Homework 04

Joseph Blubaugh
jblubau1@tamu.edu

STAT 608-720

1. In order get confidence intervals with transformed variables we have to first back transform the coefficients so that they are in the same scale as the original data

2. The hat matrix projects the values of x onto y. The trace of the matrix shows the influence that each value of x has on y. It makes sense that all values are either 1 or 0 because the variables are indicator variables whos value can only be 1 or 0.

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1_m & 0_{m+1} & 0_n \\ 0 & 0 & 0_m & 1_{m+1} & 1_n \end{bmatrix}$$

3. a)

$$\begin{aligned} \hat{e} &= (I-H)y \\ &= (y - Hy) \\ &= y - X\hat{B} \end{aligned}$$

b)

$$\begin{aligned} (I-H)' &= (I-H) \\ (I-H)(I-H) &= (I-H) \\ \sum &= \sigma^2 I \end{aligned}$$

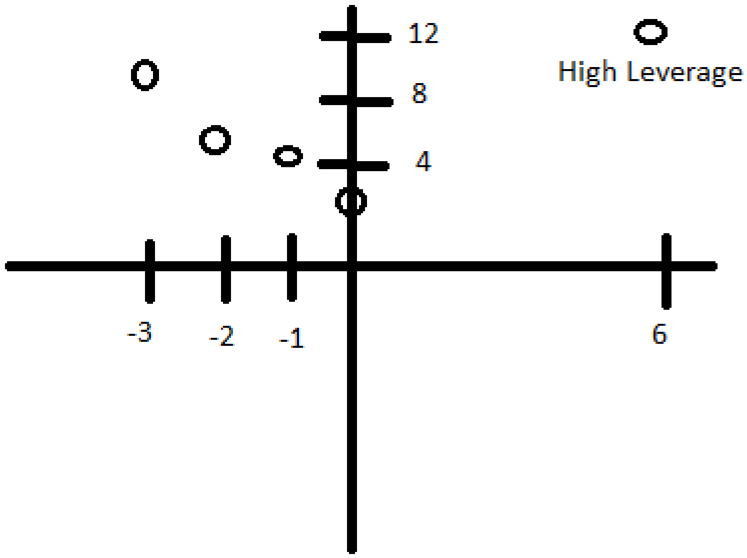
$$(I-H)' \sum (I-H) = \sigma^2 (I-H)$$

c)

$$\begin{aligned} \sigma^2(I-H) &= I\sigma^2 - H\sigma^2 \\ &= -H\sigma^2 \end{aligned}$$

4. a)

$$\begin{aligned} H' &= (X(X'X)^{-1}X')' \\ &= X\left[(X'X)^{-1}\right]'X' \\ &= X\left((X'X)'\right)^{-1}X' \\ &= X(X'X')^{-1}X' \\ &= H \end{aligned}$$



b) h_{ii} is garunteed to be a fraction because the individual errors are divided by the sum of all errors of x for each iteration so the maximum that leverage could be is 1.

$$
\begin{aligned}
h_{ii} &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} \\
SXX &= \sum (x_i - \bar{x})^2 \\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}
\end{aligned}
$$

c) When x_i or x_j are equal to to $\bar{x}$, $h_{ij} = \frac{1}{n}$, for all other values h_{ij} is greater than $\frac{1}{n}$

d) When variables are independent their covariance is 0. When variables are independent, their errors will be independent. The reason that there are a small ammount of covariance is because of the differences between x_i, y_i and $\bar{x}, \bar{y}$. These small differences divided over N cause covariance to be slightly different from zero.

5. a) Design matrix

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

b)

$$\begin{aligned}
e_1 &\to (7.2 + .5(-3)) = 7.2 - 1.5 \to 5.7 - 10 = -4.3 \\
e_2 &\to (7.2 + .5(-2)) = 7.2 - 1 \to 6.2 - 6 = .2 \\
e_3 &\to (7.2 + .5(-1)) = 7.2 - .5 \to 6.8 - 5 = 1.3 \\
e_4 &\to (7.2 + .5(-0)) = 7.2 - 0 \to 7.2 - 3 = 4.2 \\
e_5 &\to (7.2 + .5(-6)) = 7.2 - 3 \to 4.2 - 12 = -7.8
\end{aligned}$$

c) observation 5 is very close to a bad leverage point, but looking at the data graphically it looks like a horrible observation that should be looked at closer.

$$\begin{aligned}
SXX &= 50 \\
h_{ii} &= \frac{1}{5} + \frac{-3^2}{50} = \frac{9}{50} \\
h_{ii} &= \frac{1}{5} + \frac{-2^2}{50} = \frac{4}{50} \\
h_{ii} &= \frac{1}{5} + \frac{-1^2}{50} = \frac{1}{50} \\
h_{ii} &= \frac{1}{5} + \frac{0^2}{50} = \frac{0}{50} \\
h_{ii} &= \frac{1}{5} + \frac{6^2}{50} = \frac{36}{50}
\end{aligned}$$

d)

$$\begin{aligned}
Var(\hat{e}) = & \frac{(5.7 - 10)^2}{4} = 4.61 \\
& + \frac{(6.2 - 6)^2}{4} = .01 \\
& + \frac{(6.8 - 5)^2}{4} = 0.81 \\
& + \frac{(7.2 - 3)^2}{4} = 4.4 \\
& + \frac{(4.2 - 12)^2}{4} = 15.21 \\
& = 25.06
\end{aligned}$$

e) It seems that the points with the highest error have smaller impacts when it comes to standardized residuals.

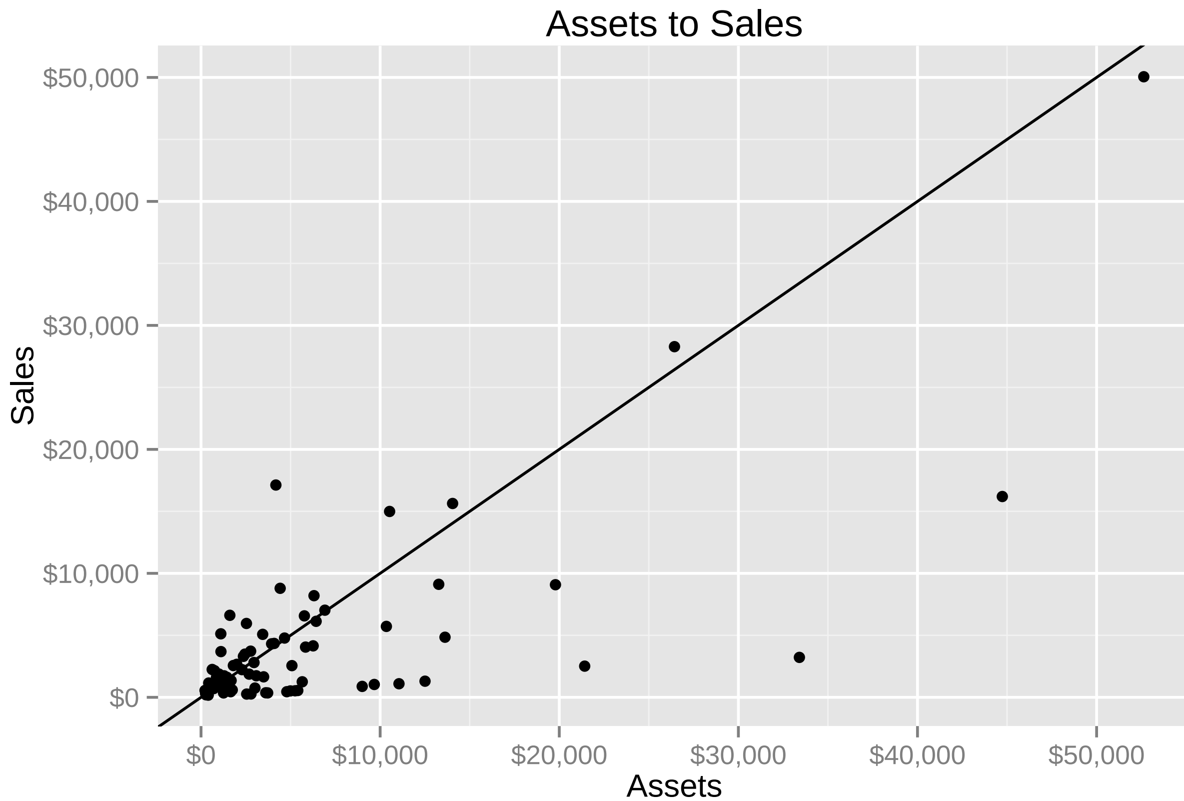
$$
\begin{aligned}
r_1 &= \frac{-4.3}{10\sqrt{1-9/50}} \\
r_2 &= \frac{.2}{10\sqrt{1-4/50}} \\
r_3 &= \frac{1.3}{10\sqrt{1-1/50}} \\
r_4 &= \frac{4.2}{10\sqrt{1-0}} \\
r_5 &= \frac{-7.8}{10\sqrt{1-36/50}}
\end{aligned}
$$

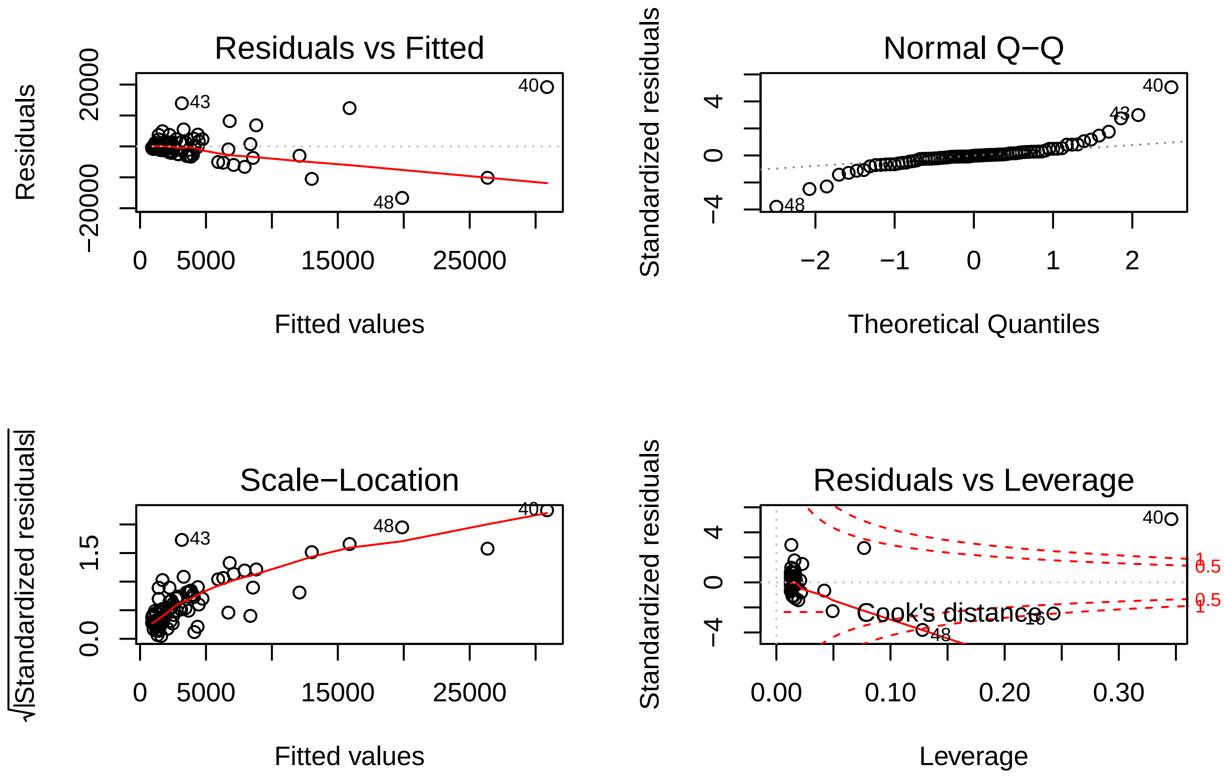
f) High leverage points are points that are very close and very far from $\bar{x}$. The highest leverage point is a good leverage point because it is close to $\bar{x}$.

6.

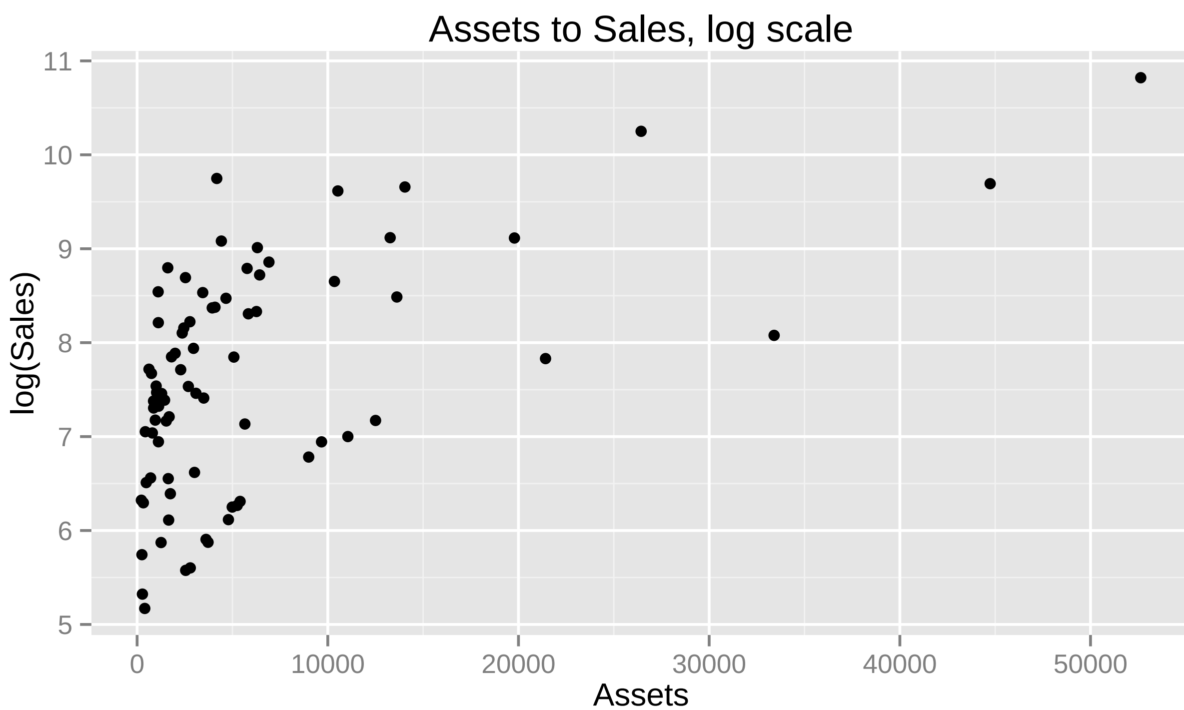
$$
\begin{aligned}
Var(Y) &= \mu^2 \\
log(Var(Y)) &= log(\mu^2) \\
log(Var(Y)) &= 2log(\mu) \\
\frac{log(Var(Y))}{2} &= log(\mu)
\end{aligned}
$$

7. a) A plot of the data and initial model suggests that a log transformation might be a better fit than no transformation.

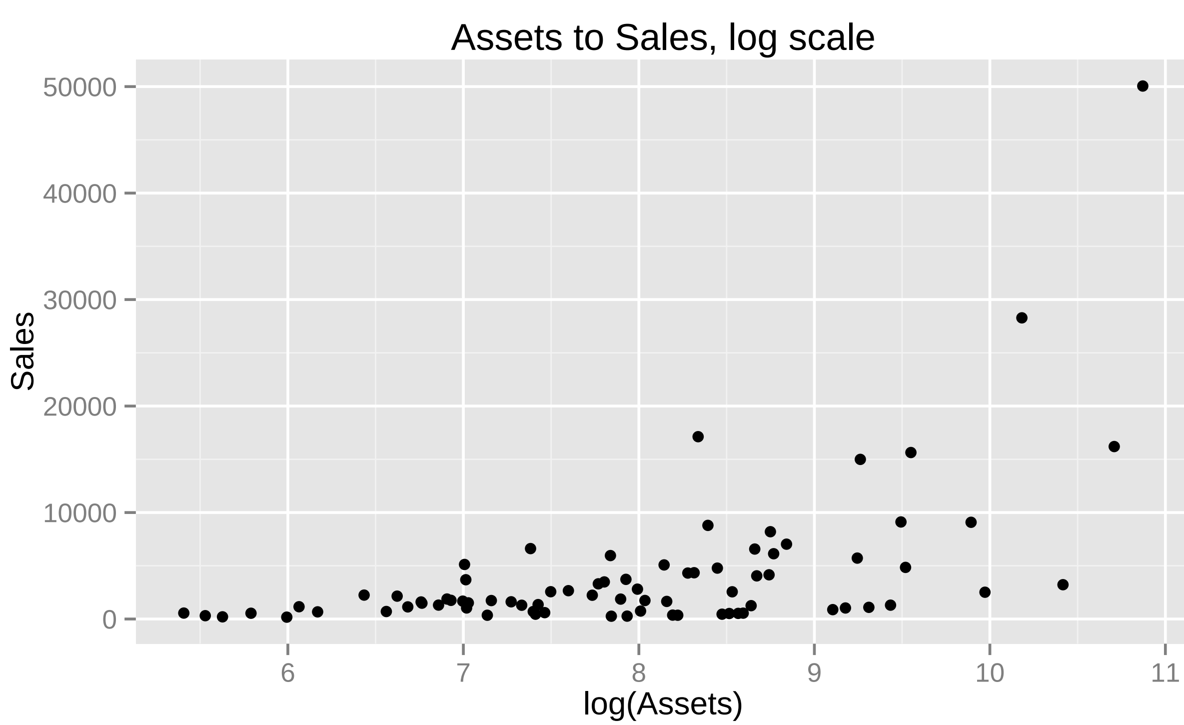

Assets to Sales
$50,000
$40,000
$30,000
$20,000
$10,000
$0
Sales
$0
$10,000
$20,000
$30,000
$40,000
$50,000
Assets



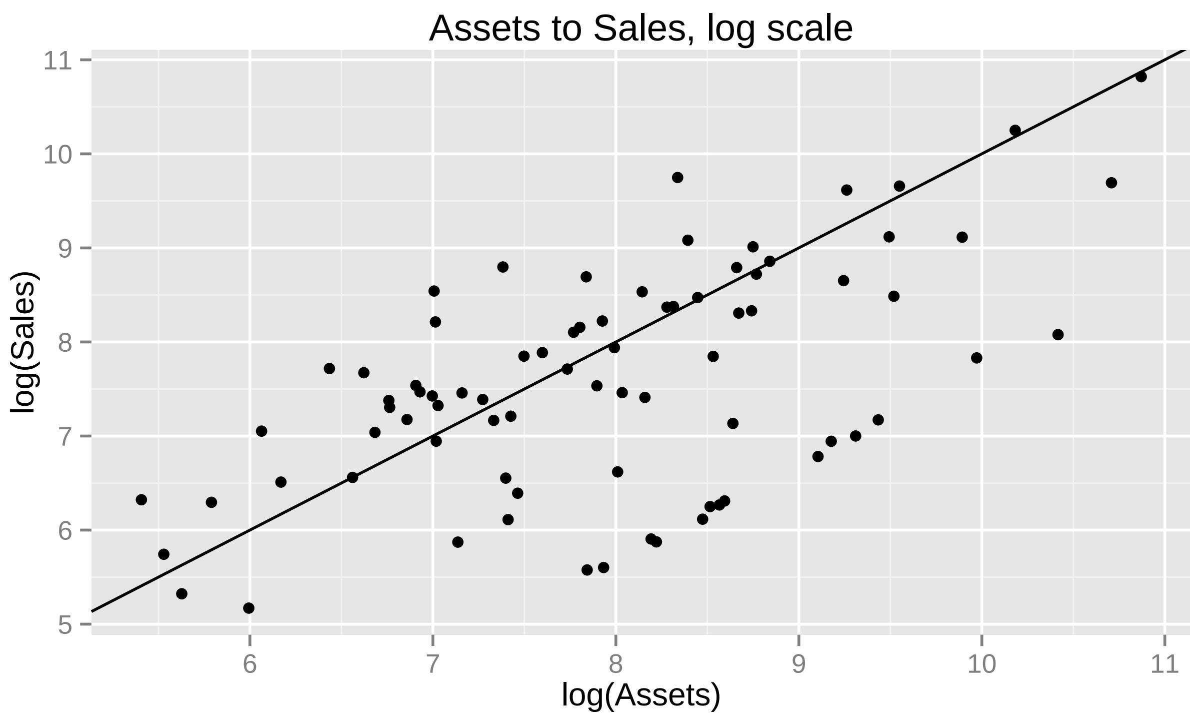
b) Using a log transformation on sales takes care of the extreme values for sales, but leaves the extreme values of assets which looks exponential with only one axis transformed

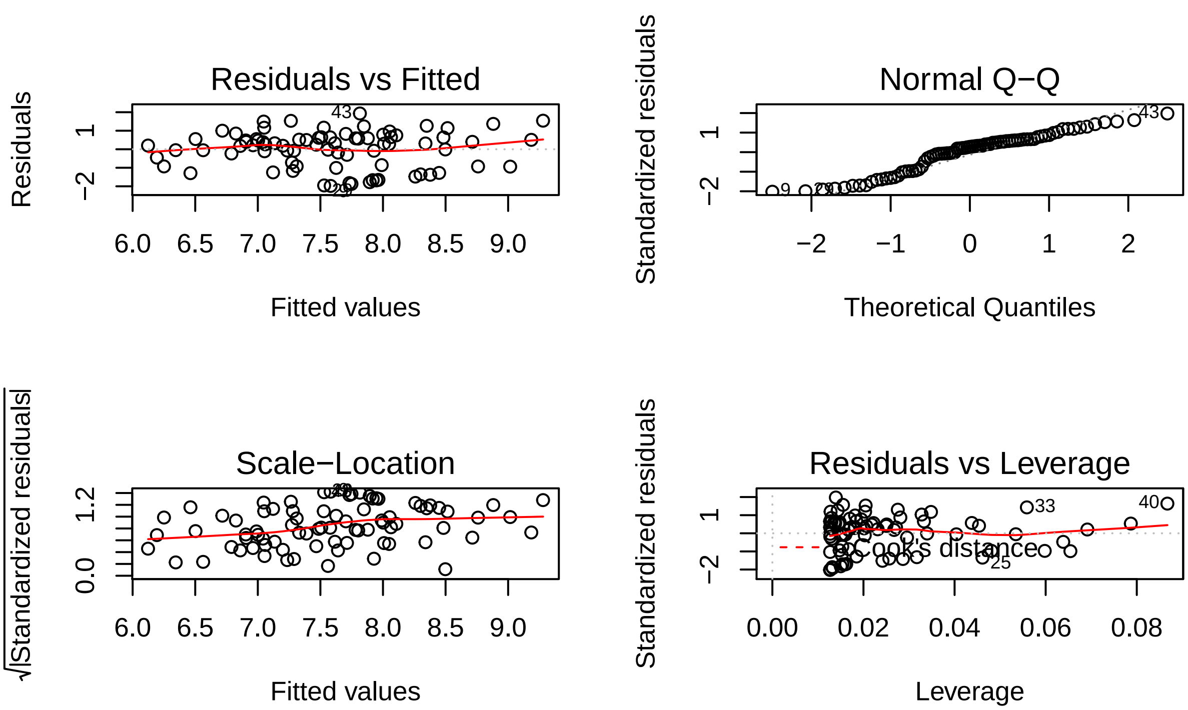


c) Using a log transformation on assets takes care of the extreme values of assets but leaves the extreme values of sales which also looks exponential.



d) Overall the R^2 is lower for this model meaning it explains less of the variance than the non-transformed model. Most of the diagnostic plots look okay, but one problem with this model is the distribution of the residuals which do not appear to normally distributed.





e) Model 2 is preferable to model one overall. The errors are distributed better and there are no bad high leverage points in the second model. Even though model 2 has a lower R^2 it is still a better fitting model.

f) A one percent increase in assets on average will result in .557 percent increase in sales.

g) The average prediction for a company with assets of $6,571M is sales of $3,220M with a 95 percent confidence interval about the mean between $2,461M and $4,211M