STAT 3A03 Applied Regression Analysis With SAS

Assignment 4 Due at 5pm on Thursday November 16, 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 that a we have a continuous response variable Y and a single categorical variable X with values $1, \ldots, p$. It is of interest to test if the levels of the response vary among categories of the predictor. Define the indicator variables

$$E_j = \begin{cases} 1 & \text{if } X = j \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } j = 1, ..., p$$

and consider fitting the linear model

$$Y = \gamma_1 E_1 + \gamma_2 E_2 + \cdots + \gamma_p E_p + \varepsilon$$

a) State the assumptions of this model and give an interpretation for the parameters.

b) Let $\boldsymbol{X}$ be the design matrix for this model. Give the form of $\boldsymbol{X}^t\boldsymbol{X}$ and its inverse and then use the results for multiple regression to obtain the least squares estimates for $\gamma_1, \ldots, \gamma_p$.

c) Give the variance of $\hat{\gamma}_j$ and the correlation between $\hat{\gamma}_j$ and $\hat{\gamma}_k$ when $j \neq k$.

d) State the null hypotheses that are being tested in the usual t-tests that are produced as part of the regression output and also the F test from the ANOVA for this model. Comment on whether these are appropriate tests.

Q. 2 The data in `GeneExp.txt` gives some data from a genetics experiment on mice. There were 3 strains of mice used of both genders. The response variable is the expression, which is a measure of the activity, of a certain gene.

a) Fit an appropriate model to test if there is a difference between the mean expression levels across the three strains and report your findings.

b) Test the hypothesis that there is a gender effect on gene expression in each strain and report your findings.

c) Test the hypothesis that the gender effect is the same for all three strains and report your findings.

d) What model would you choose for this data? Justify your answer.

e) For the model you chose in (d), write a brief report for the biologist giving an interpretation of the results and commenting on the standard assumptions for this model.

Q. 3 Textbook 5.7 (The data are in `Fertilizer.txt`)

Q. 4 The data in `Students.txt` contains Age, Height, Weight and Sex (0=Male, 1=Female) for students in a university statistics course.

a) Fit a model to predict Weight from Height for all students without taking gender into account. Check the usual assumptions of the linear model. Give an interpretation of the fitted model.

b) The following code will plot Height against Weight using different symbols for each gender.

```
PROC GPLOT Data=S3A3.Student;
  PLOT Weight*Height=Sex;
run;
```

Construct a plot of the studentized residuals from your model against the fitted values using different symbols for males and females and comment on whether there is any evidence of a different pattern of residuals by gender.

c) Construct a single model that allows separate intercepts and slopes of Height for males and females. Write a short report to interpret the results including comments on the usual assumptions. Give the fitted regression lines for males and females in this analysis.

d) Carry out a statistical test to compare the models in part (a) and part (c), stating your null and alternative hypotheses clearly and describing your conclusions.

e) Based on the model fitted in part (c), test the hypothesis that the lines for males and females are parallel giving the null and alternative hypotheses and the value of the test statistic. Give the distribution of the test statistic under the null hypothesis and the p-value for the test. What is your conclusion?