

## Final Exam

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 for full credit.

**Name:**

**ID Number:**

**Section:**

### (POTENTIALLY) USEFUL FORMULAS

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

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

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

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

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

$$\mu = E(X)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$E(b_j) = \beta_j$$

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

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

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

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

$$F^* = \frac{n-k}{k-1} \frac{R^2}{1-R^2} \text{ (for testing all regressors)}$$

$$F^* = \frac{n-k}{k-g} \frac{ESS_r - ESS_u}{ESS_u} = \frac{n-k}{k-g} \frac{R_r^2 - R_u^2}{1-R_u^2} \text{ (for testing subset of regressors)}$$

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

## SECTION I: MULTIPLE CHOICE (60 points)

1. If an additional explanatory variable is added to a regression:
  - (a) The  $R^2$  for the regression may decrease.
  - (b) The adjusted  $R^2$  for the regression will either stay the same or increase.
  - (c) The  $R^2$  for the regression will either stay the same or increase.
  - (d) Both (b) and (c).
    - (c) The  $R^2$  should never go down when we add another regressor, it should either stay the same or increase. The adjusted  $R^2$  can go up or down.
2. We would expect the coefficient of variation for height measured in inches:
  - (a) To be larger than the coefficient of variation for height measured in meters.
  - (b) To be smaller than the coefficient of variation for height measured in meters.
  - (c) To be equal to the the coefficient of variation for height measured in meters.
  - (d) (a), (b) and (c) could all be true.
    - (c) The coefficient of variation is a unitless measure of dispersion. Changing the units for the variable will rescale the mean and the standard deviation by the same factor. When we divide the standard deviation by the mean to get the coefficient of variation, this factor will cancel out and we'll end up with the same value for the coefficient of variation as before the change of units.
3. Suppose we ran a regression of student's high school GPA on parents' income. Which of the following would not lead to a biased coefficient for parents' income?
  - (a) Measurement error in parents' income.
  - (b) The fact that student ability is correlated with parents' income and student's GPA.
  - (c) Measurement error in student's GPA.
  - (d) Selecting students only from honors classes.
    - (c) While measurement error in the dependent variable would affect the precision of the coefficient on parents' income, it would not lead to a bias.
4. A histogram is useful for showing:
  - (a) The distribution of the observed values for a single variable.
  - (b) The relationship between a single dependent variable and a single regressor.
  - (c) The correlation between two variables.
  - (d) The precision of an estimated slope coefficient.
    - (a) A histogram shows the distribution of a single variable. All of the other answers relate to relationships between multiple variables, something that won't be captured by a single histogram.
5. Inclusion of an irrelevant variable in a regression:
  - (a) Will tend to bias the coefficients for the other variables downward and increase their standard errors.
  - (b) Will tend to bias the coefficients for the other variables upward and increase their standard errors.

- (c) Will not bias the coefficients for the other variables but will increase their standard errors.
  - (d) Will not bias the coefficients for the other variables but will decrease their standard errors.
- (c) Irrelevant variables do not lead to biased coefficients. However, they do decrease the precision of our estimates (meaning larger standard errors).
6. If we have four different types of cars in our dataset and would like to include dummy variables for car type in a regression, how many dummy variables will we use?
- (a) Four.
  - (b) Three.
  - (c) Two.
  - (d) One.
- (b) We need to use a dummy variable for each category except one omitted category. If we used a dummy variable for every category we would run into a dummy variable trap and face problems of perfectly colinear regressors.

Use the following information to answer questions (7) and (8). Suppose we run a regression of annual medical expenses ( $MED$ ) measured in dollars on age ( $AGE$ ) measured in years, hours of exercise per day ( $HOURS$ , note that this can be no larger than 24 and no smaller than 0), and age interacted with hours of exercise and come up with the following coefficient estimates:

$$MED = 10000 + 50 \cdot AGE - 500 \cdot HOURS - 1 \cdot AGE \cdot HOURS$$

7. Based on these coefficients, if predicted medical expenses were graphed as a function of hours of exercise:
- (a) The slope of the line would be steeper for a 50-year-old person compared to a 40-year-old person.
  - (b) The slope of the line would be flatter for a 50-year-old person compared to a 40-year-old person.
  - (c) The slope of the line would be the same for a 50-year-old person and a 40-year-old person but the intercepts would be different.
  - (d) The slope of the line and the value of the intercept would be the same for a 50-year-old person and a 40-year-old person.
- (a) The slope of the line would be -540 for the forty-year-old and -550 for the 50-year-old. So the magnitude of the slope is slightly larger for the 50-year-old than the 40-year-old indicating a slightly steeper line.
8. The predicted change in medical expenses associated with an increase in age of one year will be:
- (a) Positive regardless of a person's level of exercise.
  - (b) Negative regardless of a person's level of exercise.
  - (c) Could be positive or negative depending on a person's level of exercise.

- (d) Will be independent of a person's level of exercise.
- (a) The predicted change will be  $50 - 1 \cdot HOURS$ .  $HOURS$  is between 0 and 24 so this predicted change will always be positive.
9. The sample mean:
- (a) Will be equal to the population mean on average.  
 (b) Will always be equal to the population mean.  
 (c) Is a constant.  
 (d) Has a distribution centered at zero.
- (a) The sample mean is a random variable with a distribution centered at the population mean. So the expected value of the sample mean will be equal to the population mean but a single realization of the sample mean will most likely not be exactly equal to the population mean.
10. The error sum of squares:
- (a) Will be larger the better our model fit is.  
 (b) Will be larger than the total sum of squares when we have a misspecified model.  
 (c) Will always be less than or equal to the total sum of squares.  
 (d) (a) and (c).
- (c) The error sum of squares can never be larger than the total sum of squares. The total sum of squares is capturing the total amount of observed variation in  $y$ . The error sum of squares is capturing the amount of variation left over after we take out the component explained by our independent variables. At worst, the independent variables explain none of the variation in  $y$  and the error sum of squares will be equal to the total sum of squares.
11. Which of the following is always true?
- (a) The correlation between two variables has the same sign as the covariance.  
 (b) The magnitude of the correlation between two variables will be larger than the magnitude of the covariance.  
 (c) The magnitude of the correlation between two variables will be smaller than the magnitude of the covariance.  
 (d) The correlation between  $x$  and  $y$  will have the same sign as the  $R^2$  for a regression of  $y$  on  $x$ .
- (a) Correlation and covariance always have the same sign. The  $R^2$  is always positive, so in cases where the correlation between  $x$  and  $y$  is negative, the correlation and the  $R^2$  will have opposite signs.
12. Suppose that the true relationship between happiness ( $HAPPY$ , measured on a scale of 0 to 100) and wealth ( $WEALTH$ , measured in dollars) is given by:

$$HAPPY = \beta_1 + \beta_2 WEALTH + \varepsilon$$

where  $\varepsilon$  is an error term that satisfies all of our assumptions. If our measure of wealth contains some random measurement error, the magnitude of the coefficient we estimate for wealth:

- (a) Will be smaller than  $|\beta_2|$  on average.
  - (b) Will be larger than  $|\beta_2|$  on average.
  - (c) Will be equal to  $|\beta_2|$  on average.
  - (d) The answer depends on the sign of  $\beta_2$ .
    - (a) Measurement error in the independent variable will bias our coefficient toward zero, meaning that our estimated coefficient will tend to have a smaller magnitude than the true population coefficient.
13. We are more likely to reject the null hypothesis that a particular coefficient  $\beta_j$  from a multivariate regression is equal to zero when:
- (a) The standard error for the coefficient is very small.
  - (b) The  $R^2$  for the regression is close to zero.
  - (c) The estimated coefficient is very close to zero.
  - (d) The estimated coefficient divided by its standard error is very small.
    - (a) A small standard error leads to large t-stat and a smaller p-value making us more likely to reject the null hypothesis that the coefficient is equal to zero.
14. When regressing  $y$  on  $x$ , an estimated coefficient for  $x$  of 5 that is statistically significant tells us:
- (a) That a one unit increase in  $x$  is associated with a five unit increase in  $y$ .
  - (b) That a one unit increase in  $x$  causes a five unit increase in  $y$ .
  - (c) That either an increase in  $x$  causes an increase in  $y$  or an increase in  $y$  causes an increase in  $x$ .
  - (d) (a) and (c).
    - (a) Our regression coefficient tells us that there is an association between  $x$  and  $y$  but does not tell us about the direction of causality or even if one variable causes the other (it is possible that there is some third variable that is causing the changes in both  $x$  and  $y$ ).
15. Suppose that we regress  $y$  on  $x_2$  but don't include  $x_3$  leading to an omitted variable bias for the coefficient for  $x_2$ . The sign of the bias will depend on:
- (a) The sign of the correlation between  $x_2$  and  $x_3$ .
  - (b) The sign of the correlation between  $x_3$  and  $y$ .
  - (c) Both (a) and (b).
  - (d) Neither (a) nor (b).
    - (c) There will be an upward bias if the sign of the correlation between  $x_2$  and  $x_3$  is the same as the sign of the correlation between  $x_3$  and  $y$ . There will be a downward bias if the signs of the correlations are different.
16. Suppose we have data on food expenditures and income in the year 2007 for 200 different households. This is an example of:
- (a) Univariate data.
  - (b) Cross-section data.
  - (c) Time series data.

- (d) None of the above.
- (b) Each observation corresponds to a different household and all of the households are observed at the same point in time. This is an example of cross-section data. If we observed the same household over several time periods, it would be time series data. If we observed several households over several time periods, we would have panel data.
17. Suppose that a person's health increases with an increase in the amount of calories consumed at low levels of calories but decreases with an increase in the amount of calories consumed at high levels of calories. Which equation would best model this relationship?
- (a)  $HEALTH = \beta_1 + \beta_2 CALORIES$   
 (b)  $\ln(HEALTH) = \beta_1 + \beta_2 CALORIES$   
 (c)  $\ln(HEALTH) = \beta_1 + \beta_2 \ln(CALORIES)$   
 (d)  $HEALTH = \beta_1 + \beta_2 CALORIES + \beta_3 CALORIES^2$
- (d) We need to fit a curve that first rises and then falls. We would need a polynomial to do this.
18. Increasing the sample size used for a multivariate regression will tend to:
- (a) Make the standard errors for the individual coefficients larger.  
 (b) Make the confidence intervals for the individual coefficients wider.  
 (c) Increase the size of the standard errors relative to the magnitude of the coefficients.  
 (d) Decrease the standard errors of the individuals coefficients.
- (d) A larger sample size will tend to give us more precise estimates for the coefficients.
19. Which of the following are random variables:
- (a) The estimated coefficients in a multivariate regression.  
 (b) The sample mean.  
 (c) The standard error of a coefficient in a bivariate regression.  
 (d) All of the above.
- (d) All of these will have different values depending on the sample we happen to draw from the population.
20. The \_\_\_\_\_ the difference between the  $R^2$  for the unrestricted model and the  $R^2$  for the restricted model, the \_\_\_\_\_ we are to reject the null hypothesis that the coefficients for a subset of regressors are all equal to zero.
- (a) Larger, less likely.  
 (b) Smaller, more likely.  
 (c) Larger, more likely.  
 (d) The  $R^2$  values are irrelevant.
- (c) If there is a large difference between the  $R^2$  for the restricted model and for the unrestricted model, then the subset of regressors is improving the fit of our model making it likely that at least one of the coefficients is different than zero.

21. If  $\ln y$  is regressed on  $\ln x$ , the coefficient on  $\ln x$  tells us:
- (a) The predicted change in  $y$  with a one unit change in  $x$ .
  - (b) The predicted percentage change in  $y$  with a one percent change in  $x$ .
  - (c) The predicted change in  $y$  with a one percent change in  $x$ .
  - (d) The predicted percentage change in  $y$  with a one unit change in  $x$ .
- (b) Recall that differences in logs can be interpreted as percent changes. So the change in  $\ln y$  with a one unit change in  $\ln x$  (which is what the slope coefficient is telling us) can be thought of as the percent change in  $y$  with a one percent change in  $x$ .
22. Suppose that we regress height on a dummy variable equal to one if a person is female and equal to zero if a person is male. The coefficient for the dummy variable gives us an estimate of:
- (a) The average height of females.
  - (b) The average height of males.
  - (c) The difference between the average height of females and the average height of males.
  - (d) The difference in the variance of female heights and the variance of male heights.
- (c) The average height of females will be the intercept plus the coefficient on the dummy variable. The average height for males will just be equal to the intercept (since the dummy variable equals 0 for men). The difference between these is just the slope coefficient.
23. Which of the following will not decrease the standard error of a coefficient in a bivariate regression of  $y$  on  $x$ ?
- (a) Greater variation in the observed  $x$  values.
  - (b) A larger number of observations.
  - (c) A smaller error sum of squares.
  - (d) Larger residuals.
- (d) Larger residuals would indicate a larger standard error (larger residuals imply a larger  $s_e^2$ ).
24. Which of the following would be the best type of graph to display changes in the unemployment rate over time?
- (a) A line graph.
  - (b) A bar chart.
  - (c) A histogram.
  - (d) A pie chart.
- (a) A line graph with the unemployment rate on the vertical axis and time on the horizontal axis would give a clear picture of changes in unemployment over time.
25. Which of the following scenarios violates our assumptions for multivariate hypothesis testing?
- (a) Two of the regressors have a correlation of .80.
  - (b) Two of the regressors have a correlation of 0.

- (c) The correlation between one of the regressors and the error term is .55.
- (d) (a) and (c).

(c) Our assumptions are violated if any of the regressors are correlated with the error term. Regressors can be correlated with each other (and often are) as long as they aren't perfectly correlated.



## SECTION II: SHORT ANSWER (40 points)

1. The following page contains the results from three different regressions using the same sample of standardized test score data for California high schools. Each observation corresponds to a single high school. The variables are the following (all percentages are measured from 0 to 100, not from 0 to 1):

- *SCORE* - average test score for the high school (higher scores mean students performed better)
- *PCT\_SD* - percentage of students who are socioeconomically disadvantaged (meaning the students come from poor households)
- *PCT\_NONHSGRAD* - percentage of students who have parents that did not graduate high school
- *PCT\_LEARNERS* - percentage of students who are learning English as a second language
- *PCT\_CRED* - percentage of teachers who are fully credentialed

- (a) Explain why the coefficient on *PCT\_SD* differs between Regression A and Regression B. Be certain that your explanation specifically addresses why the coefficient is *greater* (more positive) in Regression B than it is in Regression A. (4 points)

In regression A, we are omitting the *PCT\_NONHSGRAD* variable and this is leading to an omitted variable bias for the *PCT\_SD* variable. The percentage of students who are socioeconomically disadvantaged is likely to be positively correlated with the percentage of students with parents who did not graduate from high school. From the results of regression B, we can see that there is a negative correlation between *PCT\_NONHSGRAD* and *SCORE*. So by omitting *PCT\_NONHSGRAD*, we are getting an downward bias on the coefficient for *PCT\_SD*. The coefficient is picking up the negative association between *PCT\_SD* and test scores but also the effect of an increase in the percentage of disadvantaged students leading to an increase in the percentage of students with parents who didn't graduate high school which has a negative impact on test scores. This is why the coefficient on *PCT\_SD* is more negative in regression A than it is in regression B.

- (b) In Regression B, is the coefficient for *PCT\_NONHSGRAD* significantly different than zero at a 5% significance level? Note, even if you do not need to do any calculations, you must still explain how you arrived at your answer. (3 points)

The numbers we need are all included in the regression output. To test whether a single coefficient is significantly different than zero, we can use the reported p-value for that coefficient. If this p-value is less than our chosen significance level, we will reject the null hypothesis that the coefficient is equal to zero. In the case of *PCT\_NONHSGRAD*, the p-value is .000065. Since this value is less than .05, we would conclude that the coefficient is significantly different than zero at a 5% significance level.

- (c) For Regression B, would you reject the null hypothesis that the coefficients for *PCT\_SD* and *PCT\_NONHSGRAD* are both zero at a 5% significance level? Be certain to justify your answer even if you do not need to do any calculations. (3 points)

Once again, the numbers we need are included in the regression output. We want to test overall significance and we can do this by calculating an F-stat for the regression and the corresponding p-value. If this p-value is less than our chosen significance level, we would reject the null hypothesis that all of the coefficients are equal to zero. The reported F-stat and corresponding p-value are 417 and  $1.8 \cdot 10^{-151}$ , respectively. Since this p-value is less than .05, we reject the null hypothesis that the coefficients for both regressors are equal to zero.

- (d) Test whether the addition of *PCT\_LEARNER* and *PCT\_CRED* improved the fit of the regression at a 5% significance level. Clearly state your null and alternative hypotheses, show any calculations you may need to make, and be certain to clearly state your conclusions. (5 points)

*NOTE: There was a mistake in the values for  $FDIST()$  put on the board; decimal place was in the wrong spot for the first parameter in the  $FDIST()$  values put on the blackboard. Given this mistake, full credit will be given for answers that are set up correctly even if the numbers work out wrong because of the  $FDIST$  values.*

Our null and alternative hypotheses are the following:

$$H_o : \beta_{learner} = 0, \beta_{cred} = 0$$

$$H_a : \text{at least one of the two coefficients is different than 0}$$

To test this set of hypotheses, we need to calculate our F-stat based on the  $R^2$  values for the restricted model (regression B) and the unrestricted model (regression C):

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

$$F^* = \frac{1941 - 5}{5 - 3} \frac{.3014 - .3011}{1 - .3014}$$

$$F^* = .42$$

To get the p-value that corresponds to this F-stat, we can use Excel's  $FDIST()$  function:

$$p = FDIST(.42, 5 - 3, 1941 - 5) = .66$$

This p-value is larger than .05 so we fail to reject the null hypothesis that the coefficients on *PCT\_LEARNER* and *PCT\_CRED* are both equal to zero.

**Regression A Output: Dependent variable is SCORE**

<i>Regression Statistics</i>	
R Square	0.295305163
Adjusted R Square	0.294941731
Standard Error	102.1472355
Observations	1941

  

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	8478148.006	8478148	812.5456	1.4437E-149
Residual	1939	20231637.93	10434.06		
Total	1940	28709785.94			

  

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	775.9137107	4.831396195	160.5982	0
PCT_SD	-2.340641476	0.082112829	-28.50519	1.4E-149

**Regression B Output: Dependent variable is SCORE**

<i>Regression Statistics</i>	
R Square	0.301084321
Adjusted R Square	0.300363045
Standard Error	101.7537638
Observations	1941

  

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	8644066.395	4322033	417.4333	1.7685E-151
Residual	1938	20065719.54	10353.83		
Total	1940	28709785.94			

  

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	774.4214522	4.827200746	160.4287	0
PCT_SD	-1.947588476	0.127794267	-15.24003	1.28E-49
PCT_NONHSGRAD	-0.816419724	0.203946713	-4.003103	6.49E-05

**Regression C Output: Dependent variable is SCORE**

<i>Regression Statistics</i>	
R Square	0.301377117
Adjusted R Square	0.299933681
Standard Error	101.7849819
Observations	1941

  

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	4	8652472.518	2163118	208.7915	4.9375E-149
Residual	1936	20057313.42	10360.18		
Total	1940	28709785.94			

  

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	769.9690284	19.18397495	40.13605	7.9E-257
PCT_SD	-1.926887117	0.130714519	-14.74119	1.08E-46
PCT_NONHSGRAD	-0.717102472	0.233177276	-3.075353	0.002132
PCT_LEARNERS	-0.190856233	0.220795236	-0.864404	0.387473
PCT_CRED	0.045892944	0.189469762	0.242218	0.808637

2. Below is the regression output from running a regression of educational attainment (measured as years of secondary and post-secondary education) on a dummy variable for whether the individual was the oldest child in his or her family (*oldest*, equal to one if individual was the oldest child, zero otherwise) and a dummy variable for whether the individual was the youngest child in his or her family (*youngest*, equal to one if the individual was the youngest child, zero otherwise). The sample includes only individuals with at least two other siblings.

SUMMARY OUTPUT: Dependent variable is years of education

<i>Regression Statistics</i>				
R Square	0.018739908			
Adjusted R Square	0.008131691			
Standard Error	2.005199409			
Observations	188			

  

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	2.205128205	0.16054444	13.73531	4.71E-30
youngest	0.044871795	0.476251772	0.094219	0.925037
oldest	1.128205128	0.600702289	1.878144	0.061935

- (a) Could you run the same regression for a sample of individuals with a single sibling? If not, how would you need to modify the regression. (3 points)

If you ran the same regression using individuals with one sibling, the variables *oldest* and *youngest* would be perfectly correlated. We would need to drop one of the two dummy variables from our regression.

- (b) Explain in words how we should interpret the estimated coefficient for *youngest*. (3 points)

The coefficient on *youngest* is telling us the difference in average educational attainment for individuals who are the youngest child in their family and for individuals who are the middle child. So the estimated coefficient tells us that average educational attainment is .04 years greater for youngest children compared to middle children.

- (c) Test the hypothesis that the oldest child receives more education than a middle child at a 5% significance level. Be certain to clearly state your null and alternative hypotheses, show any necessary calculations, and state your conclusions clearly. (4 points)

Our null and alternative hypotheses are the following:

$$H_o : \beta_{oldest} \leq 0$$

$$H_a : \beta_{oldest} > 0$$

Our test statistic is:

$$t^* = \frac{b_{oldest}}{s_{b_{oldest}}} = \frac{1.13}{.60} = 1.88$$

The p-value for this t-stat is (remember that we are doing a one-sided test):

$$p = TDIST(t^*, n - k, 1) = TDIST(1.88, 188 - 3, 1) = .03$$

This p-value is less than .05 so we would reject the null hypothesis that the coefficient is less than or equal to zero in favor of the alternative hypothesis that the coefficient is greater than zero.

- (d) Suppose that a researcher claims, “Birth order accounts for the majority of the observed variation in educational attainment.” Based on these regression results, would you agree or disagree with the statement? Be certain to cite any relevant regression output that supports your answer. (3 points)

There is no evidence to support this claim. The  $R^2$  for the regression is .02, implying that variation in birth order is only explaining about two percent of the variation in educational attainment. Clearly the majority of the variation in educational attainment is not explained by variation in birth order.

3. For each scenario below, explain whether the estimated coefficient for education would be biased and if so, what the direction of the bias would be. Be certain to justify your answers.
- (a) You run a regression of income on education. You have mistakenly omitted ability from the regression which is positively correlated with both income and education. (4 points)
- Since ability is correlated with both income and education, we will get a biased coefficient for education. The correlation between education and ability has the same sign as the correlation between income and ability, leading to an upward bias on the education coefficient (an increase in education has a direct, positive impact on predicted income and an indirect, positive impact through the associated increase in ability).
- (b) You run a regression of a measure of job stress on education. The coefficient on education turns out to be negative and significant but because of survey difficulties, the education variable has problems of random measurement error. (4 points)
- There will be a bias because we have measurement error in our independent variable. Random measurement error will bias our coefficient toward zero. In this case, the coefficient is negative so this will mean an upward bias.
- (c) You run a regression of annual income on education but it turns out that your sample consists of many older people who have either retired or cut back to working part time. (4 points)
- For older people who are voluntarily cutting back on the amount they work, a change in education will be associated with a smaller change in income than it would for full time workers. Since our sample is overrepresentative of older people, our estimated coefficient on education will be lower than the true value of the coefficient for the population (a downward bias).