STAT 408/608 Homework 4 Solutions: Written Section

February 19, 2015

1. We cannot simply take the inverse transformation because the expected value of the inverse transformation of Y is not equal the value of the function when evaluated at the expected value of Y, $E[f(X)] \neq f(E[X])$.
   When we transform the data, the interval may change and when we try to back-transform the interval, we will need to add an additional term in order for the interval to agree with the original data. This is known as the correction factor and it changes for a given transformation.

2. In the previous homework problem question 3, $\hat{\beta}$ gives the average for each of the two groups.
   In this two dummy variable case, take $m = 3, n = 4$ for example, the design matrix and the hat matrix are:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, H = X(X'X)^{-1}X' = \begin{bmatrix} 1/3 & 1/3 & 1/3 & & & & \\ 1/3 & 1/3 & 1/3 & & & & \\ 1/3 & 1/3 & 1/3 & & & & \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

   $\hat{y} = Hy$, the hat matrix projects y into its average.

3. (a) In this problem, the model $Y = X\beta + e$ can be written as:
   $y = X\hat{\beta} + \hat{e} \Rightarrow \hat{e} = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = y - Hy = (I - H)y$.

   (b) Idempotent:
   $HH = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}IX' = X(X'X)^{-1}X' = H$.
   Symmetric:
   $H' = (X(X'X)^{-1}X')' = X(X'X)^{-1}X' = H$.
   $(I - H)(I - H) = I - H - H + HH = I - H$.
   Therefore, $Cov(\hat{e}) = Cov((I - H)y) = (I - H)Cov(y)(I - H)' = (I - H)\sigma^2 I(I - H)' = \sigma^2(I - H)$.

   (c) From (b), when $i \neq j, Cov(\hat{e}_i, \hat{e}_j) = (0 - h_{ij})\sigma^2 = -h_{ij}\sigma^2$.

4. (a) $H' = (X(X'X)^{-1}X')' = X(X'X)^{-1}X' = H$.

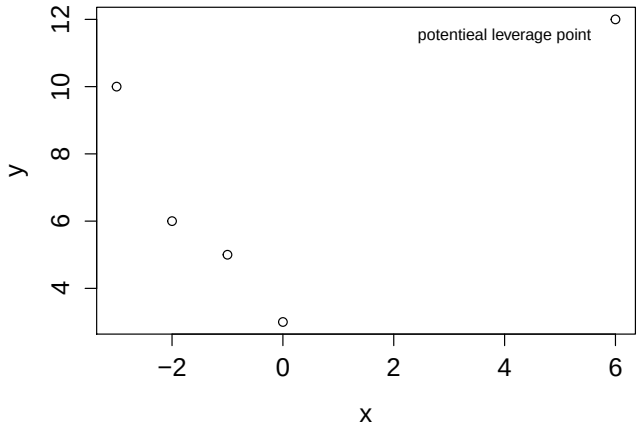

Figure 1: Scatterplot of the data

(b) H is a symmetric matrix, and it implies that $h_{ij} = h_{ji}$.
$h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \geqslant h_{ii}^2 \Rightarrow h_{ii}(1 - h_{ii}) \geqslant 0 \Rightarrow 0 \leqslant h_{ii} \leqslant 1$

(c) Please see textbook P153-154.

(d) When $n \to \infty$, the denominators of h_{ij} will be very large and $h_{ij} \to 0$. For a fixed sample size, it makes sense that the covariances are small numbers.

5. (a) Figure 1 shows the scatterplot. The design matrix:

$$X = \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 6 \end{bmatrix}$$

(b)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = 7.2 + 0.5 \begin{bmatrix} -3 \\ -2 \\ -1 \\ 0 \\ 6 \end{bmatrix} = \begin{bmatrix} 5.7 \\ 6.2 \\ 6.7 \\ 7.2 \\ 10.2 \end{bmatrix}$$

$$\hat{e} = (y - \hat{y}) = \begin{bmatrix} 4.3 \\ -0.2 \\ -1.7 \\ -4.2 \\ 1.8 \end{bmatrix}$$

(c)

$$X'X = \begin{bmatrix} 5 & 0 \\ 0 & 50 \end{bmatrix}, (X'X)^{-1} = \frac{1}{50}\begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}, H = X(X'X)^{-1}X' = \frac{1}{50}\begin{bmatrix} 19 & 16 & 13 & 10 & -8 \\ 16 & 14 & 13 & 10 & -2 \\ 13 & 12 & 11 & 10 & 4 \\ 10 & 10 & 10 & 10 & 10 \\ -8 & -2 & 4 & 10 & 46 \end{bmatrix}$$

Using the rule $h_{ii} > 4/n = 4/5 = 0.8$, we can see $h_{55} = 46/50 > 0.8$, the point (6,12) is a potential leverage point, and it is "bad" because it does not fall near the line that would fit the other points.

(d)

$$Var(\hat{e}_i) = \sigma^2(1 - h_{ii}) = \frac{\sigma^2}{50}\begin{bmatrix} 31 \\ 36 \\ 39 \\ 40 \\ 4 \end{bmatrix} = \sigma^2 \begin{bmatrix} 0.62 \\ 0.72 \\ 0.78 \\ 0.80 \\ 0.08 \end{bmatrix}$$

The residual e_4 has the highest variance.

(e) The standardized residual $= \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}} \begin{bmatrix} 1.454 \\ -0.063 \\ -0.513 \\ -1.251 \\ 1.695 \end{bmatrix}$

The point 5 has the highest standardized residual in absolute value. Because the standardized residuals are obtained by dividing the residuals by the variance of the residuals, points with low variance will tend to have high standardized residuals.

(f) Since he variance is given by $\sigma^2(I - H)$, points with high leverages will tend to have lower variances.

6. Suppose $f(Y) = log(Y), E(Y) = \mu, Var(Y) = \mu^2$. Using Taylor series expansion,
$f(Y) = f(\mu) + f'(\mu)(Y - \mu) + \cdots$
$Var(f(Y)) = (f'(\mu))^2 Var(Y) = \frac{\mu^2}{\mu^2} = 1$
Because variance of $f(Y)$ is a constant, the variance has been stabilized and the log

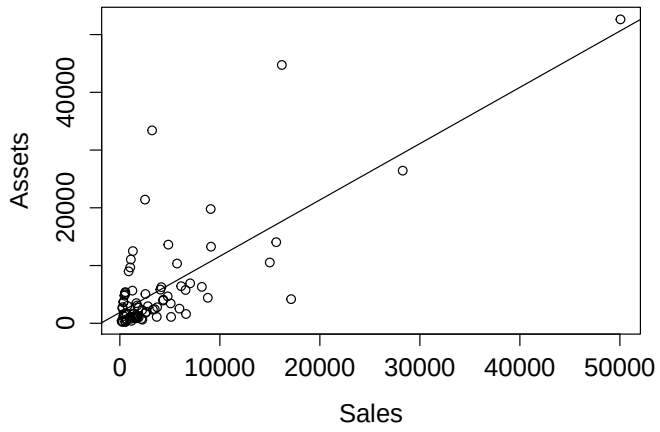

Figure 2: Scatterplot of the data and fitted line before transformation

transformation is appropriate.

7. (a) Figure 2 gives the scatterplot of sales and assets. In Figure 3, the Q-Q plot shows that the residuals are not normally distributed. The first and third plot show that most fitted values are on the left-hand side of the graph. The Scale-Location plot shows that observations 16, 48, and 54 are outliers. The residual vs leverage plot shows that observation 40 is a high leverage point, and observation 16 is an outlier.

   (b) I would like to use the box-cox transformation to see which power is the most appropriate value for doing the transformation. (We know that when $\lambda = 0$, it is the log transformation.)
   By using the R code, we find that $\lambda = -0.068$ is the most appropriate power value shown in Figure 4, which is very close to 0. Therefore, the log transformation for sales is an appropriate transformation.

   (c) Do the same procedure as (b). Figure 5 shows the power value of $\lambda = -0.043$, which is very close to 0. Therefore, the log transformation for assets is an appropriate transformation.

   (d) After log transformation, the Sales and Assets are linear related, as shown in Figure 6.
   The diagnostic plots improve a lot compared with model 1 shown in Figure 7. The residual plot is less concentrated than model 1, and Q-Q plot are more close to normal distribution. In model 2, all assumption are met: linear relationship,

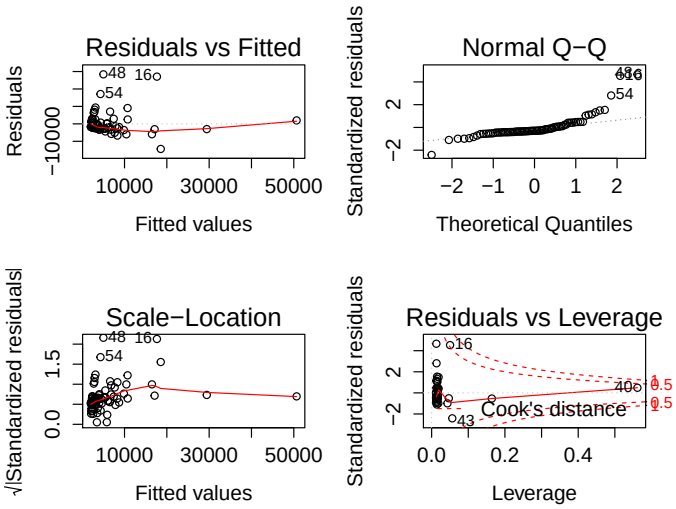

Figure 3: Diagnostic plot

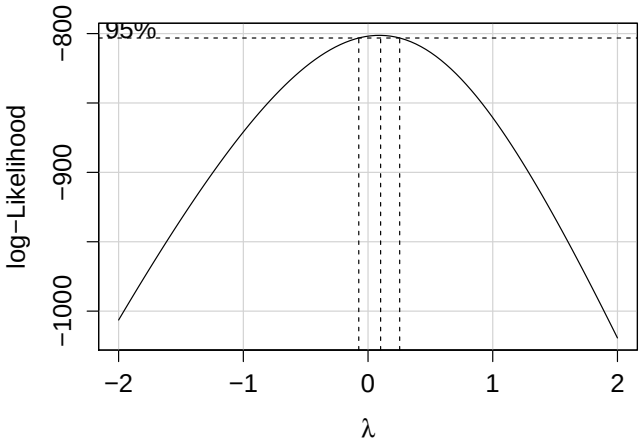

Figure 4: Box-Cox transformation for Sales

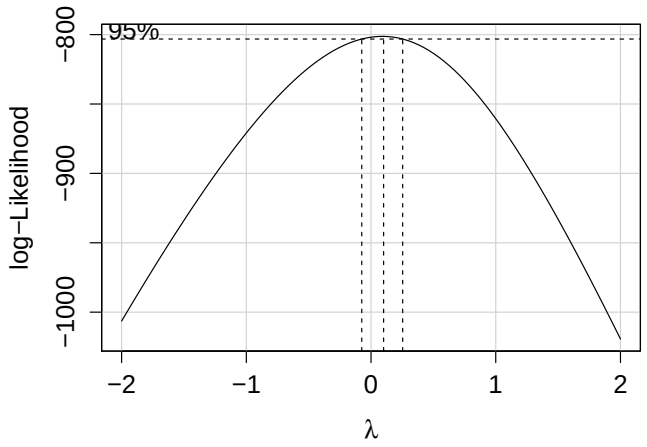

Figure 5: Box-Cox transformation for Assets

independent observations, normality of residual and constant variance.

(e) As shown in (d), model 2 met all the assumptions: linear relationship, independent observations, normality of residual and constant variance. therefore, I prefer the model 2.

(f) The slope is 0.587, which corresponds to the percentage change in assets for every percentage change in sales. This can be interpreted as that every 1% increase in sales results in an approximately 0.587% increase in assets.

(g) To transform this interval into the original units, add one half of the mean square error to each the interval and apply the exponential function, we have 95% confidence interval (6,870, 12,889).
I am 95% confident that a company with 6,571 million in sales would have the average Assets between 6,870 and 12,889 million.

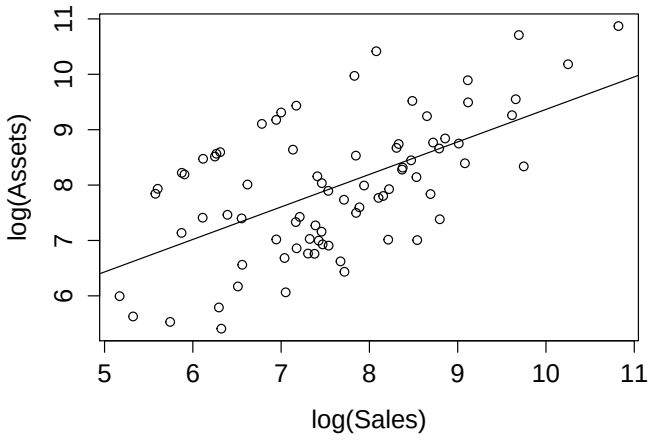

Figure 6: The relationship between Sales and Assets after transformation

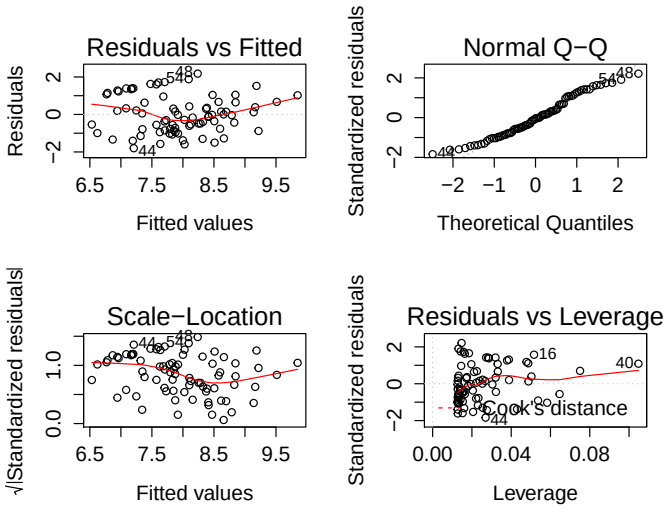

Figure 7: Diagnostic plots after transformation