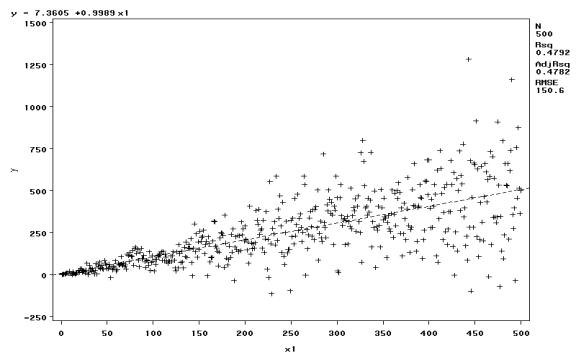
PA 541 – SAMPLE EXAM

Midterm

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\***Directions**: You are allowed to use a hand-written page of notes (double-sided) during the exam. You are not required to perform any difficult calculations and thus do not need your phone, a calculator, or your computer. Good luck.

1. A social scientist is interested in understanding the work-life balance of married women in the public sector. Specifically, the social scientist would like to explain the number of hours worked per week (outside of the home) by the number of years of formal education a woman has completed and the number of children in her family.
   1. Identify the dependent and independent variables. (2pts)
   2. Write out the regression equation for this model. (1pt)
   3. Modify the regression equation in part b above to include the squared value of education. (1pt)
   4. What is one plausible reason why we would include the squared value of education? What is it capturing and what do you hypothesize the sign of the coefficient to be? (4pts)
2. What does the following image reveal about the relationship between our independent and dependent variable and the potential violation of a regression assumption? Draw a scatterplot that would suggest that this assumption is not violated. (4pts)



1. You develop a regression model to predict a subject’s weight. In your model you use the height of the subject measured in inches and subject’s height measured in feet. Explain intuitively why ordinary least squares (OLS) cannot estimate the regression coefficients in such a regression model. (3pts)
2. A city manager was interested in predicting home values (*price*, measured in thousands of dollars) based off of the number of bedrooms in the home (*bdrms*) and the lot size (*lotsize*, the size of the land the home sits on in square feet). The following is the output from the regression model:  
     
   Pricei = β0 + β1Bedroomsi + β2LotSizei + εi  (MODEL 1)

Call:

lm(formula = price ~ bdrms + lotsize, data = house)

Residuals:

Min 1Q Median 3Q Max

-239.42 -53.18 -5.28 34.19 286.57

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 63.26224 39.61957 1.60 0.1140

bdrms 57.31285 10.88453 5.27 0.000001 \*\*\*

lotsize 0.00286 0.00090 3.18 0.0021 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.6 on 85 degrees of freedom

Multiple R-squared: 0.337, Adjusted R-squared: 0.321

F-statistic: 21.6 on 2 and 85 DF, p-value: 0.0000000263

* 1. What is the percentage of variation explained in this model? (2pt)
  2. How do we interpret the coefficient on the number of bedrooms and the lot size (please be specific and use the proper units of measurement)? (4pts)
  3. For this model, in terms of our dataset, does the intercept term make sense to interpret by itself? Why or why not? (2 pts)

* 1. What is the expected price of a 2 bedroom home that sits on a 10,000 sqft lot? (2pts)

1. An analyst working with the city manager on the preceding problem was worried that she underspecified the model. The analyst is concerned that the size of the home itself (*sqrft*), measured in square feet, should also be included in the model. The following regression model was run:  
   Pricei = β0 + β1HomeSizei + β2LotSizei + β3Bedroomsi + εi  (MODEL 2)

Call:

lm(formula = price ~ sqrft + lotsize + bdrms, data = house)

Residuals:

Min 1Q Median 3Q Max

-120.03 -38.53 -6.55 32.32 209.38

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -21.770309 29.475042 -0.74 0.4622

sqrft 0.122778 0.013237 9.28 0.00001 \*\*\*

lotsize 0.002068 0.000642 3.22 0.0018 \*\*

bdrms 13.852522 9.010145 1.54 0.1279

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.8 on 84 degrees of freedom

Multiple R-squared: 0.672, Adjusted R-squared: 0.661

F-statistic: 57.5 on 3 and 84 DF, p-value: <2e-16

* 1. Was the analyst correct in assuming that the size of the house (*sqrft*) should have been included in the regression model? (2pts)
  2. What is the increase in the percentage of variation explained in the model due to the inclusion of the variable measuring the size of the house? (2pts)
  3. Compare the estimated coefficient on the number of bedrooms based on the results from Model 1 with those from Model 2. Not only does the size of the coefficient change dramatically between these two sets of results, but so does the t-statistic. Give at least one plausible explanation why we observe such large changes in these values. (4pts)

1. Because she was concerned with some possible assumption violations (linearity, normality) the analyst wanted to run an additional model that predicts the log of the price of the home (*lprice*), based off of the logged values of house size (*lsqrft*) and lot size (*llotsize*). The output for this model is below:  
     
   ln(Pricei )= β0 + β1ln(homeSizei )+ β2ln(lotSizei ) + εi  (MODEL 3)

Call:

lm(formula = lprice ~ lsqrft + llotsize, data = house)

Residuals:

Min 1Q Median 3Q Max

-0.6533 -0.1105 -0.0065 0.1182 0.6642

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.6401 0.6019 -2.72 0.0078 \*\*

lsqrft 0.7624 0.0809 9.43 7.4e-15 \*\*\*

llotsize 0.1685 0.0385 4.38 3.4e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.185 on 85 degrees of freedom

Multiple R-squared: 0.635, Adjusted R-squared: 0.627

F-statistic: 74 on 2 and 85 DF, p-value: <2e-16

* 1. How do we interpret the coefficient on the log of house size? (2pts)
  2. How do we interpret the coefficient on the log of lot size?(2pts)
  3. Would it be possible to add the variable on the number of bedrooms to this model without it needing to be log transformed? (2pts)

1. Based on the model below, where: yi = hourly wage of a worker; Di = 1 if the worker is a male (0 for female); and Xi = years of experience. Interpret what the graph is telling us in terms of intercepts and slope terms for males and females. Please discuss each term (excluding the error) in the equation (α0, α1, β1, β2). (8pts)  
     
     
   
2. The model written below is defined as follows: yi = hourly wage of a worker; D1i = 1 if the worker is in the private sector, 0 otherwise (i.e., nonprofit or public); D2i = 1 if white, 0 otherwise; and Xi = years of experience. Please fill out the following conditional expectations. (The first one has been done for you; use this format for the rest. Please note the dummy variable interaction term.) (6pts)



* 1. Mean earnings of non-private sector, non-white worker

E(yi|xi, D1=0, D2=0) = α0 + βxi

* 1. Mean earnings of private sector, non-white worker

c. Mean earnings of non-private sector, white worker

d. Mean earnings of private sector, white worker

1. A county executive was concerned that there may be some discrimination going on regarding the pay scales based on gender and race within local government agencies. Data was collected on 525 randomly selected employees. The dependent variable is the average hourly earnings (*wage*). The independent variables are (i) years of education (*educ*), (ii) years with current employer (*tenure*), (iii) a dummy variable to indicate if the person is white or non-white (*white*, where being white =1), (iv) a dummy variable to indicate if the person is male (*Male*, where being male = 1), (v) a dummy variable to indicate if the person is married (*married*, were being married = 1), and (vi) a variable of the squared value of the years with current employer (*tenursq*). Please answer the following questions based on the output below.

Call:

lm(formula = wage ~ educ + tenure + white + Male + married +

tenursq, data = wage)

Residuals:

Min 1Q Median 3Q Max

-7.432 -1.780 -0.502 0.970 13.481

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.09711 0.70259 -4.41 1.3e-05 \*\*\*

educ 0.52367 0.04693 11.16 < 2e-16 \*\*\*

tenure 0.28122 0.04712 5.97 4.5e-09 \*\*\*

white 0.13838 0.42528 0.33 0.7450

Male 1.71591 0.26491 6.48 2.2e-10 \*\*\*

married 0.56709 0.27682 2.05 0.0410 \*

tenursq -0.00492 0.00168 -2.93 0.0035 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.94 on 518 degrees of freedom

Multiple R-squared: 0.375, Adjusted R-squared: 0.368

F-statistic: 51.8 on 6 and 518 DF, p-value: <2e-16

* 1. Is there any evidence from the model results to support the claim that there is discrimination based on race? Why or why not? (2pts)
  2. Is there any evidence to support the claim that there is discrimination based on gender? If yes, does this mean that the local government agencies are guilty of gender discrimination? Why or why not? Provide an explanation. (4pts)
  3. How do we interpret the coefficient on the dummy variable for male? (2pts)
  4. Can one tell from the model used if there is a significant difference on the return on education (i.e. the slope of education) between men and women? If yes, what is the difference? If not, what would to do to test the hypothesis that there is a significant difference between men and women regarding the return on education? (4pts)
  5. What is the interpretation of the intercept in this model? What is one reason why the intercept is negative? (2pts)
  6. What does the coefficient value and significance of tenure squared tell us? Draw a graph depicting the relationship between tenure and hourly earnings. Make the y-axis hourly earnings and the x-axis tenure. (2pts)