STAT 608 Homework 4

Jack Cunningham

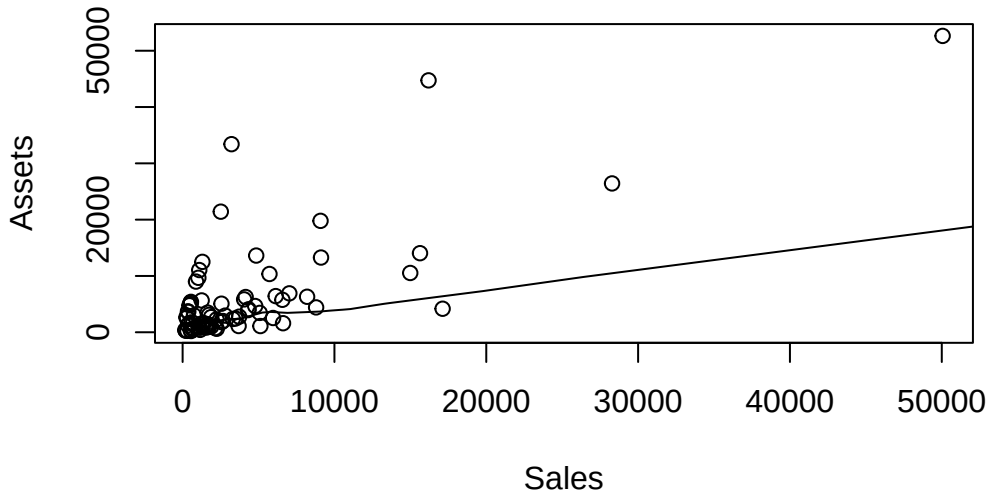
8)

Loading Company Data

```
company <- read.csv("company.csv")
attach(company)
```

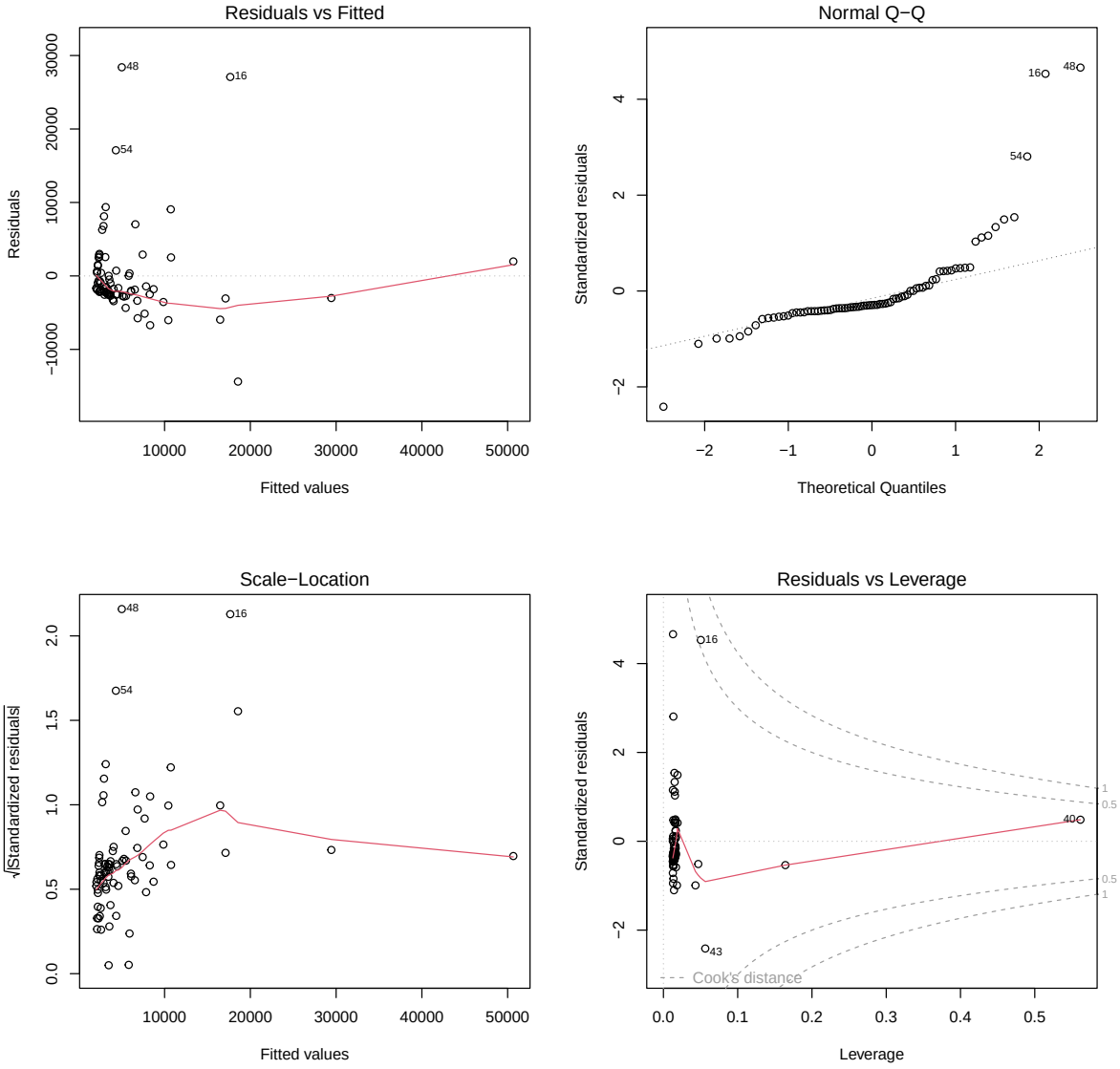
a)

```
plot(x = Sales, y = Assets)
lines(lowess(Assets, Sales))
```



From this plot we can say that it appears Sales and Assets do not have a linear relationship. The upward curving trend towards the high leverage point in the top right of the plot suggest a log transformation would be reasonable.

```
fit_1 <- lm(Assets ~ Sales)
par(mfrow = c(2,2))
plot(fit_1)
```



From the Residuals vs Fitted plot we can see that residuals do not appear to vary randomly

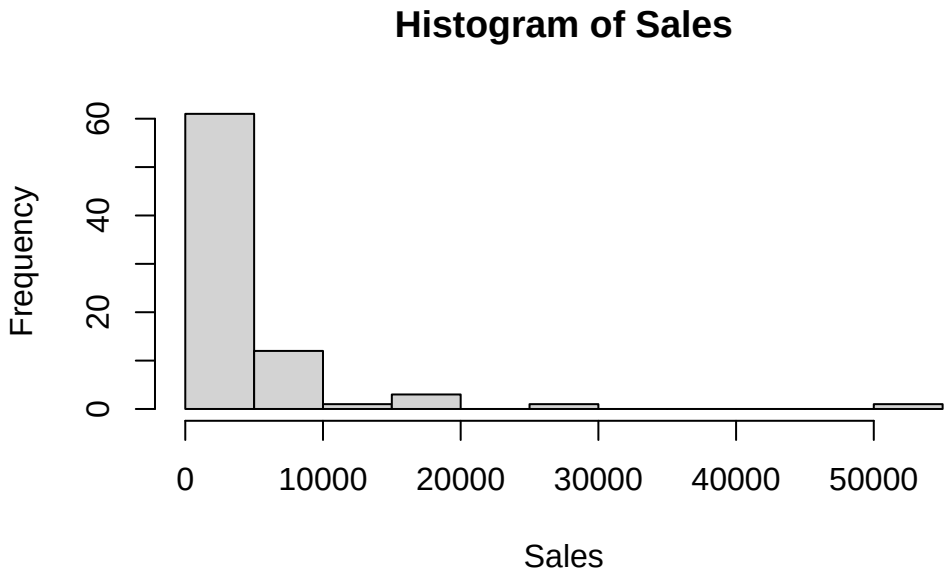
around zero. For fitted values 0 to 20,000 residuals are below zero on average.

In the normal Q-Q plot we can see the standardized residuals don't appear normally distributed. Particularly the right tail is much longer and fatter than what would be expected by the theoretical normal distribution.

In the $|\sqrt{\text{std. residuals}}|$ by Fitted values plot we can see evidence of non-constant variance. The trend line has an upward slope for Fitted values 0 to 20,000 where most observations are found, then it has a downward slope.

b)

```
hist(Sales)
```



From the histogram of Sales we can see that the distribution is highly right skewed and would be poorly fit by the normal distribution. We can use the box-cox method to transform Sales. That is the below transformation:

$$\Psi_S(X,\lambda) = \begin{cases} (X^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ log(X) & \text{if } \lambda = 0 \end{cases}$$

Then choosing between the below options for λ:

$$\lambda : \{-1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1\}$$

```
library(car)
```

```
Loading required package: carData
```

```
summary(powerTransform(Sales))
```

```
bcPower Transformation to Normality
      Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Sales   -0.0675           0      -0.2329       0.0979

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                             LRT df    pval
LR test, lambda = (0) 0.6481734  1 0.42077

Likelihood ratio test that no transformation is needed
                            LRT df       pval
LR test, lambda = (1) 170.5833  1 < 2.22e-16
```

The likelihood ratio test with H_0 : Sales is normally distributed, H_a : Sales is not normally distributed rejects the null hypothesis with a p value extremely close to zero. We need to transform Sales.

The estimated λ found by the box-cox transformation is -0.0675. From the likelihood ratio test with $H_0 : \lambda = 0, H_a : \lambda \neq 0$ we see that the null hypothesis cannot be rejected. Thus, we choose the log transformation of Sales.

```
log_sales <- log(Sales)
```

c)

```
fit_2 <- lm(Assets ~ log_sales)
summary(powerTransform(fit_2))
```

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.0166           0      -0.1688       0.1357

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
```

```
                              LRT df    pval
LR test, lambda = (0) 0.04564128  1 0.83083

Likelihood ratio test that no transformation is needed
                             LRT df       pval
LR test, lambda = (1) 159.6628  1 < 2.22e-16
```

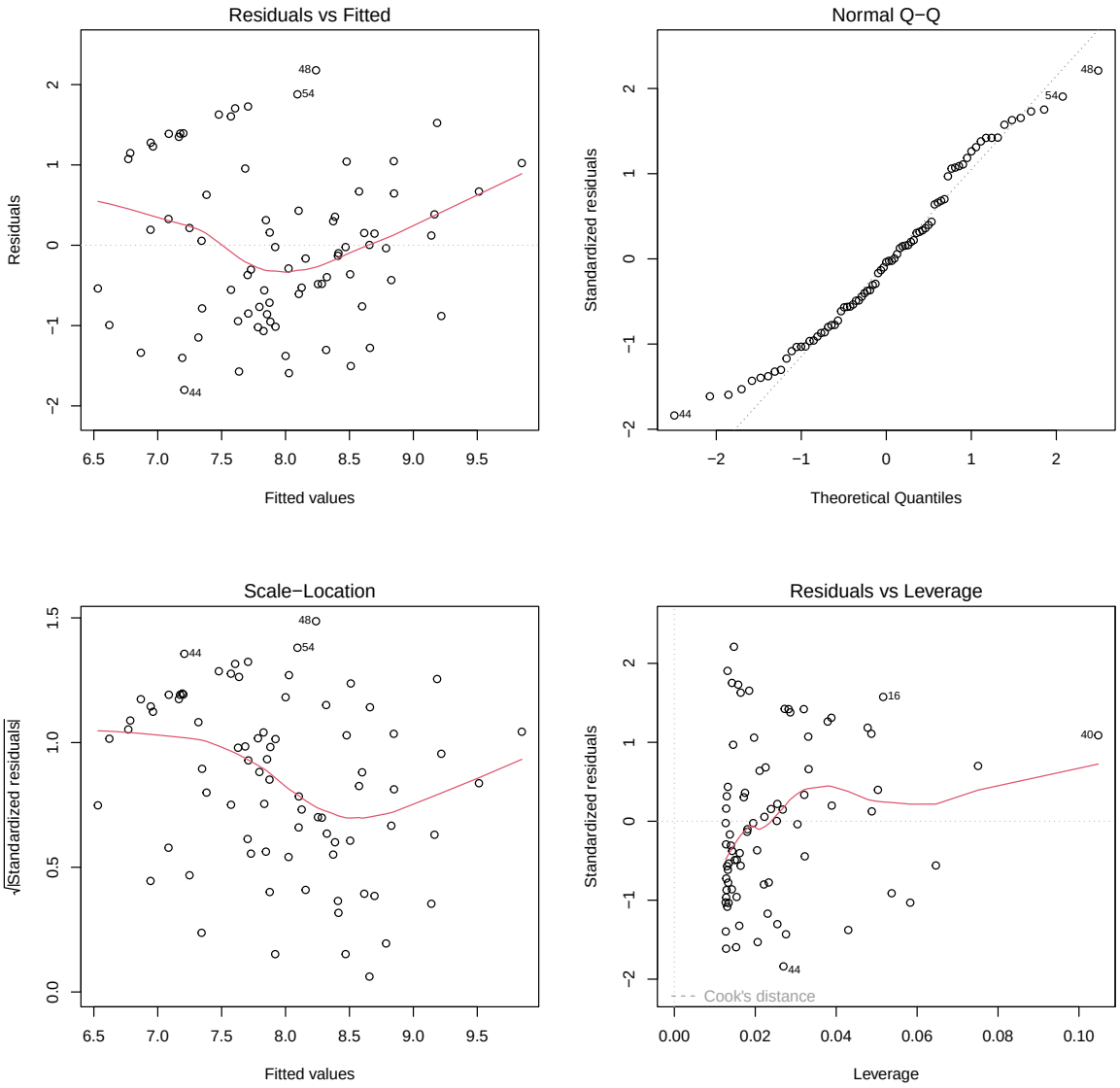
The likelihood ratio test with H_0 : Residuals are normally distributed, H_a : Resiudals are not normally distribut rejects the null hypothesis with a p value extremely close to zero. This is evidence that we need to transform Assets too.

The estimated λ found by the box-cox transformation is -0.0166. From the likelihood ratio test with $H_0 : \lambda = 0, H_a : \lambda \neq 0$ we see that the null hypothesis cannot be rejected. Thus, we choose the log transformation of Assets.

```r
log_assets <- log(Assets)
```

d)

```r
fit_3 <- lm(log_assets ~ log_sales)
par(mfrow = c(2,2))
plot(fit_3)
```



There are a few weaknesses in the model, first in the normal Q-Q plot of the standardized residuals there is deviation in the tails, both left and right are lighter than anticipated. There is also a trend in the scale-location plot, there is a sharp decrease in fitted values 7.5 to 8.5. This provides evidence of non-constant variance still being present.

e)

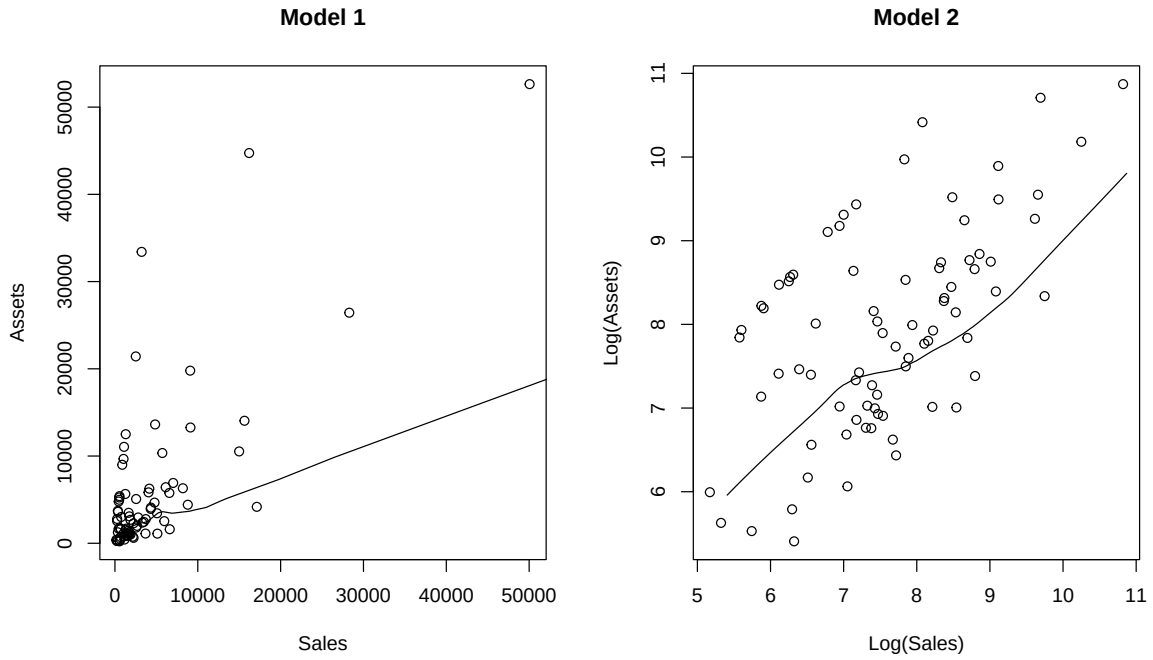
To compare the two models I will step through the assumptions we make and how reasonable they are.

Beginning with linear association in the explanatory and response variables.

```
par(mfrow = c(1,2))

plot(x = Sales, y = Assets, main = "Model 1")
lines(lowess(Assets, Sales))

plot(log_sales, log_assets, xlab = "Log(Sales)",ylab = "Log(Assets)", main = "Model 2")
lines(lowess(log_assets, log_sales))
```



In Model 1 we assume that Sales and Assets are linearly associated. In the left plot we clearly observe that this assumption is poorly met.

In Model 2 we assume that log(Sales) and log(Assets) are linearly associated. In the right plot we see that this assumption seems reasonable, a linear trend is evident.

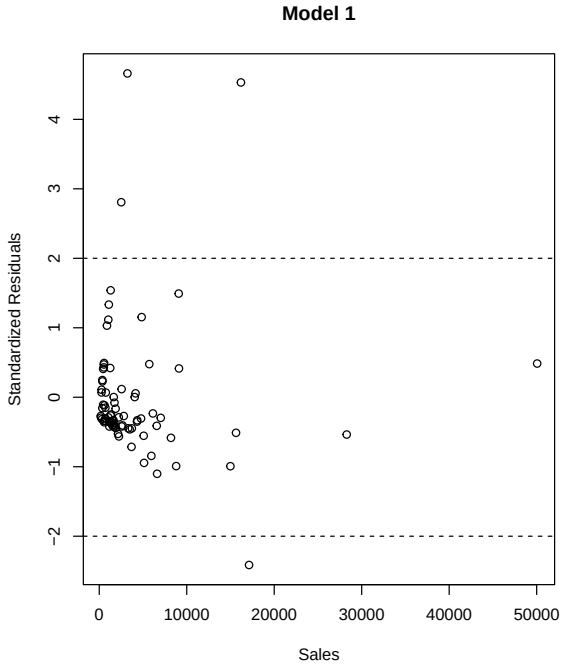
The assumption of independent errors won't be a differentiator for either model as we are using the same data.

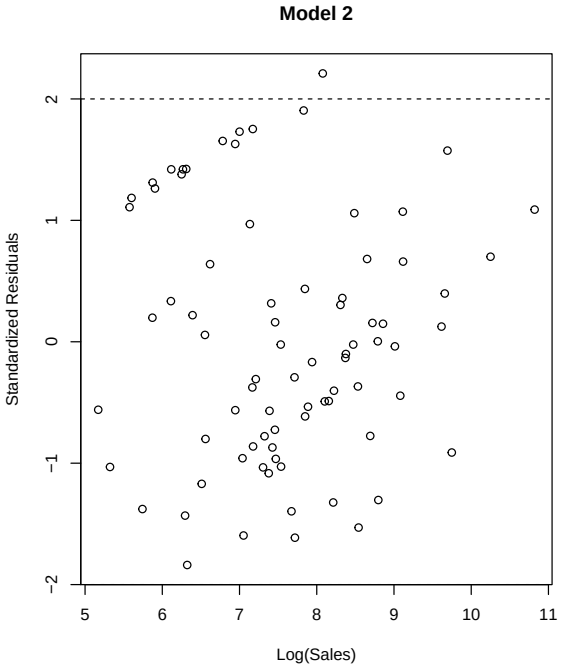
Next we look at the assumption of normality in errors.

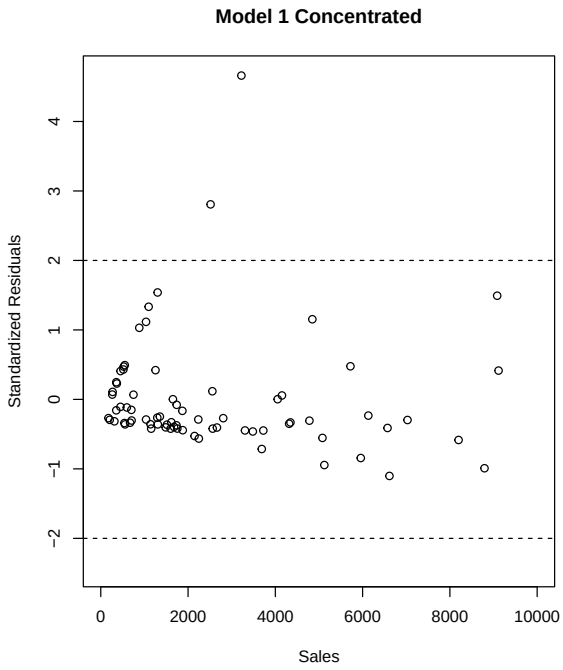
```r
par(mfrow = c(2,2))
StanRes1 <- rstandard(fit_1)
plot(x = Sales, y = StanRes1, xlab = "Sales", ylab = "Standardized Residuals", main= "Model 1
abline(h=2,lty=2)
abline(h=-2,lty=2)

StanRes2 <- rstandard(fit_3)
plot(x = log_sales, y = StanRes2, xlab = "Log(Sales)", ylab = "Standardized Residuals", main
abline(h=2,lty=2)
abline(h=-2,lty=2)

plot(x = Sales, y = StanRes1, xlab = "Sales", ylab = "Standardized Residuals", main= "Model 1
abline(h=2,lty=2)
abline(h=-2,lty=2)
```







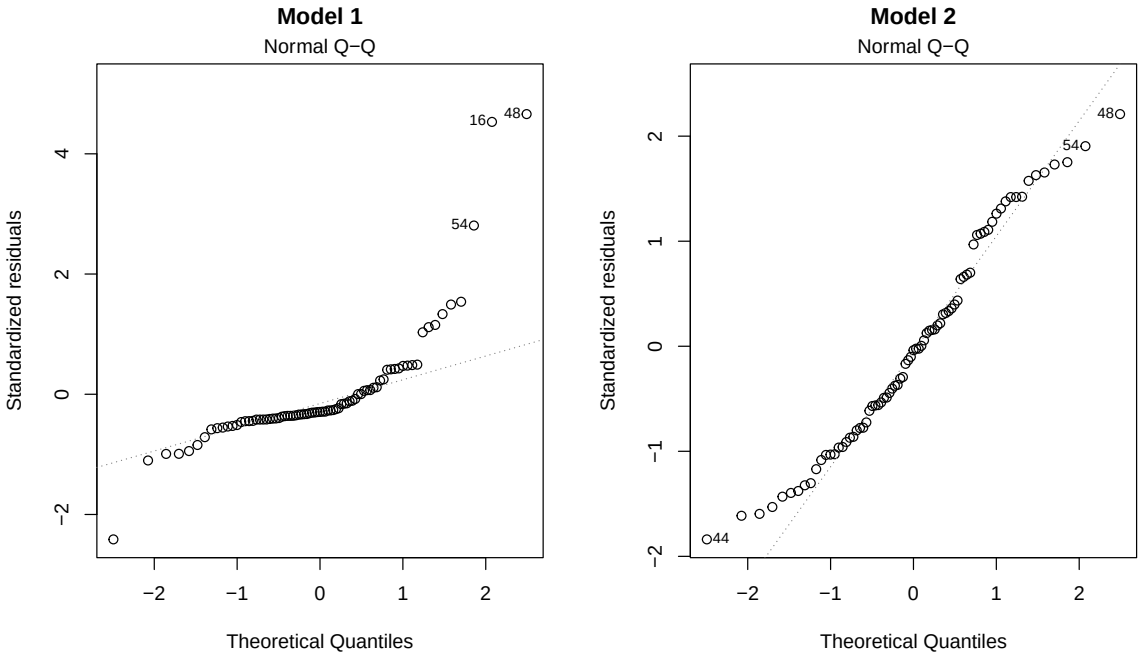
In Model 1 we can see that standardized residuals do not deviate randomly around zero how we would expect from a normal distribution. We can see this clearly if we take a closer look

at residuals for Sales 0 to 10,000. The majority of residuals in this range are beneath zero and particularly concentrated in the -.3 to -1 standardized residual range. That is not what we would expect in the normal distribution.

In Model 2 the Standardized Residuals randomly deviate around 0 as we would expect in a normal distribution.

Next we look at the respective QQ-plots:

```
par(mfrow = c(1,2))
plot(fit_1, which = 2, main = "Model 1")
plot(fit_3, which = 2, main = "Model 2")
```



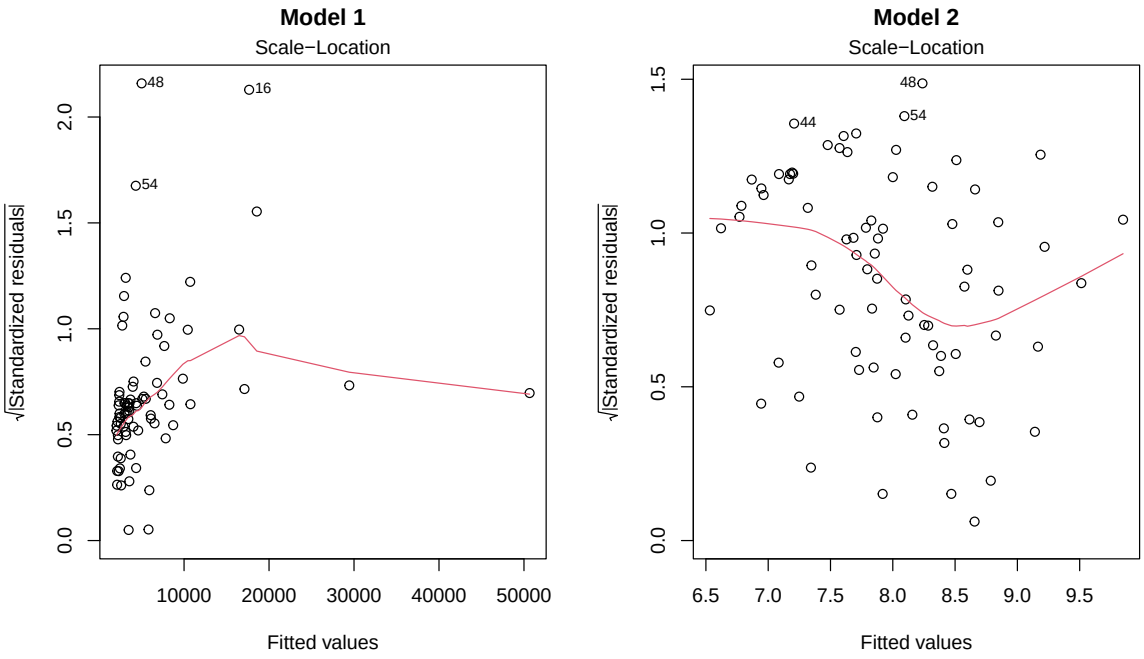
In Model 1 we see a far longer and fatter right tail than we would expect in the normal distribution.

In Model 2 we see slightly lighter tails than what we would expect in the normal distribution.

Overall the assumption of normality in errors is reasonable for Model 2 and not reasonable for Model 1.

The next assumption we check is equal variance.

```
par(mfrow = c(1,2))
plot(fit_1, which = 3, main = "Model 1")
plot(fit_3, which = 3, main = "Model 2")
```



Both models appear to have issues with this assumption as previously discussed.

Overall Model 2 is superior to Model 1, Model 2 meets the LINE assumptions well and the biggest difference between the two models is the assumption of Linear Association between the explanatory and response variables. Assets and Sales are not linearly associated which calls into question any conclusion taken away from Model 1. On the other hand log(Assets) and log(Sales) are linearly associated, so Model 2 is the more valid model.

f)

```
summary(fit_3)
```

```
Call:
lm(formula = log_assets ~ log_sales)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.80081 -0.77679 -0.03703  0.66919  2.17828

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4969     0.7181   4.870 5.83e-06 ***
log_sales     0.5870     0.0934   6.284 1.82e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.993 on 77 degrees of freedom
Multiple R-squared:  0.339, Adjusted R-squared:  0.3304
F-statistic: 39.49 on 1 and 77 DF,  p-value: 1.817e-08
```

The slope is interpreted as follows: As Sales increases by one percent we expect a .587 percent increase in Assets.

g)

#Come back to#

9)

With $E[Y] = \mu, Var(Y) = \mu^2$ we want to use a transformation that provides us with constant variance.

Using the Taylor expansion, $f(Y) = f(E[Y]) + f'(E[Y])(Y - E[Y]) + ...$, that allows us to estimate $Var(f(Y))$:

$$Var(f(Y)) = f'(E[Y])^2 Var(Y)$$

Subbing in $E[Y], Var(Y)$:

$$Var(f(Y)) = f'(\mu)^2 \mu^2$$

If we use $f(y) = log(y)$, which has derivative $f'(y) = \frac{1}{y}$ we get the following variance:

$$Var(f(Y)) = \frac{1}{\mu^2}\mu^2 = 1$$

Variance is now constant.