STAT 420
2GR, 2UG

Spring 2017

Homework 8

(Due Friday, April 7, by 4:00 p.m.)

Please submit your assignment *on paper*, following the Guidelines for Homework Write-Ups and Submissions. Please include your name (with your last name underlined), and your NetID at the top of the first page.

1. In order to compare the average GPA for the members of three activity clubs at a university, eight students were randomly chosen from each of the three clubs (Drama, Writing, Statistics), the students' GPA (y) and the average time spent studying per week (x) in hours was recorded. Consider the model

$$Y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 x + \varepsilon,$$

where

$v_1 = 1$ if a student is from Drama club, 0 otherwise,

$v_2 = 1$ if a student is from Writing club, 0 otherwise.

Let $v_3 = 1$ if a student is from Statistics club, 0 otherwise (Hint or Note: You may or may not need this notation for v_3).

```
> sum( lm( y ~ v1 + v2 + x )$residuals^2 )
[1] 8.0
> sum( lm( y ~ v1 + v2 )$residuals^2)
 [1] 14.0
> sum( lm( y ~ x + 0 )$residuals^2)
[1] 18.0
> sum( lm( y ~ x )$residuals^2)
[1] 11.0
> sum( lm( y ~ 1 )$residuals^2)
[1] 20.0
```

(a) We wish to test if the relationship between GPA and time spent studying is the same for all three cubs. Perform the appropriate test at $\alpha = 0.05$. State the null hypothesis, report the value of the test statistic, the critical value(s), and the decision.

(b) We wish to test if studying affects GPA. That is, test $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$ at $\alpha = 0.01$. Report the value of the test statistic, the critical value(s), and the decision.

(c) Suppose we suspect that the rate (i.e. the slope) of the relationship between GPA and time spent studying may be different for different clubs. Suggest an appropriate model.

(d) For your model in part (c), we wish to test if the rate (i.e. the slope) of the relationship between GPA and time spent studying is the same for the three clubs. Specify the null hypothesis H_0 in the notations of your part (c) model.

2. The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. The data frame has 97 rows and 9 columns:

`lcavol`	log(cancer volume)
`lweight`	log(prostate weight)
`age`	age
`lbph`	log(benign prostatic hyperplasia amount)
`svi`	seminal vesicle invasion
`lcp`	log(capsular penetration)
`gleason`	Gleason score
`pgg45`	percentage Gleason scores 4 or 5
`lpsa`	log(prostate specific antigen)

(Source: Andrews D.F. and Herzberg A.M. (1985) *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York.)

```
> library(faraway)
> data(prostate)
> prostate[1:5,]        ### so we can see what the data set looks like
      lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
>
> attach(prostate)
```

(a) Fit a model with `lpsa` as the response and the other variables as predictors. Print the regression summary using the `summary` function.

(b) Perform the significance of the regression test for this full model at a 5% level of significance. Specify the null and alternative hypotheses. State the value of the test statistic, *p*-value, and a decision.

(c) Compute 90% and 95% CIs for the parameter associated with `age`. Using just these intervals, what could we have deduced about the *p*-value for `age` in the regression summary? [3.1 (a) from the textbook]

(d) Check the constant variance assumption for the errors [6.3 (a) from the textbook]. Plot the residuals vs. the fitted values. Conduct the appropriate inference test. BRIEFLY comment on the results.

(e) Check the normality assumption for the errors [6.3 (b) from the textbook]. Make a histogram and a Normal Q-Q plot for the residuals. Conduct the appropriate inference test. BRIEFLY comment on the results.

(f) Check for large leverage points (that is, identify point(s) with large leverage) [6.3 (c) from the textbook]. Report the observations with large leverage and their leverage values.

(g) Check for outliers [6.3 (d) from the textbook]. Report the potential outliers you find, first without Bonferroni correction, then with a Bonferroni correction.

(h) Check for influential points [6.3(e) from the textbook]. Report the observations with large influence and their Cook's distance.

(i) Remove all predictors (i.e. predictors used in part (a)) that are not significant at a 5% level. Generate a new, smaller model for `lpsa` as explained by only those significant predictors. Test this model against the full model in part (a) using an ANOVA test. Which model is preferred? [3.1 (d) from the textbook]