STAC67: Regression Analysis

Assignment 2
(Total: 100 points)

Please submit R Markdown file for Q. 1- Q. 3 along with your submission of the assignment.

Q. 1 (20 pts) This question is to practice R to generate fake data simulation from the regression model. Use the "vote.txt" data in Assignment 1.

When you generate a random number, use R code, **set.seed(your student number)** before the R codes of generating a random number, so that we can replicate the result.

We start by assuming true regression parameters in the model. Thus, we assume that $Y_i = 46.3 + 4X_i + \epsilon_i$, with $\epsilon_i \sim N(0, 3.9^2)$. We use the predictors X (growth) that we already have from "vote.txt".

- Step 1: Simulation of the fake data
  Simulate a vector Y of fake data and put this in a data frame with the same X (growth).
- Step 2: Fitting the model and keeping the estimated regression coefficients.
- Step 3: Repeating Step 1 and Step 2, 10,000 times.

(a) (5 pts) Do Step 1 and Step 2. Obtain the least square estimates of β_0 and β_1 with the fake data.
Also, compute estimated $E(Y|X_0 = 0.1)$ and obtain 95% confidence interval for $E(Y|X_0 = 0.1)$ by hands and compare it by R built-in function.

(b) (10 pts) Do Step 3. Make a histogram of 10,000 $\hat{\beta}_0$ and 10,000 $\hat{\beta}_1$. Superimpose (overlay) its theoretical distribution on each histogram. Calculate the mean and standard deviation of 10,000 estimates each. Are the results consistent with theoretical values?

(c) (5 pts) Do Step 3. Generate 10,000, 95% confidence interval for $E(Y|X_0 = 0.1)$. What proportion of the 10,000 confidence intervals for $E(Y|X = 0.1)$ includes $E(Y|X = 0.1)$? Is this result consistent with theoretical expressions?

Q. 2 (15 pts) This question is to practice R to build a R function.

Use a following simple dataset, build a box cox transformation function in R (follow the steps described in the lecture note) and compare the result with the built-in R function.

```
x <- c(0:9)
y <- c(98, 135, 162, 178, 221, 232, 283, 300, 374, 395)
```

Q. 3 (30 pts) (5 pts each) The dataset "kidiq.csv" is posted at Quercus. It contains children's test scores (Y = **kid.score**) and mother's IQ scores (X = **mom.iq**). The data is from a survey of adult American women and their children (a subsample from the National Longitudinal Survey of Youth). We fit a regression model predicting cognitive scores of preschoolers given their mothers' IQ scores.

(a) Fit a Simple Linear Regression relating test scores (Y) to mother's IQ scores (X) using R. Construct 95 % confidence interval for the mean test scores of all kids with their mother's IQ score = 110. Compute it by hands (use R) and compare the result with the built-in R function, predict().

(b) Construct a 99% prediction interval for a new kid's test score when his or her mother's IQ score = 110. Compute it by hands (use R) and compare the result with the built-in R function.

(c) Plot the residuals versus fitted values. Comment on the plot.

(d) Obtain a normal probability plot of residuals and test the hypothesis that the errors are normally distributed with the Shapiro-Wilk test. Comment on the graph and test result with $\alpha = 0.05$.

(e) We would like to conduct the **Breusch-Pagan test** to determine whether or not the error variance varies with the level of X. Install the package, "lmtest", and use the following R codes:

```
> library(lmtest)
> bptest(lm_object)
```

What is your test result with $\alpha = 0.05$?

(f) If there is evidence of non-normality or non-constant variance of errors, obtain a Box-Cox transformation (use the built-in function), and repeat the previous parts (d) and (e).

Q. 4 (20 pts) (5 pts each) A simple linear regression was fit, relating the modulus of a tire (Y) to the amount of weeks (X) heated at 125 Celsius, with results given below:

$X_i(Weeks)$:	0	1	2	4	6	15
$Y_i(Modulus)$:	2.3	4.2	5.2	5.9	6.3	7.2

Use the simple linear regression in matrix form.

(a) Obtain the design matrix $\mathbf{X}$ and $\underset{\sim}{Y}$.

(b) Obtain the vector of estimated regression coefficients, $\hat{\underset{\sim}{\beta}}$, and the vector of fitted value, $\hat{\underset{\sim}{Y}}$, and the residual vector, $\underset{\sim}{e}$.

(c) Compute the estimated variance-covariance matrix of $\hat{\underset{\sim}{\beta}}$, $\widehat{Var}(\hat{\underset{\sim}{\beta}})$.

(d) Find the hat matrix $\mathbf{H}$. What does $\sum_{i=1}^{n} h_{ii}$ equal? Here, h_{ij} is the element in $\mathbf{H}$ in the ith row and jth column.

(e) Find the estimated variance-covariance matrix of the residual vector, $\widehat{Var}(\underset{\sim}{e})$.

Q. 5 (15 pts) (5 pts each) An engineer is interested in the relationship between steel thickness (X) and its breaking strength (Y). She obtains the following matrices from a matrix computer package:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 60 \\ 60 & 360 \end{bmatrix} \quad \mathbf{X}'\underset{\sim}{Y} = \begin{bmatrix} 120 \\ 800 \end{bmatrix} \quad \underset{\sim}{Y}'(\mathbf{I}-\mathbf{H})\underset{\sim}{Y} = \mathbf{20}, \quad \underset{\sim}{Y}'(\mathbf{H}-\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J})\underset{\sim}{Y} = \mathbf{250}$$

(a) Construct the ANOVA table based on this information.

(b) Provide 95% confidence interval for β_1.

(c) Test $H_0 : \beta_1 = 0$ vs $\beta_1 \neq 0$ with $\alpha = 0.05$.