YEARROUND - dummy variable for year round schools. This variable is equal to one if schools in a district are open year round and equal to zero if schools close for the summer.
   - NONHSGRAD - dummy variable indicating districts in which many parents are not high school graduates. This variable is equal to one if over half of the parents in the district did not graduate from high school. It is equal to zero if over half of the parents in the district did graduate from high school.

   (a) Suppose that school district A has an average class size of 30, it holds classes year round, and 80% of the parents in the district are high school graduates. School district B has an average class size of 20, it does not hold classes in the summer, and 40% of the parents in the district are high school graduates. How much higher (or lower) would district A's predicted API score be compared to district B? Show all of your calculations.

   First, notice that we have several regressions to choose from. However, our choice is fairly simple. The first regression includes all of the variables and all three of the independent variables are individually statistically significant. Given that they are all signficant, they are relevant variables and should be included in the regression. So we should focus on the results of this first regression. Getting the predicted API scores for the two different school districts is simply a matter of plugging in the appropriate values for all of the variables:

$$\widehat{API_A} = 688.49 + 3.91 \cdot 30 - 21.82 \cdot 1 - 88.76 \cdot 0$$

$$\widehat{API_A} = 783.97$$

   Notice that I used one for $YEARROUND$ because school district A holds classes year round and zero for $NONHSGRAD$ because over 50% of the parents in disctrict A are high school graduates. Now for district B:

$$\widehat{API_B} = 688.49 + 3.91 \cdot 20 - 21.82 \cdot 0 - 88.76 \cdot 1$$

$$\widehat{API_B} = 677.93$$

   Notice that I used zero for $YEARROUND$ because district B does not hold classes in the summer and one for $NONHSGRAD$ because less than half of the parents in district B are high school graduates. The difference in predicted

API scores is:

$$\widehat{API_A} - \widehat{API_B} = 783.97 - 677.93 = 106.04$$

So the predicted API score in school district A is 106.04 points higher than the predicted API score in district B.

(b) Suppose you wanted to use an F test to determine whether the $YEARROUND$ and $NONHSGRAD$ variables are jointly significant in the regression that includes $CLASSSIZE$, $YEARROUND$ and $NONHSGRAD$. In other words, you want to test the following set of hypotheses:

$$H_0\text{: } \beta_{YEARROUND} = \beta_{NONHSGRAD} = 0$$

$$H_a\text{: at least one of } \beta_{YEARROUND} \text{ and} \beta_{HSGRAD} \text{ is not equal to zero}$$

Calculate the F statistic you would use to do this test. You should calculate an exact numerical value for the F statistic.

We are testing the joint significance of $YEARROUND$ and $NONHSGRAD$ in the regression that contains all three variables. So our unrestricted model is:

$$API = \beta_1 + \beta_2 CLASSSIZE + \beta_3 YEARROUND + \beta_4 NONHSGRAD + \varepsilon$$

Notice that there are four variables in this equation, so k is equal to four. Our restricted model excludes the variables that we are testing the joint significance of:

$$API = \beta_1 + \beta_2 CLASSSIZE + \varepsilon$$

Notice that there are only two variables left in the equation, so g is equal to two. The regression results for the unrestricted model are given in the first set of results on the previous page. The regression results for the restricted model are given in the second set of results. We do not need the third set of regression results for this particular question. Given the regression results, the calculation of the test statistic is straightforward:

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

$$F^* = \frac{6426-4}{4-2}\frac{0.144 - 0.055}{1 - 0.144}$$

$$F^* = 333.85$$