STAT 3008: Applied Regression Analysis
2019-20 Term 2 Assignment #4

Due: May 4th, 2020 (Monday) at 5:30pm

This assignment covers material from Section 6.1 to 8.3 of the lecture notes.

** Please submit the hardcopy of the R-code and R-outputs for Problem 2 and 3 (Quick and dirty is good enough, R markdown NOT recommended)

You need to show your calculation in details order to obtain full scores.

* Note that the solutions will be available on May 5th (Tuesday) at 1pm, as the final term exam will be on May 7th (Thursday). No late assignment will be accepted after the solutions are posted.

Problem 1 [30 points]: Consider simple linear regression $y_i = \beta_0 + \beta_1 x_i + e_i$, with E($e_i$) =0 and Var($e_i$) = σ^2 for i = 1, 2, …, n.

(a) [11 points] By simplifying the **Hat Matrix** $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$, show that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SXX}} \quad \text{for } i = 1, 2, \ldots, n$$

[Part (b) and (c)] Suppose x_n is a leverage point, with $x_n = a - (n-1)\delta$, but $x_i = a + \delta$ for i = 1, 2, …, n-1 for some constants a and $\delta \neq 0$.

(b) [6 points] Show that h_{nn} = 1.

(c) [5 points] Compute h_{ii} as a function of n for i = 1,2, …, n-1.

(d) [8 points] Suppose n=2m+1, with $x_1 = x_2 = \cdots = x_m = a + \delta$, $x_{m+1} = x_{m+2} = \cdots = x_{2m} = a - \delta$ and $x_{2m+1} = a$. Evaluate h_{ii} as a function of n for i = 1, 2, … n

**Note: Results for part (b) and (c) should be consistent with the $\mathbf{H}$ in Ch7 page 9; Results from part (b) and (d) should provide the upper and lower bounds for Property#5 on page 7.*

Problem 2 [23 points]: Suppose we want to explain Tension by Sulfur in the dataset "baeskel.txt" using a simple linear regression,

library(car);   library(alr3);   x<-baeskel$Sulfur;   y<-baeskel$Tension

(a) [4 points] Draw a scatterplot of the data using the *"plot"* function in R. Does the plot suggest a linear relationship between the two variables?

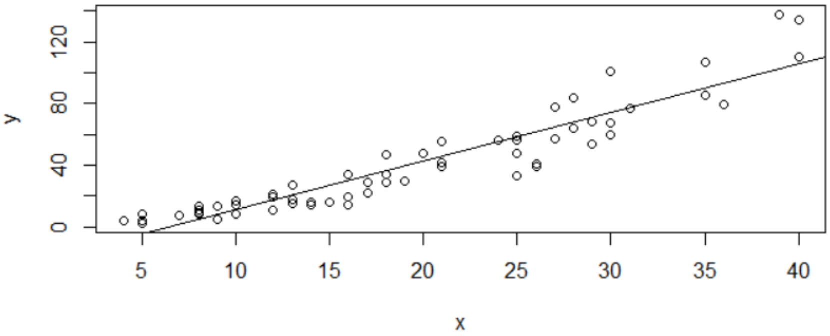
(b) [5 points] Suppose a simple linear regression $y_i = \beta_0 + \beta_1 x_i + e_i$ is fitted to the data. What is the regression equation based on OLS estimates and its R^2?

(c) [7 points] Generate the 4 residual plots based on the *"plot"* function (as in Ch7 page 30). Comment on the null plot assumption of the residuals.

(d) [7 points] Generate the table of influence diagnostics using the *"influence.measures"* function in R. What conclusion can you draw from each of the following measures?

(i) DFFITS   (ii) DFBETAS   (iii) Cook's Distance   (iv) Leverage

Problem 3 [47 points]: The data set "*stopping*" in alr3 contains hypothetical data to explain the *distance* (in feet) required to stop an automobile, based on its *speed* (miles per hour) right before the brake is applied. *library(alr3); x<-stopping$Speed; y<-stopping$Distance; plot(x,y)*



(a) [5 points] Suppose a quadratic regression is fitted to the data

(Model Q) $y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + e_i$ with $E(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$

What are the OLS estimates $\hat{\alpha}_0, \hat{\alpha}_1$ and $\hat{\alpha}_2$, and the RSS of the model?

(b) [8 points] Suppose **Model Q** is the full model. Using the "*stepAIC*" function (Ch6 p.26), show that the parsimonious model based on AIC and forward selection is

(Model P) $y_i = \alpha_0^* + \alpha_2^* x_i^2 + e_i$ with $E(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$

What are the OLS estimates $\hat{\alpha}_0^*$ and $\hat{\alpha}_2^*$, and the RSS of the model?

(c) [8 points] Use the "plot" function to obtain the residual plots (as in Ch7 p30) for **Model P**. Which of the null plot assumption (i.e. constant mean, constant variance and separated points) is invalid based on plots? Explain.

[part (d) to (e)] Suppose we apply the scale power transform $\psi_s(x, \lambda) = (x^\lambda - 1)/\lambda$ to *x*, where λ = 1.5, 2.0 and 2.5. Consider the regression model with mean function

(Model λ) $E(y \mid X = x) = \beta_0 + \beta_1 \psi_s(x, \lambda)$

(d) [7 points] In the original scatterplot (i.e *x* vs *y*), draw the fitted curves for **Model λ** with λ = 1.5, 2.0 and 2.5 based on Approach #1 on Ch8 page 11.

(e) [10 points] Compute the RSS of the 3 models (λ = 1.5, 2.0, 2.5). Show that (i) λ = 2.0 is the best model among the 3, and (ii) explain why RSS(Model λ = 2.0) = RSS(Model P).

(f) [9 points] Suppose a simple linear regression is fitted to transformed data based on power transform $\psi(u, \lambda) = u^\lambda$ as follows: $\psi(y, \lambda) = \beta_0 + \beta_1 \psi(x, \lambda) + e$.
For each of λ = 0.2, 0.4, 0.67 and 1.0, draw a scatterplot of $(\psi(x, \lambda), \psi(y, \lambda))$ with the inclusion of the corresponding fitted regression line. Which λ is able to provide the smallest number of leverage points?

\- **End of the Assignment** -