TEXAS A&M UNIVERSITY

Department of Statistics

STAT 608 - Regression Analysis

Homework II

Salih Kilicli

June 8, 2019

Question 1: Suppose you have an experiment with design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 0 & 2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

parameters γ and α and responses $y_1, ..., y_4$, i.e.

$$\begin{aligned} y_1 &= \gamma + \alpha + e_1 \\ y_2 &= 2\alpha + e_2 \\ y_3 &= \gamma + e_3 \\ y_4 &= \alpha + e_4 \end{aligned}$$

Is it true that

$$\hat{e_1} + 2\hat{e_2} + \hat{e_4} = 0 \ ?$$

Show your calculations.

Solution: **<u>Method I:</u> Yes, it is true. Let,**

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \gamma \\ \alpha \end{bmatrix}, \mathbf{E} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

then, the matrix equivalent form of the set of equations given above is:

$$Y = XB + E$$

Since

$$RSS = E^T E = (Y - XB)^T (Y - XB) = Y^T Y - 2Y^T XB + B^T X^T XB$$

minimizing RSS (i.e., solving $\frac{\partial RSS}{\partial B} = 0$ **for B) yields, the estimator:**

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{\gamma} \\ \hat{\alpha} \end{bmatrix} = (X^T X)^{-1} X^T Y$$

After calculating

$$\mathbf{X^T X} = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}, \ \mathbf{X^T X}^{-1} = \frac{1}{11} \begin{bmatrix} 6 & -1 \\ -1 & 2 \end{bmatrix} \text{ and } \mathbf{X^T X^{-1} X^T} = \frac{1}{11} \begin{bmatrix} 5 & -2 & 6 & -1 \\ 1 & 4 & -1 & 2 \end{bmatrix}$$

we obtain;

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{\gamma} \\ \hat{\alpha} \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 5y_1 - 2y_2 + 6y_3 - y_4 \\ y_1 + 4y_2 - y_3 + 2y_4 \end{bmatrix}$$

Therefore,

$$\hat{e_1} + 2\hat{e_2} + \hat{e_4} = y_1 + 2y_2 + y_4 - \hat{\gamma} - 6\hat{\alpha} = y_1 + 2y_2 + y_4 - \frac{1}{11}\Big[11y_1 + 22y_2 + 0y_3 + 11y_4\Big] = 0$$

<u>Method II:</u> From page 18, normal equations, we have:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \sum_{i=1}^{4} \hat{e_i} x_i = X^T \hat{E} = \begin{bmatrix} \hat{e_1} + \hat{e_3} \\ \hat{e_1} + 2\hat{e_2} + \hat{e_4} \end{bmatrix}$$

Question 2: A botanist is interested in the efficacy on predator bugs in reducing pests on garden plants. In particular, two species (A and B) of praying mantis are to be compared to see which devours potato beetles at a higher rate. One hundred grams of potato beetles are released into a cage containing potato foliage, and one praying mantis of each species is introduced into the cage for one week. At the end of the week, the reduction (in grams) of potato beetles is measured. Another identical cage is prepared, but this time, there are two praying mantis of Species A and one of Species B. For the third cage, there are two of Species B and one of Species A, and for the fourth cage of the study, there are two of each species introduced. Let A be the average grams of potato beetles eaten per week per praying mantis for Species A, and let B be the average for Species B. Write down a **linear model** in algebraic or matrix form (whichever you prefer) that can be used to estimate A and B. Assume that the consumption of each praying mantis is independent of others in the cage. Assume also that there is no natural attrition of potato beetles over a period of one week.

Solution : **A linear model for the problem can be defined as;**

$$\begin{aligned} y_1 &= \beta_A + \beta_B + e_1 \\ y_2 &= 2\beta_A + \beta_B + e_2 \\ y_3 &= \beta_A + 2\beta_B + e_3 \\ y_4 &= 2\beta_A + 2\beta_B + e_4 \end{aligned}$$

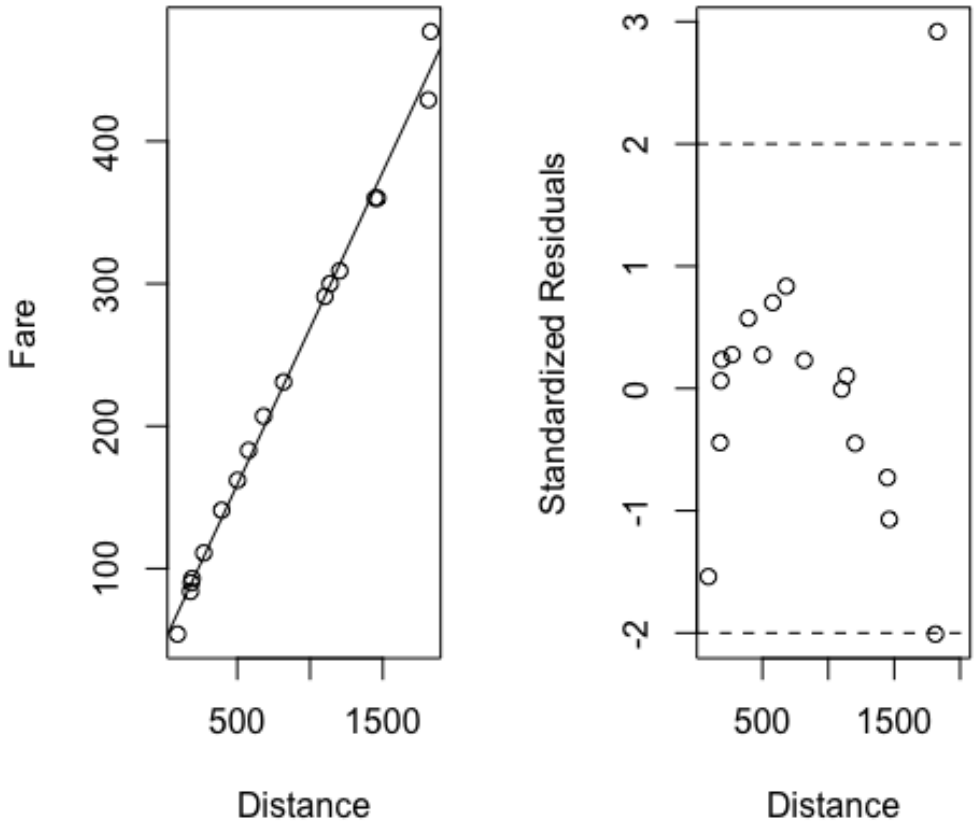
or in matrix form,

$$Y = X\beta + E$$

where,

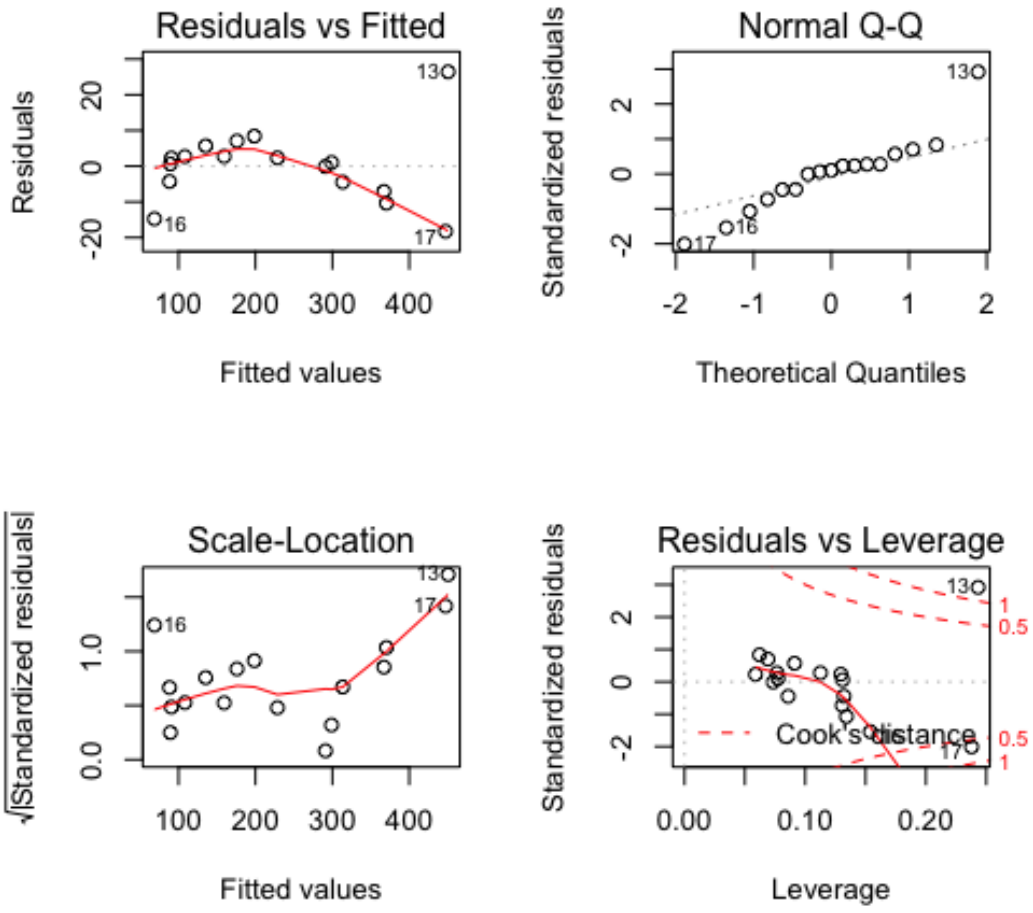
$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 2 \\ 2 & 2 \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix}, \ \textbf{and } \mathbf{E} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}.$$

Question 3: Work Exercise 1 on page 103 of our textbook.



Solution: **(a) Both the standardized residuals vs distance and the residual vs fitted values plots reveal a nonlinear trend in the residuals and there are two (or possibly more, including Data point 16) leverage points, one of which "bad" (Data point 13, also an outlier) and the other is "good" (Data point 17). The leverage points greatly reduce the confidence in the ordinary least squares assumptions. It is true that the "Distance" explain a large about of the overall variation in the "Fare", leading to a large R^2 value; however, the model needs improvements.**

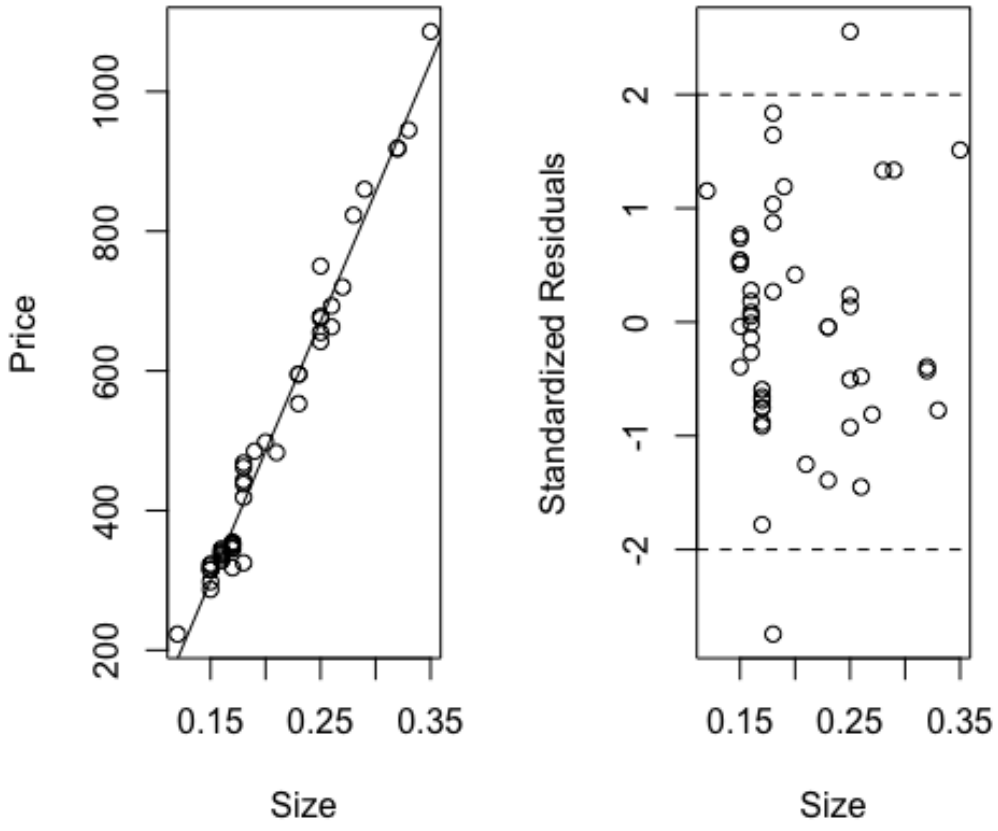
(b) Even though, the straight line seem to fit the data well and being supported by numerical results, due to outlier with Cook's distance bigger than 1, the extreme data points are need to be taken care of. Also, Scale-Location plot shows the non-constant error variance clearly that leads to incorrect results. As mentioned in the first part, Residuals vs Fitted plot implies the non-linear trend in the data points. Therefore, either by using some appropriate transformations, or ,if necessary, removing the extreme values might lead a better and more accurate model.



Question 4: Work Exercise 6 on page 112 of the textbook.

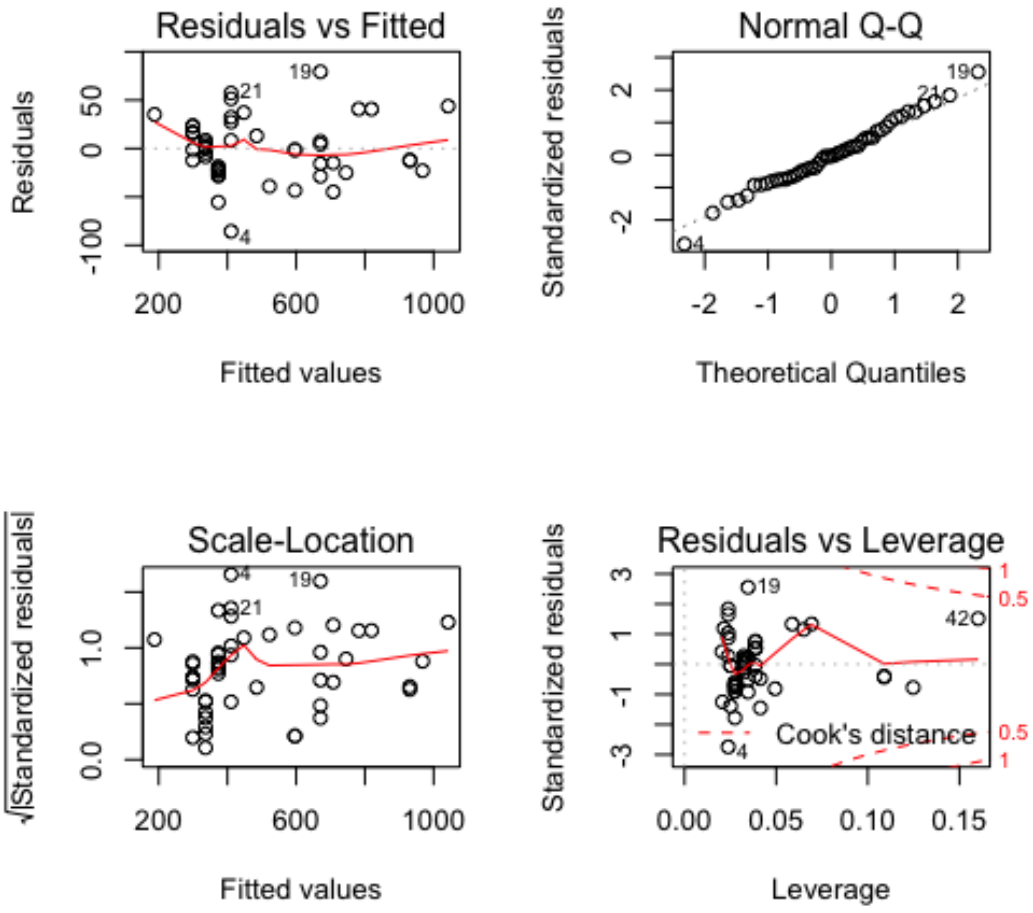
Solution: The **inverse response** plot of $\hat{y}$ vs y gives accurate information about the transformation **when the distribution of Y is skewed, and the distribution of x is elliptically symmetric (or, a stronger assumption, being normally distributed)**. However, in this problem, the distribution of x is **highly skewed**. Therefore, it leads to an incorrect estimate of λ.

Question 5: Work Exercise 1, Part I only, on page 112 of the textbook.



Solution: **(a) The model is given by $y = -258.1 + 3715x$ The model reason for the linear choice, is due to apparent linear relation between price and size data.**

(b) The model has a good summary statistics with $2e^{-16}$ p-values for the coefficients and with a 0.9785 R-squared value. However, from the Standardized Residuals vs Size plot it is apparent that we have 2 "bad" leverage points (also outliers) since their leverage values lie outside of the $[-2, 2]$ interval. Also, the Residuals vs Fitted plot shows a slight non-linear trend. Moreover, Scale-Location plot shows an increasing trend, which implies a non-constant error variance. The model still can be improved either by adding a nonlinear term or/and applying transformations to get a constant error variance.



Question 6: When $Y(>0)$ has mean and variance both equal to μ it is shown on pages 76 -77 of our textbook that the appropriate transformation of Y to stabilize variance is the square root transformation. Now, suppose that Y has mean equal to μ and variance equal to μ^2. Find the transformation $Z = f(Y)$ of Y that makes the variance of Z approximately equal to 1.

Solution: **Let $Z = f(Y)$ and consider the Taylor expansion of $f(Y)$ only with linear terms;**

$$Z = f(Y) \approx f(E[Y]) + f'(E[Y])(Y - E[Y])$$

Taking variance of the both sides of the equation above yields;

$$1 \approx Var(Z) = Var(f(Y)) \approx [f'(E[Y])]^2 Var(Y)$$

Since $E(Y) = \mu$ and $Var(Y) = \mu^2$, we obtain;

$$f'(\mu) = \frac{1}{\mu}$$

Integrating both sides of the equation above with respect μ;

$$\int f'(\mu) d\mu = \int \frac{1}{\mu} d\mu$$

$$f(\mu) = log(\mu) + C$$

Therefore, the transformation $Z = f(Y) = log(Y) + C$, where C is a constant (of integration, might be included in the error while regressing as well).