University of California, Davis
Department of Statistics
Midterm 2– Spring 2021

Course Title: Applied Statistical Methods: Regression Analysis
Date of Examination: 5/21/2021
Teacher's name: Jairo Fúquene-Patiño
Student's name: ____________________
Course Code: STA 108

Sections: B01 and B02.
Time duration: 1 hour for Q.1. The deadline for Q.2 is 5/24/2021, 5:00 p.m.
Total marks: 100

Q.1) **Plastic hardness**. Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours? and Y is hardness in Brinell units (40 points).

i	Y_i	X_i
1	199	16
2	205	16
3	196	16
4	200	16
5	218	24
6	220	24
7	215	24
8	223	24
9	237	32
10	234	32
11	235	32
12	230	32
13	250	40
14	248	40
15	253	40
16	246	40

Assume first-order regression model, i.e.,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: is the value of the response variable in the $i-th$ trial.
- β_0 and β_1 are the parameters.
- X is a known constant, namely, the value of the predictor variable in the $i-th$ trial.
- ϵ_i is a random error term with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$ and follows a normal distribution $N(\epsilon_i; 0, \sigma^2)$

- ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $cov(\epsilon_i, \epsilon_j) = 0$ for all i, j; with $i \neq j$ and $i = 1, ..., n$)

Using matrix methods obtain the following (You can use the software R or R Studio) . **You need to show your solution, please submit your R code**:

- 15 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, $\boldsymbol{b}$ and SSE.
- 10 Compute $s\{b_0\}$, $s\{b_1\}$ and $s\{b_0, b_1\}$.
- 5 Find the matrix of the quadratic form for SSR.
- 10 Compute a 95% confidence interval for the parameter β_1.

40 marks

Q.2) The data set (Demographic.txt) provides selected county demographic information with 17 variables for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. For each geographic region, regress the number of serious crimes (Y) against population density (X_1: total population divided by land area), per capita income ($X2$), percent high school graduates (X_3), total personal income (X_4) and percent unemployment (X_5). Use first-order regression model with 5 predictor variables (60 points).

1. State the estimated regression functions.
2. Are the estimated regression functions similar for the four regions? Discuss.
3. Compute MSE and MSR for each model and using $\alpha = 0.01$ compute the p-value for the two alternatives in formula

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

versus

$$H_a\text{: not all } \beta_k, (k = 1, ..., 5), \text{ equal to zero.}$$

   state the decision rule and conclusion. Are these measures similar for the four regions?, discuss.
4. Obtain the residuals for each fitted model and prepare a box plot of the residuals for each fitted model.
5. State the conclusions.

60 marks