Ecn 102 - Analysis of Economic Data
University of California - Davis March 19, 2010
Instructor: John Parman

Final Exam - Solutions

You have until 5:30pm 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

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

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

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

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

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

$\mu = E(X)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$F = \frac{ESS_r - ESS_u}{ESS_u}\frac{n-k}{k-g}$

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

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

SECTION I: MULTIPLE CHOICE (60 points)

1. Suppose we regress SAT score on parent's education and parent's income. If we run the regression again but also include the student's GPA as an additional regressor:
   (a) The R^2 for the regression will either stay the same or increase.
   (b) The adjusted R^2 for the regression will either stay the same or increase.
   (c) Both (a) and (b) are true.
   (d) Neither (a) nor (b) is true.

   (a) When adding an additional regressor, our fit should be at least as good as before, so the R^2 for the regression should either stay the same or increase. Adjusted R^2 may decrease if the additional regressor had little to no explanatory power.

2. Suppose we have a sample of the heights of Davis students and want to use the sample mean to get a confidence interval for the mean height in the population. Which of the following would increase the width of this confidence interval?
   (a) Switching from a 95% confidence interval to a 90% confidence interval.
   (b) Increasing the sample size used to calculate the sample mean.
   (c) Switching from a 95% confidence interval to a 99% confidence interval.
   (d) All of the above.

   (c) The smaller we make α, the wider our confidence interval will get. A larger sample size would make the confidence interval narrower.

3. Suppose we can reject the null hypothesis that $\beta_2 \geq 0$ at a 5% significance level where β_2 is the slope coefficient from a bivariate regression. Which of the following is definitely true?
   (a) Our test statistic was negative.
   (b) We can reject the null hypothesis that $\beta_2 = 0$ at a 5% significance level.
   (c) We can reject the null hypothesis that $\beta_2 \geq 0$ at a 2.5% significance level.
   (d) We can reject the null hypothesis that $\beta_2 < 0$ at a 5% significance level.

   (a) The critical value for a lower one-tailed hypothesis test will be negative and we will reject the null when the test statistic is more negative than the critical value.

4. Suppose we regress y on x_2. Which of the following would lead to a biased coefficient for x_2?
   (a) There is a variable x_3 that is correlated with y but not with x_2
   (b) There is a variable x_3 that is correlated with x_2 but not with y.
   (c) y is measured with some random error.
   (d) x is measured with some random error.

   (d) An omitted variable will bias the coefficient on x_2 only if it is correlated with both x_2 and with y. Measurement error in x_2 will bias the coefficient on x_2 since it will lead to errors that are negatively correlated with x_2. Measurement error in y will decrease the precision of the estimated slope coefficient but will not bias the coefficient.

5. When testing the significance of a subset of regressors, the R^2 of the unrestricted model will always be:

(a) Greater than or equal to the R^2 of the restricted model.
(b) Less than or equal to the R^2 of the restricted model.
(c) Equal to the R^2 of the restricted model.
(d) It could be greater than, less than or equal to the R^2 of the restricted model.

(a) The unrestricted model contains all of the regressors in the restricted model plus additional regressors. The unrestricted model can achieve the same fit as the restricted model by simply having the coefficients on the additional regressors set to zero. More likely is that these coefficients will be nonzero and the fit will improve.

6. Suppose that the true population model of the relationship between weight (W) and hours of exercise per day (H) is:

$$W = 200 - 10H + \varepsilon$$

where ε meets all of our assumptions. If hours of exercise is measured with some random, mean zero error, which of the following statements about the estimated slope coefficient $\tilde{b_2}$ is true?

(a) $E(\tilde{b_2}) = -10$.
(b) $E(\tilde{b_2}) = 10$.
(c) $E(\tilde{b_2}) > -10$.
(d) $E(\tilde{b_2}) < -10$.

(c) The slope coefficient will be biased toward zero due to the measurement error. In this case, that means that the expected value of $\tilde{b_2}$ will be between -10 and 0.

7. Doubling the value of the largest observation in a sample of incomes will:

(a) Increase the median.
(b) Increase the mode.
(c) Increase the mean.
(d) All of the above.

(c) Doubling the largest observation will increase the average value of the variable but will not change the position or value of the 50th percentile of the distribution of values.

8. Suppose that the risk of catching the flu is high for young children and the elderly but low for teenagers and younger adults. Which of the following equations would be the best choice for modeling the relationship between flu risk (R) and age (A)?

(a) $R = \beta_1 + \beta_2 A + \varepsilon$.
(b) $ln(R) = \beta_1 + \beta_2 ln(A) + \varepsilon$.
(c) $R = \beta_1 + \beta_2 A + \beta_2 A^2 + \varepsilon$.
(d) $R = \beta_1 + \beta_2 ln(A) + \varepsilon$.

(c) Based on the description, flu risk is first decreasing with age and then increasing with age. We need a polynomial to fit this type of parabolic curve.

9. Suppose that snow depth measured in feet is included as a regressor in a multivariate regression and the magnitude of the estimated coefficient for snowdepth is 5. If we rerun the

regression using snow depth measured in inches, the new estimated coefficient on snowfall will be:

(a) Larger than 5.
(b) Smaller than 5.
(c) Still equal to 5.
(d) Not enough information.

(b) The coefficient is giving us the change in y with a change in snowdepth of one foot. The change in y with a change in snowdepth of one inch will be $\frac{1}{12}$ of this value.

10. When regressing annual work hours on income, a researcher finds that the variance of the residuals increases as work hours increases. This will affect:

(a) The expected value of the slope coefficient for income.
(b) The magnitude of the standard error for the slope coefficient for income.
(c) Both (a) and (b).
(d) Neither (a) nor (b).

(b) This is a case of heteroskedasticity. Heteroskedasticity will change the standard errors of our estimates but will not bias the coefficients.

11. The histogram for hours of study per week based on a sample of 400 Davis students is symmetric and centered at 15 hours. Which of the following statements is true?

(a) The sample median is 15 hours.
(b) The sample mean is 15 hours.
(c) The skewness for the sample is zero.
(d) All of the above.

(d) Because the distribution is symmetric, 50 percent of the observations will be to the right of 15 hours and 50 percent will be to the left of 15 hours, making 15 hours the median. The symmetry will also lead to the mean being equal to the 15 hours (for every observation that is larger than 15, there is a corresponding observation that is smaller than 15 by the same amount). For a symmetric distribution, skewness is zero.

12. When running a bivariate regression, which of the following is not possible?

(a) The error sum of squares is larger than the total sum of squares.
(b) The error sum of squares is equal to the total sum of squares.
(c) The error sum of squares is zero.
(d) The error sum of squares is positive.

(a) The largest the error sum of squares can ever be is the magnitude of the total sum of squares. If it were larger you could achieve a better fit by simply setting all of your slope coefficients to zero.

13. Which of the following would definitely not lead to the error term being correlated with a regressor x?

(a) Random measurement error in x.
(b) An omitted variable correlated with x.
(c) Choosing an incorrect functional form for the regression equation.
(d) Random measurement error in y.

(d) If the measurement error in y is truly random, then it will be independent of the value of x. So adding this measurement error into the error term will leave the error term uncorrelated with x.

14. Adding an irrelevant variable to a regression will:

(a) Have no effect on the regression results.
(b) Tend to bias the coefficients for the other regressors.
(c) Lower the R^2.
(d) None of the above.

(d) Including an irrelevant variable may increase our standard errors. It will also lower the adjusted R^2 for the regression.

15. Suppose we run a regression with GPA as the dependent variable and SAT score as the independent variable. Which of the following statements is definitely true?

(a) The sign of the estimated slope coefficient will be the same as the sign of the correlation between GPA and SAT score.
(b) The sign of the estimated slope coefficient could be different than the sign of the correlation between GPA and SAT score if there are omitted variables.
(c) The magnitude of the slope coefficient will be equal to the magnitude of the correlation between GPA and SAT score.
(d) The slope coefficient will be statistically significant.

(a) The slope coefficient is a function of the correlation between GPA and SAT score. The signs will be the same. Even if the sign of the true relationship is the opposite of the estimated coefficient due to omitted variable bias, the sign of the correlation will match up with the sign of the estimated coefficient (the correlation does not control for the omitted variable either).

16. Which of the following is not a measure of central tendency?

(a) The mean.
(b) The mode.
(c) The sample range.
(d) The median.

(c) The sample range is a measure of dispersion. If the entire sample distribution was shifted, the center of the distribution would certainly shift but the sample range would stay exactly the same.

17. Suppose we use an F test after running a multivariate regression to test the null hypothesis that $\beta_3 = \beta_4 = 0$ and get an F statistic that is larger than the critical value for a 5% significance level. We would conclude that:

(a) $\beta_3 \neq \beta_4$.

(b) $\beta_3 > 0$ or $\beta_4 > 0$.
(c) $\beta_3 \neq 0$ and $\beta_4 \neq 0$.
(d) None of the above.

(d) We would reject the null hypothesis that $\beta_3 = \beta_4 = 0$ in favor of the alternative hypothesis that at least one of the coefficients is different from zero. We cannot say anything about whether both are different than zero or whether they are different from each other.

18. Which of the following would make you more likely to reject the hypothesis that an individual slope coefficient is equal to zero?

(a) A larger standard error for that slope coefficient.
(b) A smaller t statistic for that slope coefficient.
(c) A smaller F statistic for the regression.
(d) A larger value for the ratio of the coefficient to its standard error.

(d) When testing whether a slope coefficient is different than zero, the test statistic is simply the ratio of the coefficient to the standard error. We are more likely to reject the null hypothesis that the coefficient is equal to zero when this test statistic is larger in magnitude.

19. Suppose that we include a dummy variable for male and a dummy variable for female in a regression. This will create a:

(a) Omitted variable bias problem.
(b) Heteroskedasticity problem.
(c) Multicollinearity problem.
(d) Homoskedasticity problem.

(c) The dummy variables will be perfectly collinear (the value of one always tells us exactly what the value of the other one is).

20. The distribution of the sample mean will:

(a) Be centered at zero.
(b) Have a smaller variance for smaller sample sizes.
(c) Centered at the population mean.
(d) (b) and (c).

(c) The sample mean is normally distributed with a mean equal to the population mean and a variance that decreases as sample size increases.

21. Suppose the R^2 for a bivariate regression is equal to 1. This tells us that:

(a) The correlation between the dependent and independent variables is equal to 1.
(b) The slope coefficient is equal to 1 or -1.
(c) The error sum of squares is equal to the total sum of squares.
(d) The dependent and independent variables are perfectly correlated.

(d) An R^2 of 1 tells us that the error sum of squares is equal to zero and the variables are perfectly correlated. It does not tell us whether the correlation is equal to 1 or -1.

22. Suppose we ran a regression of Y on X 1000 times using 1000 different samples and made a histogram of the resulting slope coefficient values. Which of the following is true about the distribution shown on the histogram?
    (a) It would be centered at zero.
    (b) It would look like a normal distribution.
    (c) All of the observations would be located at the true value of the slope coefficient.
    (d) It would be right skewed.

    (b) The estimated slope coefficient is simply a random variable. It will be distributed normally with a mean equal to the true population value of the slope coefficient.

23. Which of the following depends on the units variables are measured in?
    (a) Correlation.
    (b) Coefficient of variation.
    (c) Estimated slope coefficient.
    (d) t statistic.

    (c) The estimated slope coefficient is in the units of y divided by the units of x. Changing the units of either y or x will rescale the slope coefficient.

24. Suppose the size of your social network grows exponentially over time. Which of the following equations would be the most appropriate for modeling social network size (S) as a function of time (T):
    (a) $S = \beta_1 + \beta_2 T + \varepsilon$.
    (b) $ln(S) = \beta_1 + \beta_2 ln(T) + \varepsilon$.
    (c) $S = \beta_1 + \beta_2 ln(T) + \varepsilon$.
    (d) $ln(S) = \beta_1 + \beta_2 T + \varepsilon$.

    (d) If S grows exponentially, then S increases by a constant percentage for every one unit change in time. This can be modeled with a log-linear equation.

25. Suppose that we want to test whether eye color influences the likelihood of being hired for a job. Our dataset includes five different values for the eye color variable. If we want to regress the probability of being hired on eye color, we will:
    (a) Convert each eye color to a number and include this new variable in for eye color number in the regression.
    (b) Create dummy variables for each eye color and include all of the dummy variables as regressors.
    (c) Create dummy variables for each eye color and include four of the dummy variables as regressors.
    (d) Create dummy variables for each eye color and include three of the dummy variables as regressors.

    (c) We always include one fewer dummy variable than the total number of categories. If we didn't do this, we would run into the dummy variable trap and have a perfect collinearity problem (one of the dummy variables could be rewritten in terms of the other dummy variables).

SECTION II: SHORT ANSWER (40 points)

1. (14 points) Suppose that the number of traffic accidents (N) is a function of the number of cars on the road (C) and the average speed of cars on the road (S). The true population relationship between accidents, cars and average speed is given by:

$$N = 1000 + 50C + 25S + \varepsilon \tag{1}$$

where ε is a random error that meets all of our standard assumptions. The number of cars on the road is negatively correlated with average speed due to the increased congestion associated with additional cars. The true population relationship between the number of cars and average speed is given by:

$$S = 80 - 4C + \nu \tag{2}$$

where ν is a random error that meets all of our standard assumptions.

(a) If you ran a regression with N as the dependent variable and C and S as the independent variables, what would the expected value of the estimated slope coefficient for C be? Assume that you include a constant term in your regression.

Given that ε meets all of our standard assumptions, the estimated coefficient will be unbiased. So its expected value will be equal to the true population value of 50.

(b) If you ran a regression with N as the dependent variable and C as the only independent variable, what would the expected value of the estimated slope coefficient for C be? Assume that you include a constant term in your regression.

By omitting S from the regression equation, S enters the error term making the error term correlated with C and creating an omitted variable bias. The expected value of the estimated coefficient for C will be equal to the true value plus a bias term that captures the indirect effect of S on N:

$$E(\tilde{b_c}) = \beta_c + \beta_s \cdot \gamma_c$$

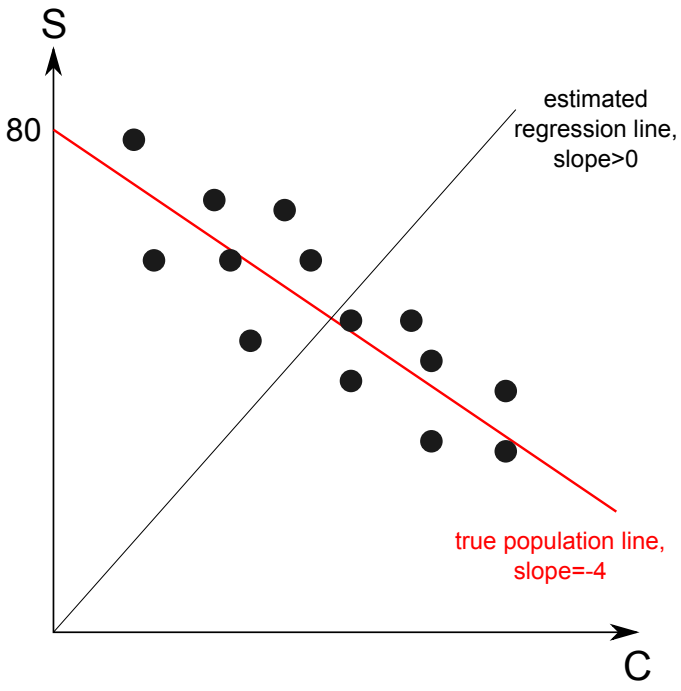
$$E(\tilde{b_c}) = 50 + 25 \cdot (-4)$$

$$E(\tilde{b_c}) = -50$$

(c) Suppose that you ran a regression with average speed as the dependent variable and number of cars on the road as the independent variable but you forced the intercept to be zero (in other words, you do not include a constant term). Will the the expected value of the estimated slope coefficient be greater than, equal to or less than the true population value of the slope coefficient? Include a written explanation and a scatter plot showing speed as a function of number of cars to illustrate your answer. Assume that we always observe positive numbers of cars and positive average speeds.

We know that all of our data points will have positive values for number of cars and average speed, so they will all lie above and to the right of the origin on a graph with S on the vertical axis and C on the horizontal axis. We are forcing our regression line to pass the origin and through this scatter of data points above and to the right of the origin. This means that we will get a positive

slope for the regression line. Given that the true value of the slope coefficient is negative, the estimated slope will certainly be greater than the true value. This situation is depicted on the graph below.



SUMMARY OUTPUT: height as dependent variable

Regression Statistics	
Multiple R	0.747320875
R Square	0.558488491
Adjusted R Square	0.533259262
Standard Error	0.357993439
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	11.34802735	2.837007	22.13657	7.68637E-12
Residual	70	8.971151182	0.128159		
Total	74	20.31917853			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	68.41515323	0.082797499	826.2949	2.3E-141
northeast	-0.731809115	0.102140602	-7.16472	6.25E-10
south	-0.260178112	0.137428521	-1.89319	0.062467
west	0.256275028	0.141697586	1.808605	0.074807
typhoiddeaths	0.02171783	0.009977496	2.176682	0.03288

Omitted region dummy variable is midwest.

SUMMARY OUTPUT: height as dependent variable

Regression Statistics	
Multiple R	0.260580275
R Square	0.06790208
Adjusted R Square	0.055133615
Standard Error	0.509357157
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.379714485	1.379714	5.317952	0.023948764
Residual	73	18.93946405	0.259445		
Total	74	20.31917853			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	68.09642269	0.07802453	872.7566	2E-148
typhoiddeaths	0.026346173	0.011424714	2.306068	0.023949

2. (14 points) For this problem, use the regression output shown on the previous page. Both regressions use the same data set. The dataset is a sample of 75 cities. *height* is a variable giving the average height in inches of adult males from the city. *tyhpoiddeaths* is a variable giving the number of typhoid deaths per 1,000 people in the city. The variables *northeast*, *south* and *west* are all dummy variables that are equal to one if the city is in that region and zero otherwise. All cities are located either in the Northeast, the South, the West or the Midwest.

   (a) Based on the regression results, what is the difference in the average male height between a city in the South and a city in the Northeast.

$$E(height|south = 1) = b_1 + b_2 northeast + b_3 south + b_4 west + b_5 typhoiddeaths$$

$$E(height|south = 1) = b_1 + b_2 \cdot 0 + b_3 \cdot 1 + b_4 \cdot 0 + b_5 typhoiddeaths$$

$$E(height|south = 1) = b_1 + b_3 + b_5 typhoiddeaths$$

$$E(height|northeast = 1) = b_1 + b_2 northeast + b_3 south + b_4 west + b_5 typhoiddeaths$$

$$E(height|northeast = 1) = b_1 + b_2 \cdot 1 + b_3 \cdot 0 + b_4 \cdot 0 + b_5 typhoiddeaths$$

$$E(height|northeast = 1) = b_1 + b_2 + b_5 typhoiddeaths$$

$$E(height|south = 1) - E(height|northeast = 1) = b_1 + b_3 + b_5 typhoiddeaths - b_1 - b_2 - b_5 tyhpoiddeaths$$

$$E(height|south = 1) - E(height|northeast = 1) = b_3 - b_2$$

$$E(height|south = 1) - E(height|northeast = 1) = (-.26) - (-.73)$$

$$E(height|south = 1) - E(height|northeast = 1) = .47$$

   So the average height in a southern city is .47 inches greater than the average height in a northeastern city.

   (b) What is the average male height for a city in the West with no typhoid deaths?

$$E(height|west = 1, typhoid = 0) = b_1 + b_2 \cdot 0 + b_3 \cdot 0 + b_4 \cdot 1 + b_5 \cdot 0$$

$$E(height|west = 1, typhoid = 0) = b_1 + b_4$$

$$E(height|west = 1, typhoid = 0) = 68.42 + .26 = 68.68$$

   (c) Based on the first set of regression results, can you reject the null hypothesis that the coefficient for typhoid deaths is less than or equal to zero at a 5% significance level? Be certain to justify your answer.

   Notice that the p-value for the typhoid deaths coefficient is .033. This value corresponds to a two-tailed test and means that would reject the null hypothesis that the coefficient is equal to 0 at a 5% significance level ($.05 > .033$) and that our t-statistic is to the right of $t_{.025,n-k}$ (since our coefficient is positive). For an upper one-sided test, we would reject the null if the t-statistic is to the right of $t_{.05,n-k}$. Notice that $t^* > t_{.025,n-k} > t_{.05,n-k}$, so we will reject the null hypothesis that the coefficient is less than or equal to zero at a 5% significance level.

(d) Calculate the test statistic you would use to test the following set of hypotheses:

$$H_o\text{: } \beta_{ne} = \beta_s = \beta_w = 0$$

$$H_a\text{: at least one of } \beta_{ne},\ \beta_s \text{ and } \beta_w \text{ is different than zero}$$

We are testing the significance of a subset of regressors. This requires calculating an F statistic:

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

$$F^* = \frac{.56 - .07}{1 - .56}\frac{75-5}{5-2}$$

$$F^* = 25.98$$

(e) Explain how you would use your test statistic from part (d) to decide whether or not to reject the null hypothesis. Be as specific as possible.

We could take either the p-value approach or the critical value approach. For the p-value approach, we would use FDIST() in Excel to calculate the p-value associate with our F statistic and our degrees of freedom. We would have Excel calculate FDIST(25.98, 3, 70). The result would be our p-value. We would reject the null hypothesis if this p-value is less than our chosen significance level α.

For the critical value approach, we would need to calculate the critical value corresponding to our chosen significance level α. We could do this in Excel by calculating FINV(α, 3, 70). If the resulting critical value is less than our F statistic, we would reject the null hypothesis.

3. (12 points) Suppose that we are interested in the relationship between hours of weekly exercise and resting heart rate. The more a person exercises on average, the lower his or her resting heart rate is. For individuals who don't exercise at all, males have a lower resting heart rate on average than females. The decrease in resting heart rate from an additional hour of exercise per week is bigger for males than females.

   (a) Write down the regression model you would use to estimate the relationship between resting heart rate, gender and weekly exercise. Resting heart rate should be your dependent variable. Provide clear definitions of all variables you include in your model.

   Our regression model will have to include resting heart rate, weekly exercise and a variable capturing gender. Since gender is a categorical variable, we will need to use a dummy variable. We have two values for gender (male and female) so we will need one dummy variable. Let's make our dummy variable for male, so it equals one if gender is male and equals zero if gender is female. This leaves gives us the following set of variables:

   - R - resting heart rate
   - E - amount of weekly exercise
   - M - dummy variable equal to one if male, zero if female

   Our regression equation will have R as the dependent variable. E and M will be independent variables. We also need to include an interaction term between E and M since the marginal effect of E on R depends on the value of M. This gives us the following regression model:

$$R = \beta_1 + \beta_2 E + \beta_3 M + \beta_4 E \cdot M + \varepsilon$$

   (b) Based on the information given above, what are the expected signs for each of your coefficients in the regression model you specified in part (a)?

   Notice that β_2 is the marginal effect of exercise on resting heart rate for females (since the interaction term will be zero). According to the problem, more exercise lowers the resting heart rate, so β_2 should be negative. The marginal effect of exercise on heart rate is larger (more negative) for males than females. This marginal effect is captured by $\beta_2 + \beta_4$, so β_4 should be negative. For individuals exercising the same amount, the difference between the average male heart rate and the average female heart rate will be β_3. We are told that males have a lower heart rate than females that exercise the same amount. So β_3 should be negative. Finally, heart rate has to be positive overall, so the constant term β_1 should be positive (if it were negative, we would predict that a female who does not exercise has a negative heart rate). To summarize:

$$\beta_1 > 0$$

$$\beta_2 < 0$$

$$\beta_3 < 0$$

$$\beta_4 < 0$$

Note that if you used a dummy variable equal to one for females and zero for males, your signs for β_3 and β_4 would be reversed. The signs for β_1 and β_2 would stay the same.

(c) Suppose people tend to make random mistakes when measuring their heart rate. What effects will this have on the estimation results when you run the regression model specified in part (a)?

Measurement error in the dependent variable will not bias our coefficients. So the expected values of the coefficients will stay the same. However, the measurement error does add variance to the error term which will lead to less precise estimates of the coefficients (larger standard errors).