**Final Exam - SPEA V506 (Summer 2018)**

Turn in your answers as a word document (you may include pictures or screenshots of hand-written work if that is easier. Show your work for all of questions below and make your final answer clear (circle, bold, etc..). For Part II, please include both your R code and the relevant output.

You may use any of the Canvas materials for this exam but are not allowed to consult other web sites or materials, or to discuss the exam with any other students. If you have questions about the exam, please email me or the TAs. Please complete the following agreement form and turn this in as part of answer document.

**Take Home Final Exam Agreement Form**

**Date:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Directions:**  **Read and initial next to each statement, type your name below as indicated**. Your initial located next to each item below signifies your understanding and compliance with the instructions for this take home examination.

**\_\_\_\_\_\_ I understand that I may not copy or distribute this exam to anyone.**

**\_\_\_\_\_\_ I understand that I must return the examination by the stated deadline July 1, 2018 11:59 pm EST and that late exams will not be accepted.**

**\_\_\_\_\_\_ I understand this is a test and I cannot copy answers from other students. I understand that I may not discuss this exam with other students (current or previous). Furthermore, I cannot seek assistance, of any kind, from anybody other than the TAs or Instructor.**

**\_\_\_\_\_\_\_ I understand that I CAN consult resources including the textbook for the class, my own hand-written notes, the course presentations, or any of the resources on the Canvas course site. I understand that I CANNOT consult the web beyond Canvas.**

**\_\_\_\_\_\_ I understand that violation of this agreement will result in an F on this exam and it cannot be replaced, it will be averaged in (as a 0%) with my other class scores. Violations of this agreement will be handled according to IU’s policies on academic misconduct.**

**(Type your name here-on the line): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Part I – Short Answer (40 pts)**

Answer each of the questions below as completely as possible, showing work wherever appropriate, i.e. whenever computing values).

1. In OLS regression, what does the method of *least squares* refer to? (2 pts)

2. What does the *F* statistic in an OLS model (just one model) indicate? What is the null hypothesis regarding this statistic? (2 pts)

3. What does the *Pearson correlation coefficient* tell us about the relationship between two variables? Suppose that this coefficient for two variables is 0.8. What would that indicate about the relationship between the two variables? What if the coefficient was -0.8. What if it was 0.1? (3 pts)

4. Suppose that in a bivariate regression model, the effect of an independent variable on the dependent variable is positive and statistically significant, but when a second variable is added to the model, the effect of the original independent variable becomes negative and statistically significant. Explain why this would occur. What implications does this have for model specification in terms of omitting independent variables? (5 pts)

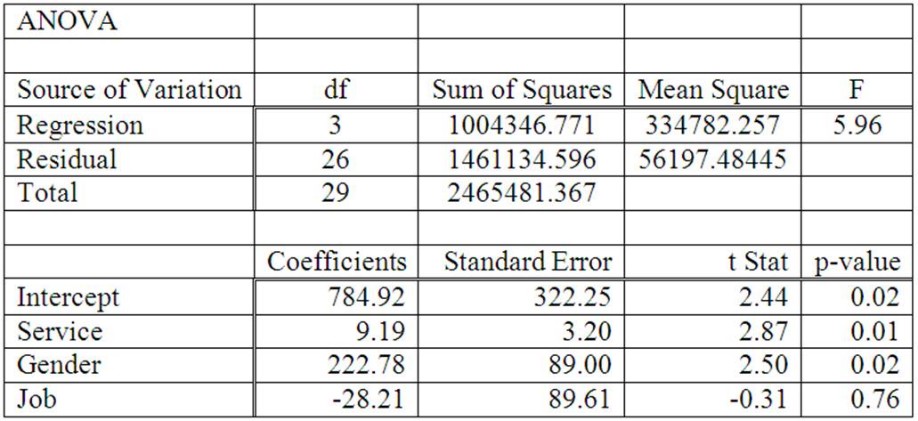
5. Given the least squares regression equation, Ŷ = 1,202 + 1,133X, when X = 3, what does Ŷ equal? (2 pts)

6. It has been hypothesized that overall academic success for college freshmen as measured by grade point average (GPA) is a function of IQ scores = X1, and hours spent studying each week = X2. Suppose the regression equation is:

Ŷ = -5.7 + 0.055X1 +0.107X2

What is the predicted GPA for a student with an IQ of 108, and 32 hours spent studying per week? (2 pts)

7. A manager at a local bank analyzed the relationship between monthly salary and three independent variables: length of service (measured in months), gender (0 = female, 1 = male), and job type (0 = clerical, 1 = technical). The table below summarizes the regression results. Interpret the coefficients and t-statistics for the Job and Gender variables. (3 ps)



8. A multiple regression model includes the interaction term (*X1*)(*X2*), and the resulting coefficient is positive and statistically significant. What does this term imply about the effect of X1 on Y? (2 pts)

9. The sales manager of a large automotive parts distributor wants to estimate as early as April the total annual sales. According to the manager of the distribution warehouse, several factors are related to annual sales (measured in millions of dollars) (*sales*), including the number of retail outlets in the region stocking the company’s parts (*outlets*), the number of automobiles in the region registered as of April 1 (measured in millions) (*cars*), the total personal income for the first quarter of the year (measured in billions of dollars) (*income*), the average age of automobiles in years (*age*), and the number of supervisors at the distribution warehouse (*bosses*). The data for all these variables were gathered for a recent year.

Consider the following correlation matrix.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *sales* | *outlets* | *cars* | *income* | *age* |
| *outlets* | 0.899 |  |  |  |  |
| *cars* | 0.605 | 0.775 |  |  |  |
| *income* | 0.964 | 0.825 | 0.409 |  |  |
| *age* | -0.323 | -0.489 | -0.447 | -0.349 |  |
| *bosses* | 0.286 | 0.183 | 0.395 | 0.155 | 0.291 |

A. Which single variable has the strongest correlation with the dependent variable? (2 pts)

B. Is there evidence of multicollinearity? If so, between which variables? (2 pts)

10. Using the same data described above, the following multivariate regression equation was estimated:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coef | SE Coef | t statistic | p |
| Intercept | -19.672 | 5.422 | -3.63 | 0.022 |
| Outlets | -0.000629 | 0.000213 | -2.96 | 0.018 |
| Cars | 1.7399 | 0.5530 | 3.15 | 0.035 |
| Income | 0.40994 | 0.04385 | 9.35 | 0.001 |
| Age | 2.0357 | 0.65247 | 3.12 | 0.014 |
| Bosses | -0.0344 | 0.01128 | -3.05 | 0.016 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA | |  |  |  |  |
| SOURCE | DF | Sum of Squares | Mean Square | F | P |
| Regression | 5 | 1593.81 | 318.76 | 140.36 | 0.000 |
| Residual Error | 4 | 9.08 | 2.27 |  |  |
| Total | 9 | 1602.89 |  |  |  |

A. State the null hypothesis concerning the statistical significance of the overall regression, test this hypothesis, and interpret the results. (use a .05 level of significance) (2 pts)

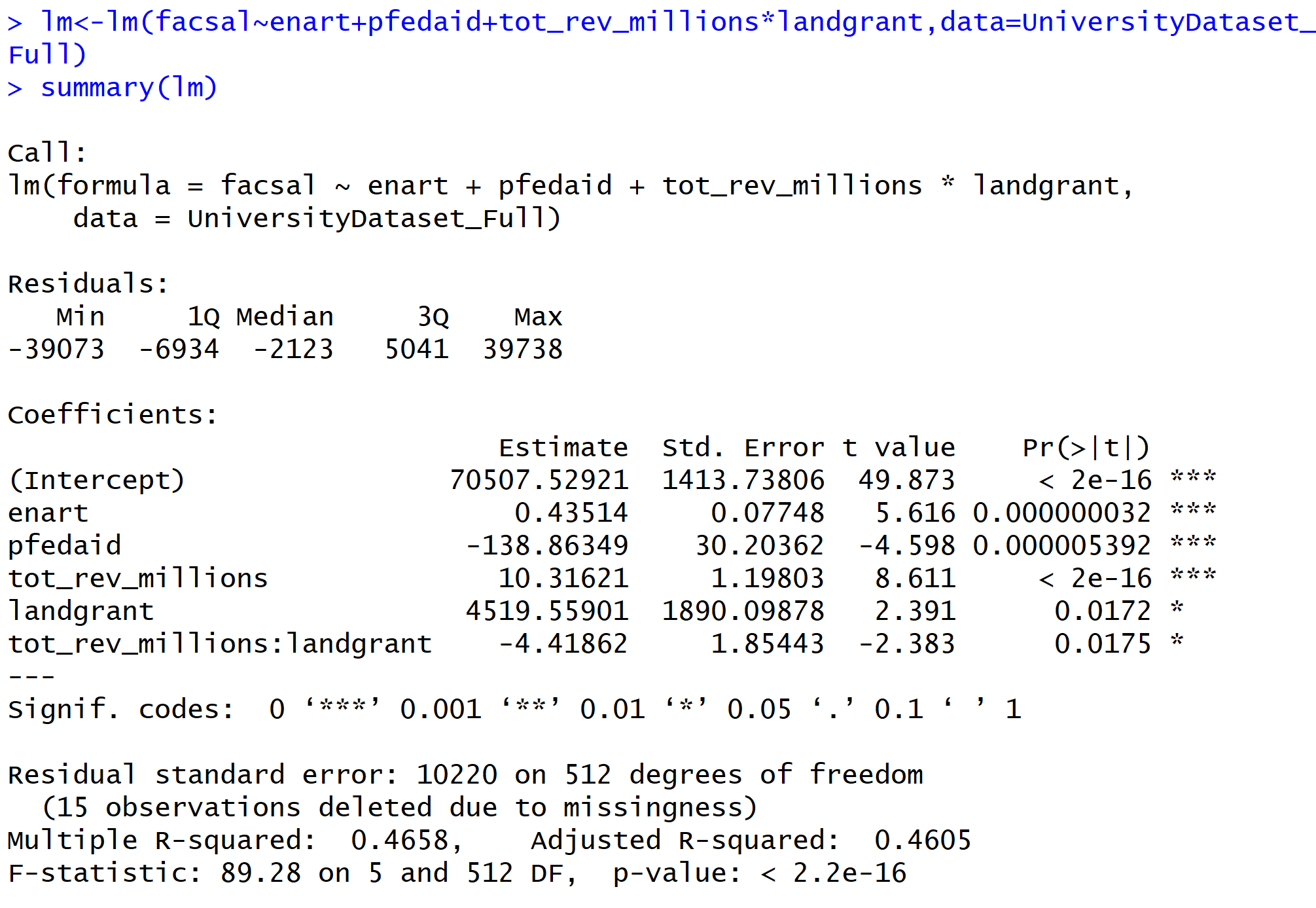
B. What is the value of the coefficient of determination, i.e., r-squared (\*note for this example, you do ***NOT*** need to compute adjusted r-squared\*)? Based on this value, what percent of the variation in the dependent variable is explained by the regression equation? (2 pts)

C. Interpret the results (both statistical significance and magnitude of effect) for each of the independent variables in the model. (Use a .05 level of significance) (4 pts)

D. What would be the projected value in annual sales if the following were true?   
*outlets =* 1739, *cars = 9.27*, *income*  = 85.4, *age*  = 3.5, and *bosses*  = 9.0   
If these values are outside the range of values used for the regression, would this be a reliable forecast? Why or why not? (2 pts)

11. Suppose you are a research assistant to a young assistant professor who is very interested in predicting faculty salaries at US public colleges and universities. To do so, you estimate a linear regression model with faculty salaries (facsal) as the dependent variable, and you include four independent variables: total enrollment (enart), % of students receiving federal aid (pfedaid), whether an institution is a Landgrant college or university (landgrant), and total revenues in millions of dollars (tot\_rev\_millions). Note that landgrant is a dichotomous (dummy) variable while the other variables are all continuous variables.

Because you suspect that there may be an interactive effect between Landgrant institutions (landgrant) and total revenues in $ millions (tot\_rev\_millions), you include an interaction term in the model for these two variables (landgrant\*tot\_rev). The the output of the linear regression model is below.

A) Interpret the base coefficients for landgrant, and tot\_rev\_millions, as well as the interaction term (tot\_rev\_millions:landgrant), both in terms of statistical significance and in terms of their impact on the dependent variable. (5 pts).

B) Based on your regression model, what is the predicted faculty salary for the following two hypothetical institutions (2 pts):

Total Enrollment = 20,000

% Receiving Federal Financial Aid = 25

Total Revenues = $30 Million

Landgrant = 1

Total Enrollment = 20,000

% Federal Receiving Financial Aid = 25

Total Revenues = $30 Million

Landgrant = 0

**Part II – R Programming (60 pts)**

Include both the output and relevant R code in your answer.

1. Download the statelife\_exp.csv file and load it into R. This dataset contains the following variables:

State: name of state

Population: population estimate as of July 1, 1975

Income: per capita income (1974)

Illiteracy: illiteracy (1970, percent of population)

Life Exp: life expectancy in years (1969–71)

Murder: murder and non-negligent manslaughter rate per 100,000 population (1976)

HS Grad: percent high-school graduates (1970)

Frost: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city

Area: land area in square miles

Estimate an OLS model where life expectancy is the dependent variable, and the following are independent variables: murder rate, population, percent high-school graduates.

A) Construct a table containing summary descriptive statistics for all of the variables in the model (both IVs and DV), and construct histograms for each variable. (6 pts)

B) Interpret the results of the model, both substantively and statistically. Be sure to discuss both the model as a whole as well (F statistic and adjusted r-squared), as well as the results for each independent variable in the model (coefficient, t-statistic, and p-value). (10 pts)

C) Estimate a new model with the same DV and set of IVs listed earlier, but add a new independent variable for the mean number of days with minimum temperature below freezing. Interpret the results of this model (same procedure as in part B). (10 pts)

D) Next, conduct a nested F-test to determine if adding this new variable improved the explanatory power of the model. (3 pts)

2) Download the violent\_dissent.csv file and load it into R. This data contains information about violent internal dissent, democracy and state repression for 1984. The variables in the dataset are as follows:

ccode: Numeric COW country code

pop: population in thousands

rgpdpch: Real GPD per capita

democracy: Behavioral measure of democracy including legislative competition/participation, competitiveness of executive recruitment and participation, along with other factors. This is a continuous measure. Higher numbers indicate greater levels of democracy.

vdissdum: Binary violent dissent variable (dummy variable) based on data regarding terrorist events and fatalities, as well as guerrilla wars and riots. 0=country is peaceful; 1=country is in state of violent conflict

repression: Violations of physical integrity rights. A continuous measure based on human rights data regarding torture, forced disappearances, extrajudicial killings and political imprisonment. Higher numbers indicate greater levels of repression.

Estimate an OLS regression model where repression is the dependent variable, and pop, rgdpch, democracy, and vdissdum are the independent variables.

A) Interpret the results of the model, both substantively and statistically. Be sure to discuss both the model as a whole (F statistic and adjusted r-squared), as well as the results for each independent variable in the model (coefficient, t-statistic, and p-value). (10 pts)

B) Some hypothesize that the effect of democracy on repression is conditional and works differently in peaceful countries than in countries in a state of violent conflict. To test this hypothesis, add a multiplicative interaction term to the model for democracy and vdissdum (democracy\*vdissdum). Interpret the model results (using the same procedure as part A). (10 pts)

C) Next, conduct a nested F-test to determine if adding the interaction term for democracy\*vdissdum improved the explanatory power of the model. (3 pts)

D) Finally, based on the model estimated in part B, construct an effects plot to show the effect of democracy on repression for peaceful and for violent countries. Be sure to include a legend and appropriately label your axes. (8 pts)