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

Homework No. 1

1 Instructions

Due Date: Friday Feb 07th, 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. 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**: The data set *prostate* from the *faraway* library, is from a study on 97 men with who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data. Commentfile:///C:/Users/lbravo/Documents/UIUC2020 on your results.

2. **Problem 2**: Show that for the SLR model, the coefficient of determination R^2 is equal to the square of the correlation coefficient r_{XY}^2.

3. **Problem 3**: The manager of the purchasing department of a large company would like to develop a regression model to predict the average amount of time it takes to process a given number of invoices. Over a 30-day period, data are collected on the number of invoices processed and the total time taken (in hours). The data are available in the file *invoices.txt*. The following model was fit to the data: $Y = \beta_0 + \beta_1 x + e$, where Y is the processing time and x is the number of invoices.

   (a) Plot the data and comment on the results.

   (b) Find a 95% confidence interval for the start-up time, ie., β_0.

   (c) Suppose that a best practice benchmark for the average processing time for an additional invoice is 0.01 hours (pr. 0.6 minutes). Test the null hypothesis $H_0 : \beta_1 = 0.01$ against a two-sided alternative. Interpret your result.

   (d) Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.

4. **Problem 4 (GR)**: Straight line regression through the origin

   In this question we shall make the following assumptions:

(1) Y is related to x by the simple linear regression model $Y_i = \beta x_i + e_i (i = 1, 2, \ldots, n)$, i.e. $E(Y|X = x_i) = \beta x_i$

(2) The errors $e_1, e_2, \ldots, e_n$ are independent from each other.

(3) The errors $e_1, e_2, \ldots, e_n$ have a common variance.

(4) The errors are normally distributed with a mean 0 and variance σ^2 (especially when the sample size is small), i.e., $e|X \sim N(0, \sigma^2)$.

In addition, since the regression model is conditional on X we can assume that the values of the predictor variable $x_1, x_2, \ldots, x_n$ are known fixed constants.

(a) Show that the least squares estimate of β is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(b) Under the above assumptions show that:

(i) $E[\hat{\beta}] = \beta$

(ii) $Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

(iii) $\hat{\beta}|X \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$

5. **Problem 5**: A story by James R. Hagerty entitled With Buyers Sidelined, Home Prices Slide published in the Thursday October 25, 2007 edition of the Wall Street Journal contained data on so-called fundamental housing indicators in major real estate markets across the US. The author argues that... *prices are generally falling and overdue loan payments are piling up.* Thus, we shall consider data presented in the article on:

Y= Percentage change in average price from July 2006 to July 2007 (based on the S& P/Case-Shiller national housing index); and

x= Percentage of mortgage loans 30 days or more overdue in latest quarter (based on data from Equifax and Moody's).

The data are available in the file *indicators.txt*. Fit the following model to the data: $Y = \beta_0 + \beta_1 x + e$. Complete the following tasks:

(a) Calculate the R^2 and adjusted R^2 for the SLR model. Provide an interpretation of both quantities.

(b) Find a 95% confidence interval for the slope of the regression model, β_1 . On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.

(c) Use the fitted regression model to estimate $E(Y|X = 4)$. Find a 95% confidence interval for $E(Y|X = 4)$. Is 0% a feasible value for $E(Y|X = 4)$? Give a reason to support your answer.

6. **Problem 6 (GR)**: In this problem we want to test that the identity: $SST = SSF + RSS$. In order to do that, test the following identities:

(a) Show that $(y_i - \hat{y}_i) = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})$

(b) Show that $(\hat{y}_i - \bar{y}) = \hat{\beta}_1(x_i - \bar{x})$

(c) Utilizing the fact that $\hat{\beta}_1 = \frac{SXY}{SXX}$, show that $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

(d) Finally test $SST = SSF + RSS$