STAT 3A03 Applied Regression With SAS

Assignment 2 Due at **5pm** on Thursday October 19, 2017

Dropboxes for assignment submission are outside HH-105. Your assignment **MUST** be deposited in the appropriate dropbox for your lab section.

N.B. Late assignments will not be accepted

Q. 1 Suppose we have a response variable Y and two predictors X_1 and X_2 and that the observed values of X_1 and X_2 have sample correlation $r_{1,2} = 0$. Define the following notation

$$S_{1,1} = \sum_{i=1}^{n}(x_{i1} - \overline{x}_1)^2 \qquad S_{2,2} = \sum_{i=1}^{n}(x_{i2} - \overline{x}_2)^2 \qquad S_{1,2} = \sum_{i=1}^{n}(x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2)$$

$$S_{1,y} = \sum_{i=1}^{n}(x_{i1} - \overline{x}_1)(y_i - \overline{y}) \qquad S_{2,y} = \sum_{i=1}^{n}(x_{i2} - \overline{x}_2)(y_i - \overline{y})$$

a) Adjusting Y for X_1 amounts to calculating the residuals from the linear model with dependent variable Y and covariate X_1. Show that the residuals from this model can be written as

$$e_{i1} \;=\; (y_i - \overline{y}) - \frac{S_{1,y}}{S_{1,1}}(x_{i1} - \overline{x}_1)$$

b) Now consider adjusting X_2 for X_1. What is the value of the slope of the regression line? Give the residuals from the model.

c) The additional effect of X_2 after adjusting for X_1 can be found by considering the residuals from (a) as the dependent variable and those from (b) as the covariate. Show that, in this case, the estimated slope of this regression is exactly equal to the slope for the regression of Y on X_2 ignoring X_1.

d) Conversely, suppose that $|r_{1,2}| = 1$, show that the residuals from part (b) are all equal to 0 and hence the regression in part (c) cannot be solved.

Q. 2 Textbook Question 3.5. The data are in the file `Examination.txt` on the website. Use SAS to fit the models and include your code and SAS output for each of the models in your solution.

Q. 3 The following are the design matrix $\boldsymbol{X}$, response vector $\boldsymbol{Y}$ and $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ for a multiple regression model.

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 2 \\ 1 & 0 & 3 \\ 1 & 0 & 4 \\ 1 & 0 & 5 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 1 & 4 \\ 1 & 1 & 5 \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} 5 \\ 2 \\ 6 \\ 5 \\ 9 \\ 0 \\ 4 \\ 5 \\ 8 \\ 7 \end{pmatrix} \quad (\boldsymbol{X}^t\boldsymbol{X})^{-1} = \begin{pmatrix} 0.65 & -0.20 & -0.15 \\ -0.20 & 0.40 & 0.00 \\ -0.15 & 0.00 & 0.05 \end{pmatrix}.$$

a) Calculate the least squares estimates of the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

b) Calculate the fitted values from your model and hence calculate an unbiased estimate of the error variance σ^2.

c) Construct 95% confidence intervals for β_1 and β_2.

d) Conduct a test of the hypothesis $H_0 : \beta_2 = 1$ against the alternative $H_1 : \beta_2 \neq 1$.

e) Give a point estimate and 95% confidence interval for the quantity $\mu = \beta_0 + \beta_1 + 3\beta_2$.

f) Predict the response value for a new observation with $x_1 = 0$ and $x_2 = 5$ and give a 95% prediction interval.

Q. 4 A realtor took a sample of 24 house sales and tried to fit a model to the sale prices of the houses as a function of the following variables

Tax	Property taxes in thousands of dollars
Age	Age of the house in years
Bed	Number of Bedrooms
Bath	Number of Bathrooms
Space	Living space in thousands of square feet
Lot	Lot size in thousands of square feet

The response variable was the sale price of the house in thousands of dollars.

a) Here is a partial ANOVA Table from fitting this model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	a	d	112.643245	g
Error	b	e	f	
Corrected Total	c	831.50958		

Give the complete ANOVA table by filling in the missing values a, b, c, d, e, f and g.

b) State the null and alternative hypothesis that the F Value (g) is testing. Give the conclusion of the test for this fitted model.

c) Give an estimate of the error standard deviation σ.

d) What is the sample correlation coefficient between the 24 observed sale prices and those that the realtor would predict from the model?

e) Here is another ANOVA Table which uses only the two variables taxes and number of bedrooms to predict the sale price of a house.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	2	636.52647	318.26323	34.28
Error	21	194.98312	9.28491	
Corrected Total	23	831.50958		

State carefully the null and alternative hypotheses that test if these two variables alone are sufficient to predict the sale price of a house. Calculate the appropriate test statistic and give its distribution when H_0 is true. What is your conclusion from this test?