STAT 425 **Applied Regression and Design** **Spring 2020**

Homework No. 2

1 Instructions

Due Date: Friday Feb 21st, in Class
Homework presentation should be neat. You can submit the homework on loose leaf paper or submit through Compass. R codes should be submitted through Compass. **Please submit your digital files in .pdf or .html format**, so it is easier for the graders to open your files. If you submit in paper and more than one sheet of paper is used, the assignment should be stapled together. You must show all work for full credit. If you feel it would help, you are encouraged to work together on Homework, but you have to present assignments individually using your own words. The aim of the Homework is to learn the material and practice for the exams. **Late assignments will not be accepted**. Graduate students should attempt **all** problems. Undergraduate students can skip problems marked as GR.

2 Problems

1. **Problem 1**: Using the *sat* data from the *faraway* library:

   (a) Fit a model with *total* sat score as the response and *expend, ratio and salary* as predictors. Test the hypothesis that $\beta_{salary} = 0$. Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Do any of these predictors have an effect on the response?.

   (b) Now add takers to the model. Test the hypothesis that $\beta_{takers} = 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test is equivalent to the t-test.

2. **Problem 2**: For the *prostate* data from the *faraway* library, fit a model with *lpsa* as the response and the other variables as predictors:

   (a) Compare 90% and 95% CIs for the parameter associated with *age*.

   (b) Remove all predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

   (c) Compute and display a 95% joint confidence region for the parameters associated with *age* and *lph*. Plot the origin on this display. The location of the origin on the display tell us the outcome of a certain hypothesis test. State that test and its outcome.

   (d) In class we discussed a permutation test corresponding to the $F-$test for the significance of a set of predictors. Execute the permutation test corresponding to the t-test for *age* in this model. (*Hint*: summary(g)\$coef$[4, 3]$ gets you the t-statistic you need if the model is called g.)

3. **Problem 3**: In the *punting* data from the *faraway* library we find average distance and hang times of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.

   (a) Fit a regression model with Distance as the response, and the right and left strengths and flexibilities as predictors. Which predictors are significant at the 5% level?

   (b) Use an F-test to determine whether collectively these four predictors are significant at the 5% level.

   (c) Relative to the model in (a), test whether the right and left strength have the same effect.

   (d) Construct a 95% confidence region for $(\beta_{RStr}, \beta_{LStr})$. Explain how the test in (c) relates to this region.

   (e) Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response, in comparison to using individual left and right strengths.

   (f) (**GR**) Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.

   (g) (**GR**) Test for the left-right symmetry by performing the tests in (c) and (f) simultaneously.

   (h) (**GR**) Fit a model with *Hang* as the response, and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

4. **Problem 4 (GR)**: Find a formula relating R^2 and the F-test for the regression.

5. **Problem 5**: For the *prostate* data, fit a model with *lpsa* as the response and the other variables as predictors.

   (a) Suppose a new patient with the following values arrives:

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
1.44692	3.62301	65	0.3001	0	-0.79851	7	15

   Predict *lpsa* for this patient along with an appropriate 95% CI.

   (b) Predict the last question for a patient with the same values except that he is age 20. Explain why the CI is wider.

   (c) For the model of the previous question, remove all predictors that are not significant at the 5% level. Now recompute the predictions of the previous question. Are the CIs wider or narrower? Which predictions would you prefer?

6. **Problem 6 (GR)**: The multiple linear regression model can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

   where $Var(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$ and $\boldsymbol{I}$ is the $(n\times)$ identity matrix so that $Var(\boldsymbol{Y}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}$.

The fitted or predicted values are given by:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

Show that $Var(\hat{\boldsymbol{Y}}|\boldsymbol{X}) = \boldsymbol{\sigma^2 H}$