STAT 608 HW 3

Jack Cunningham (jgavc@tamu.edu)

9/23/24

1.

a)

This model cannot be used for inference or prediction. We make this conclusion by analyzing the standardized residual by distance plot. If a model is well fit we would expect to see the standardized errors randomly distributed around 0 and constant variance. For this model there is a clear quadratic pattern for the standardized residuals and two outlier points that, if valid, hint towards non-constant variance. These discrepancies call into question any conclusion made from this model.

b)

The ordinary straight line regression model does not fit the data well. To improve this model the practitioner should analyze the two outlier values, that appear to possibly be leverage points as well, and assess their validity. In the event these two points are valid a transformation should be considered to address the non-constant variance. Additionally a quadratic term should be considered to address the quadratic pattern in the standardized residuals.

2)

The general form of a confidence interval or prediction interval is a sample statistic ± margin of error. For both the sample statistic of $\hat{\beta}$ and the prediction interval's point estimate of $E[Y|X = x^\star]$ the term $\bar{y}$ is involved. We know that $\bar{y} = E[Y]$. The core reason why we can't just inverse the transformation of the end points is that generally $E[g^{-1}(Y)] \neq g^{-1}[E(Y)]$.

Here's a particular example. Imagine $g(X)$ is a decreasing function and we attempt to simply inverse the transformation of the end points of a prediction interval. Let X have a lower prediction bound of L and a upper prediction bound of U, so $U > L$. If we just took the inverse of bound sides of the bound we would have $(g^{-1}(L), g^{-1}(U))$. But consider that $g^{-1}(X)$ is a decreasing function, we would end up with $g^{-1}(L) > g^{-1}(U)$, that is certainly not a valid interval.

3)

In question 3 we have the design matrix X of:

$$X = \begin{bmatrix} 1_m & 0_m \\ 0_{n-m} & 1_{n-m} \end{bmatrix}$$

We use $H = X(X^TX)^{-1}X^T$:

$$(X^TX)^{-1} = \begin{bmatrix} \frac{1}{m} & 0 \\ 0 & \frac{1}{n-m} \end{bmatrix}$$

$$X(X^TX) = \begin{bmatrix} (\frac{1}{m})_m & 0_m \\ 0 & (\frac{1}{n-m})_{n-m} \end{bmatrix}$$

$$H = X(X^TX)^{-1}X^T = \begin{bmatrix} \frac{1}{m} & \frac{1}{m} & \frac{1}{m} & ... & \frac{1}{m} & 0 & 0 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{m} & \frac{1}{m} & \frac{1}{m} & ... & \frac{1}{m} & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 & \frac{1}{n-m} & \frac{1}{n-m} & \frac{1}{n-m} & \frac{1}{n-m} & \frac{1}{n-m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & ... & 0 & \frac{1}{n-m} & \frac{1}{n-m} & \frac{1}{n-m} & \frac{1}{n-m} & \frac{1}{n-m} \end{bmatrix}$$

The projection matrix projects the response variable y onto the column space of X. In doing this each prediction $\widehat{y_i}$ is the closest point on the column space of X to the response variable y. This makes sense because in least squares regression we want to minimize the sum of mean squared error $\hat{e}^2$, geometrically $\hat{e}_i$ is the vector between response variable y_i and $\widehat{y}_i$ which rests on the column space of X. Choosing the closest point by definition finds the shortest vector e_i, this minimizes $\hat{e}^2$.

4.

a)

We know $\hat{e} = y - \widehat{y}$ and that $\widehat{y} = Hy$. Where H is the projection matrix onto y. So:

$$\hat{e} = y - \widehat{y} = y - Hy = (I - H)y$$

b)

We know the below fact where A is a constant matrix and b is a vector of random variables.

$$Var(Ab) = AVar(b)A^T$$

So then, using that $(I - H)$ is a matrix and y is a vector of random variables with $Var(y) = \Sigma$.

$$Var[(I-H)y] = (I-H)Var(y)(I-H)^T = (I-H)\Sigma(I-H)^T$$

We use the fact that $\Sigma = \sigma^2 I$ to continue to simplify:

$$(I-H)\Sigma(I-H)^T = (I-H)\sigma^2 I(I-H)^T = \sigma^2(I-H)(I-H)^T = \sigma^2(II^T - IH^T - HI^T + HH^T)$$

Using the fact that H,I are symmetric we get.

$$\sigma^2(I - 2H + HH)$$

We can prove that $HH = H$, we use that $H = X(X^TX)^{-1}X^T$. Then:

$$HH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

So we get the final result for the covariance matrix of errors:

$$\sigma^2(I-H)$$

c)

We know that the entries in H are h_{ij} in each (i,j) position. And I is the identity matrix and thus diagonal, so its entries are$I_{ij} = 0, i \neq j$ and $I_{ij} = 1, i = j$. Then we can say the below:

$$Cov(\hat{e_i}, \hat{e_j}) = \sigma^2(I_{ij} - h_{ij}) = -h_{ij}\sigma^2, i \neq j$$

5.

a)

We know that $H = X(X^TX)^{-1}X^T$. Then using the rule $(AB)^T = B^TA$ and $((AB)^{-1})^T = ((AB)^T)^{-1}$we have:

$$H^T = (X(X^TX)^{-1}X^T)^T = X((X^TX)^T)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

b)

H is idempotent, $HH = H$. So:

$$HH = \begin{bmatrix} h_{11} & h_{12} & ... & h_{1n} \\ h_{21} & h_{22} & ... & h_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ h_{n1} & h_{n2} & ... & h_{nn} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & ... & h_{1n} \\ h_{21} & h_{22} & ... & h_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ h_{n1} & h_{n2} & ... & h_{nn} \end{bmatrix} = \begin{bmatrix} \sum_{i=j}^n h_{1j}h_{j1} & \sum_{j=1}^n h_{1j}h_{j2} & ... & \sum_{j=1}^n h_{1j}h_{jn} \\ \sum_{j=1}^n h_{2j}h_{j1} & \sum_{j=1}^n h_{2j}h_{j2} & ... & \sum_{j=1}^n h_{2j}h_{jn} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=1}^n h_{nj}h_{j1} & \sum_{j=1}^n h_{nj}h_{j2} & ... & \sum_{j=1}^n h_{nj}h_{jn} \end{bmatrix} = H$$

$H = H^T$ so $h_{ij} = h_{ji}$. So the diagonal elements of H are $h_{ii} = \sum_{j=1}^n h_{ij}^2$. We can also write this as $h_{ii} = h_{ii}^2 + \sum_{j=1, j\neq i}^n h_{ij}^2$. So we can conclude that $h_{ii} \geq h_{ii}^2 \geq 0$. Additionally the only time that a square of a number is less than or equal to the number itself is when it is less than 1. So therefore $0 \leq h_{ii} \leq 1$.

c)

We have $H = X(X^TX)^{-1}X^T$. Where X is the design matrix for simple linear regression.

So:

$$(X^TX)^{-1} = \frac{1}{SXX} \begin{bmatrix} \sum_{i=1}^n x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$X(X^TX)^{-1} = \frac{1}{SXX} \begin{bmatrix} \sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}x_1 & -\bar{x} + x_1 \\ \vdots & \vdots \\ \sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}x_n & -\bar{x} + x_n \end{bmatrix}$$

$$H = X(X^TX)^{-1}X^T = \frac{1}{SXX} \begin{bmatrix} \sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}x_1 + (-\bar{x} + x_1)x_1 & ... & \sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}x_1 + (-\bar{x} + x_1)x_n \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}x_n + (-\bar{x} + x_n)x_1 & \vdots & \sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}x_n + (-\bar{x} + x_n)x_n \end{bmatrix}$$

Then we can generalize an entry of H, h_{ij} as:

$$h_{ij} = \frac{1}{SXX}(\sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}x_i + (-\bar{x} + x_i)x_j)$$

We can create the desired $1/n$ term by adding and subtracting $n\bar{x}^2/n = \bar{x}^2$. We know that $\sum_{i=1}^n x_i^2 + n\bar{x}^2 = SXX$:

$$h_{ij} = \frac{1}{SXX}(\frac{1}{n}(\sum x_i^2 - n\bar{x}^2) + \bar{x}^2 - \bar{x}x_i - \bar{x}x_j + x_ix_j)$$

$$h_{ij} = \frac{1}{n} + \frac{\bar{x}^2 - \bar{x}x_i - \bar{x}x_j + x_i x_j}{SXX}$$

We can factor the numerator into $(x_i - \bar{x})(x_j - \bar{x})$. We get our desired result:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}$$

d)

We know that the covariance between two residuals is the below:

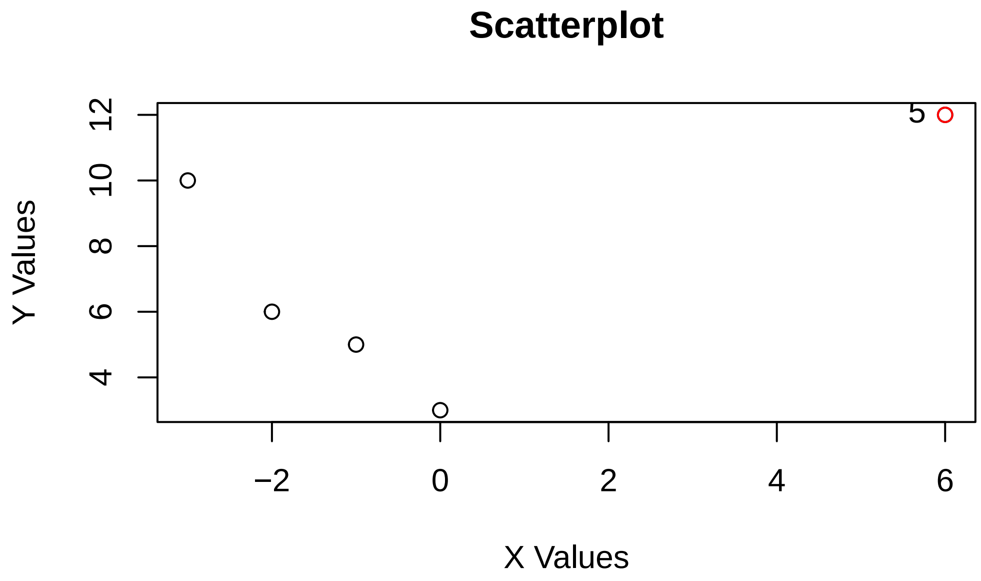
$$Cov(\hat{e_i}, \hat{e_j}) = -h_{ij}\sigma^2$$

From part c we have the formula for h_{ij}. Holding n fixed we can see that h_{ij} is small when and x_i and x_j are close to the mean of x, $\bar{x}$, so $Cov(\hat{e_i}, \hat{e_j}) \approx 0$ in those cases. This makes sense because we've previously seen that points of low leverage don't have a high impact on the regression line, so they will have low influence on the residuals for different x_i.

6)

a)

The fifth point appears to be a bad leverage point, so a leverage point and an outlier. The first point is a appears to be a leverage point but a good one, it generally fits the pattern set by the next three points.



b)

The predicted values $\hat{y}$ are $[5.7, 6.2, 6.7, 7.2, 10.2]^T$ so the residuals are:

$$\hat{e} = y - \hat{y} = \begin{bmatrix} 10 \\ 6 \\ 5 \\ 3 \\ 12 \end{bmatrix} - \begin{bmatrix} 5.7 \\ 6.2 \\ 6.7 \\ 7.2 \\ 10.2 \end{bmatrix} = \begin{bmatrix} 4.3 \\ -0.2 \\ -1.7 \\ -4.2 \\ 1.8 \end{bmatrix}$$

c)

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

```
x_bar <- mean(data$x)
SXX <- sum((data$x-x_bar)^2)
n <- length(data$x)
data$leverage <- 1/n + (data$x - x_bar)^2/SXX
data$leverage
```

```
[1] 0.38 0.28 0.22 0.20 0.92
```

The leverage for each point is:

$$h_{ii} = \big[.38, .28, .22, .20, .92\big]$$

The rule for a leverage point is $h_{ii} > \frac{4}{n}$. With $n = 5$ the rule is $h_{ii} > .8$. So the only leverage point is point 5, $(6, 10.2)$. This would be considered a bad leverage point as it doesn't fit the downward sloping pattern of the rest of the data.

d)

The variance of each residual is $Var(\hat{e_i}) = \sigma^2(1 - h_{ii})$ as previously worked through in Question 4. We need to estimate σ^2 with $\hat{\sigma}^2 = MS_{\text{Error}} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e_i}^2$.

```
data$squared_residuals <- data$residuals^2
squared_sigma_est <- sum(data$squared_residuals)/(n - 2)
data$residuals_variance <- squared_sigma_est*(1-data$leverage)
data$residuals_variance
```

```
[1]  8.742 10.152 10.998 11.280  1.128
```

e)

The formula for standardized residuals is below:

$$r_i = \frac{\hat{e_i}}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

```
data$std_residuals <- data$residuals/(sqrt(squared_sigma_est*(1 - data$leverage)))
data$std_residuals
```

```
[1]  1.4543303 -0.0627703 -0.5126159 -1.2505318  1.6947980
```

The final point $(6, 10.2)$ has the largest standardized residual. This appears to conflict with part b because in part b this same point had a relatively small residual compared to the other 4 points.

f)

The highest leverage point has the smallest variance due to the relationship between $Var(\hat{e_i})$ and h_{ii}.

$$Var(\hat{e_i}) = \sigma^2(1 - h_{ii})$$

As h_{ii} increases $Var(\hat{e}_i)$ decreases.

7.

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 $Var(f(Y))$ as below:

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

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

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

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

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

Variance is now constant.

8.

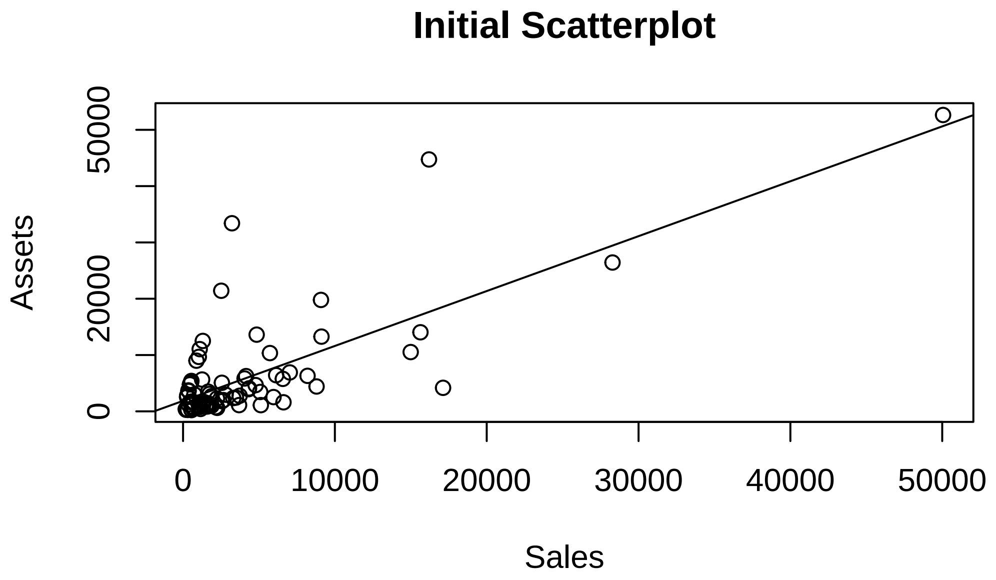
```
company <- read.csv("company.csv")
head(company, 5)
```

```
                    Company Assets Sales Market_Value Profits Cash_Flow
1              Air Products   2687  1870         1890   145.7     352.2
2             Allied Signal  13271  9115         8190  -279.0      83.0
3    American Electric Power  13621  4848         4572   485.0     898.9
4 American Savings Bank FSB   3614   367           90    14.1      24.6
5                       AMR   6425  6131         2448   345.8     682.5
  Employees         sector
1      18.2          Other
2     143.8          Other
3      23.4         Energy
4       1.1        Finance
5      49.5 Transportation
```

```
attach(company)
```

a)

```
plot(x = Sales, y = Assets, main = "Initial Scatterplot")
abline(lsfit(Sales, Assets))
```

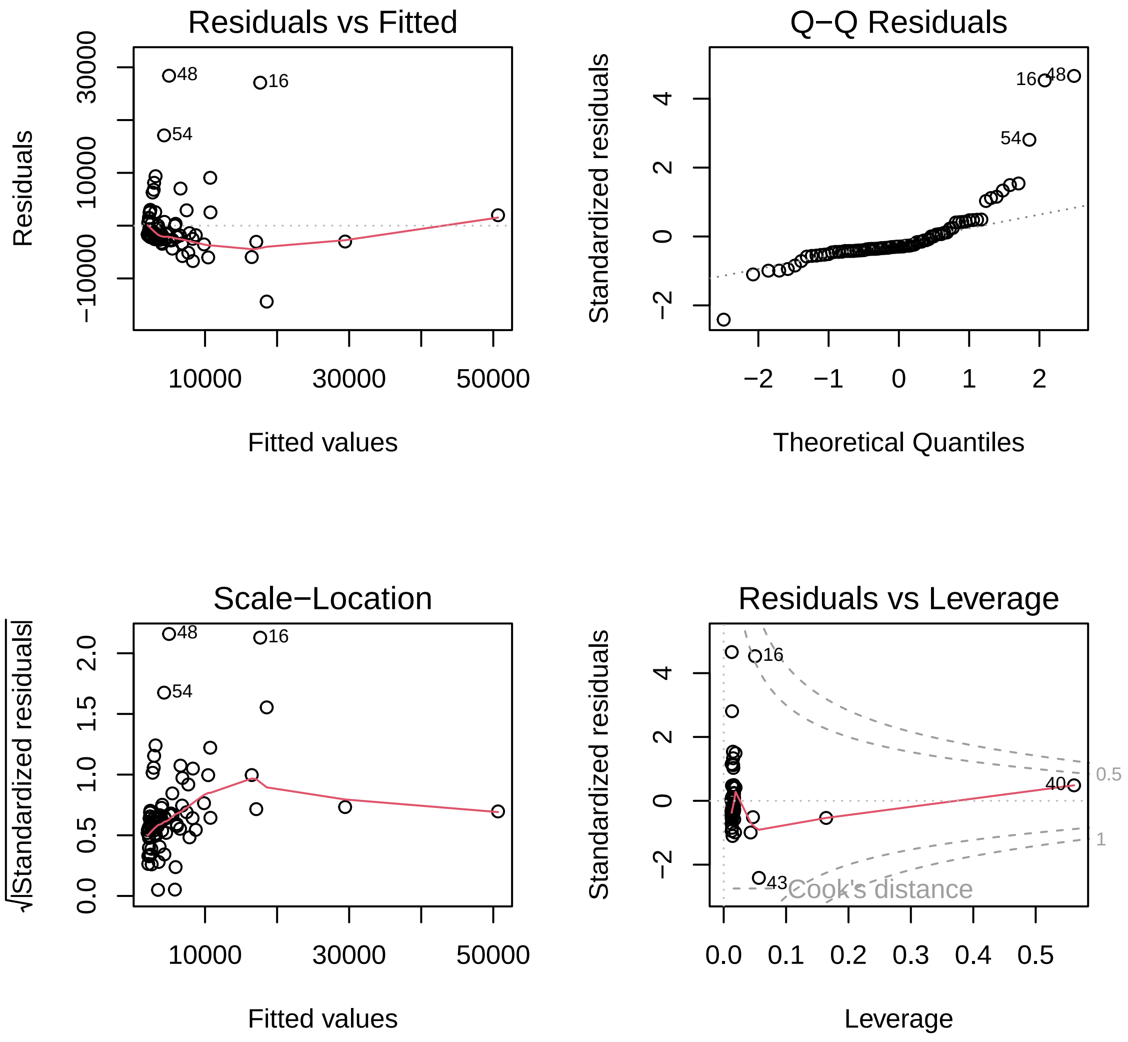


```
first_fit <- lm(Assets ~ Sales)
summary(first_fit)
```

```
Call:
lm(formula = Assets ~ Sales)

Residuals:
     Min       1Q   Median       3Q      Max 
-14382.5  -2559.0  -1789.2    685.2  28397.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.867e+03  8.045e+02   2.321   0.0229 *  
```

```
Sales       9.748e-01  9.903e-02   9.844 2.87e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6133 on 77 degrees of freedom
Multiple R-squared:  0.5572,    Adjusted R-squared:  0.5515
F-statistic:  96.9 on 1 and 77 DF,  p-value: 2.875e-15
```

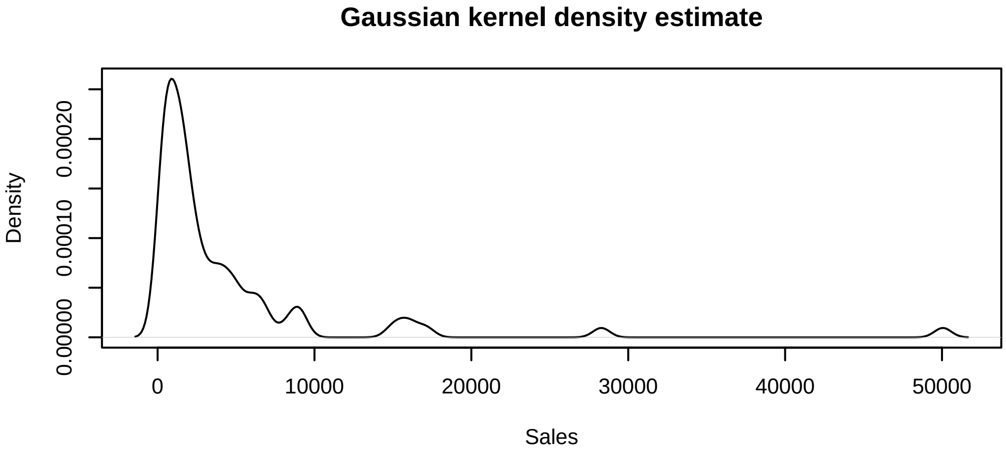
```
par(mfrow = c(2,2))
plot(first_fit)
```

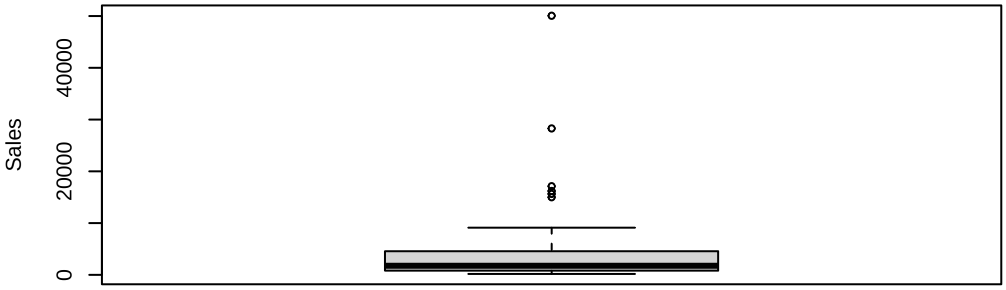


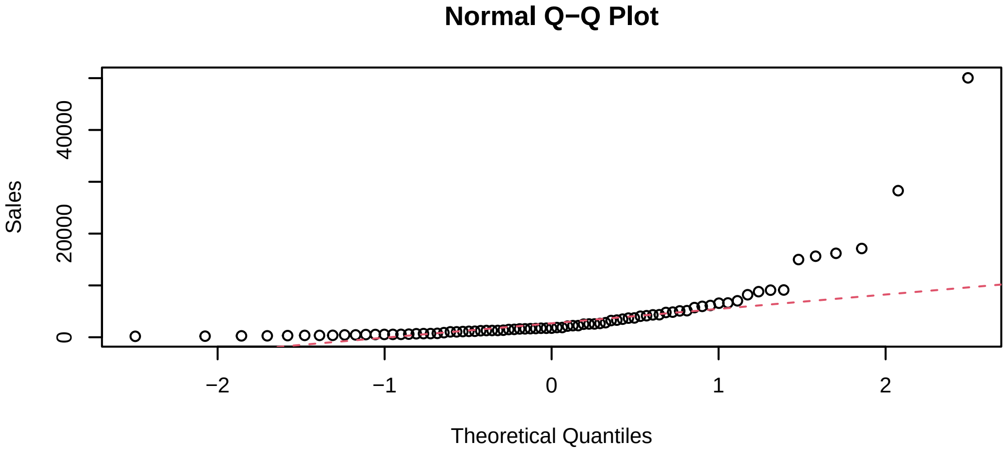
Normality in the errors is being violated. We can see a long and fat tail in the QQ plot of Standardized Residuals by Theoretical Quantiles. In the Residuals vs Fitted plot we can a pattern where many of the smaller fitted values have negative residuals, indicating the linear fit isn't working very well. There is evidence of non-stationary variance as in the densely populated area of the scale-location plot the trend line increases.

b)

```r
par(mfrow=c(3,1))
plot(density(Sales,bw="SJ",kern="gaussian"),type="l",
main="Gaussian kernel density estimate",xlab="Sales")
boxplot(Sales,ylab="Sales")
qqnorm(Sales, ylab = "Sales")
qqline(Sales, lty = 2, col=2)
```


Gaussian kernel density estimate
Density
0.00000
0.00010
0.00020
0
10000
20000
30000
40000
50000
Sales

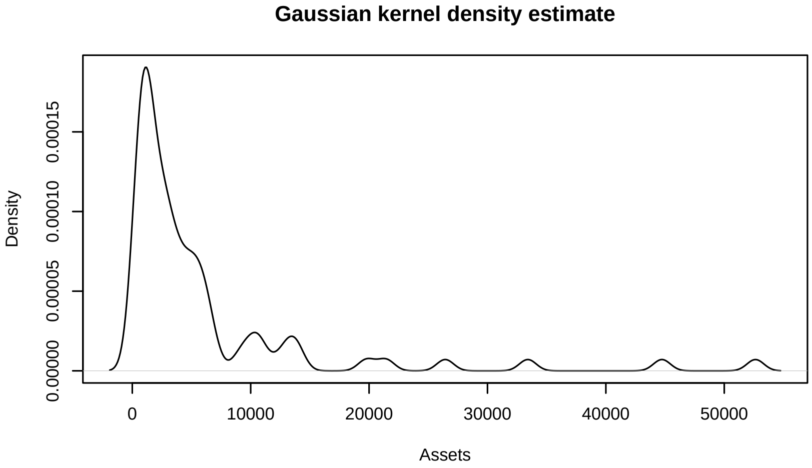

Sales
0
20000
40000

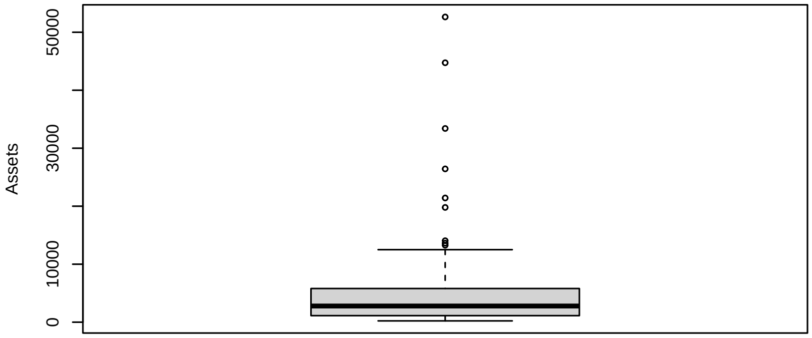

Normal Q−Q Plot
Sales
0
20000
40000
−2
−1
0
1
2
Theoretical Quantiles

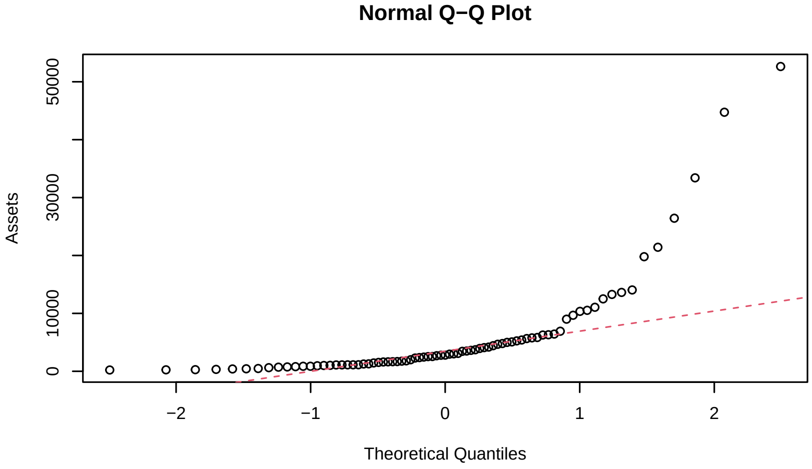
From these three plots we can see the right skew in values of Sales The normal QQ-Plot shows that the distribution of Sales isn't close to normal. We will now attempt to transform Sales so g(Sales) can approximate a normal distribution.

Before we decide how to go about this, let's take a look at the distribution of Assets. Since typically financial datasets contain many skewed variables, if Assets is skewed we will be able to plan our approach.

```r
par(mfrow=c(3,1))
plot(density(Assets,bw="SJ",kern="gaussian"),type="l",
main="Gaussian kernel density estimate",xlab="Assets")
boxplot(Assets,ylab="Assets")
qqnorm(Assets, ylab = "Assets")
qqline(Assets, lty = 2, col=2)
```


Gaussian kernel density estimate
Density
0.00000
0.00005
0.00010
0.00015
0
10000
20000
30000
40000
50000
Assets


Assets
0
10000
30000
50000


Normal Q−Q Plot
Assets
0
10000
30000
50000
−2
−1
0
1
2
Theoretical Quantiles

Assets is also skewed. With this in mind we will take approach 1 from the textbook on page 94. First we will transform X so the transformed version of X is as normal as possible. To do this we will use the Box-Cox transformation. Then we will fit $Y = g(\beta_0 + \beta_1 \Psi(x, \lambda_x) + e)$ and use an inverse response plot to find the transformation g^{-1} for Y.

So we begin with the Box-Cox transformation of X. So we are using the following family:

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

We will choose λ that minimizes the residual sum of squares from fitting $E[Y|X] = \alpha_0 + \alpha_1 \Psi(x, \lambda)$.

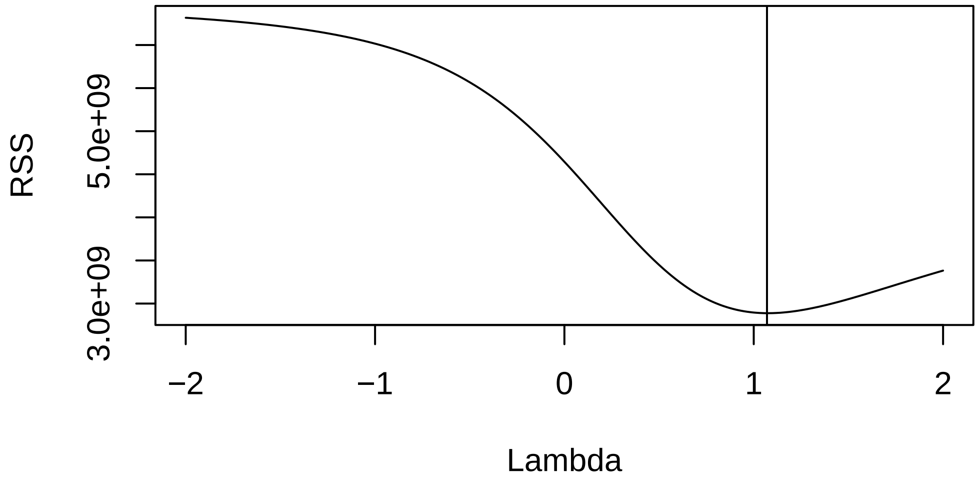
```
RSS <- rep(0,400)
lambda <- seq(-2,2,.01)
for(i in 1:length(lambda)){
  if(lambda[i]!=0){
    psi <- (Sales^lambda[i]-1)/lambda[i]
  }
  else{
    psi <- log(Sales)
  }
  fit <- lm(Assets~psi)
  RSS[i] <- sum(fit$residuals^2)
}
lambda_chosen <- lambda[which.min(RSS)]
lambda_chosen
```

```
[1] 1.07
```

```
psi_chosen <- (Sales^lambda_chosen-1)/lambda[i]
```

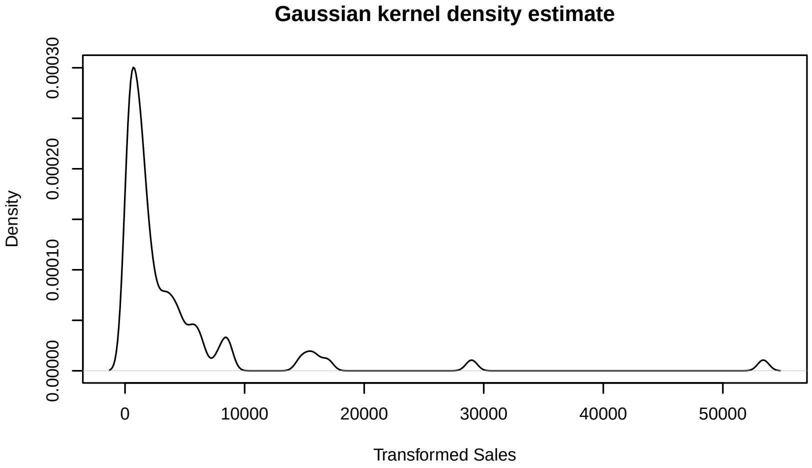
The λ that minimizes residual sum of squares is 1.07. A plot of RSS by λ is below:

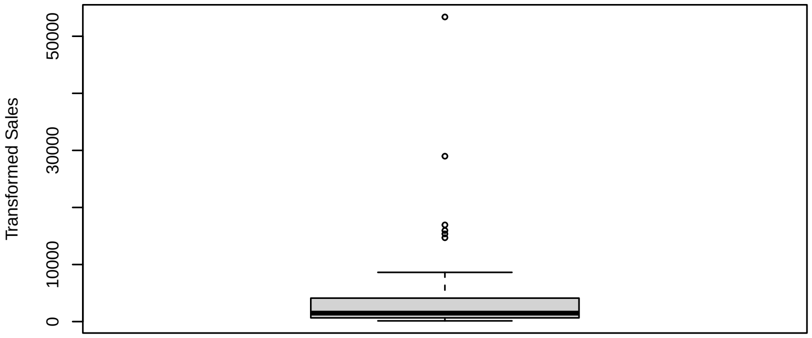
```
plot(x = lambda, y = RSS, type = "l", xlab = "Lambda", ylab = "RSS")
abline(v = lambda_chosen)
```

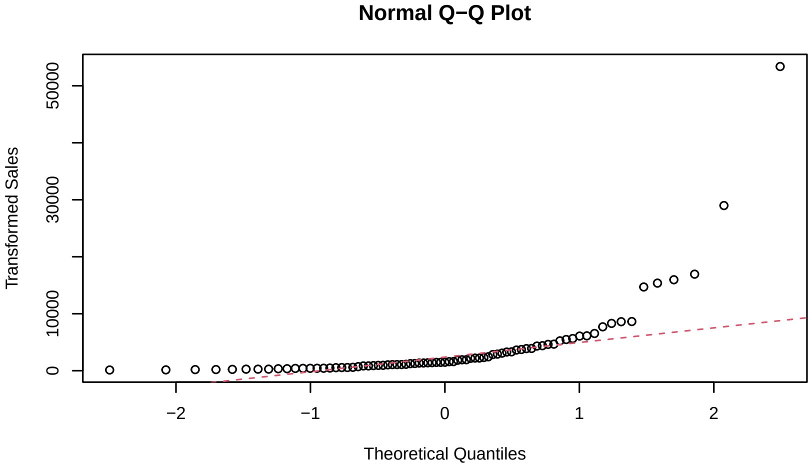


We can see how normal the transformed X has become.

```
par(mfrow=c(3,1))
plot(density(psi_chosen,bw="SJ",kern="gaussian"),type="l",
main="Gaussian kernel density estimate",xlab="Transformed Sales")
boxplot(psi_chosen,ylab="Transformed Sales")
qqnorm(psi_chosen, ylab = "Transformed Sales")
qqline(psi_chosen, lty = 2, col=2)
```


Gaussian kernel density estimate
Density
0.00000
0.00010
0.00020
0.00030
0
10000
20000
30000
40000
50000
Transformed Sales


Transformed Sales
0
10000
30000
50000


Normal Q−Q Plot
Transformed Sales
0
10000
30000
50000
−2
−1
0
1
2
Theoretical Quantiles

The transformed version of Sales is closer to normal compared to before, but still not great.

c)

We continue the procedure we laid out in part b. First we fit the regression model with our transformed X.

```r
library(car)
```

```
Warning: package 'car' was built under R version 4.3.3
```

```
Loading required package: carData
```

```
Warning: package 'carData' was built under R version 4.3.3
```

```r
fit_transformed <- lm(Assets~psi_chosen)
power_transform <- powerTransform(fit_transformed)
power_transform
```

```
Estimated transformation parameter
        Y1
0.09050312
```

```r
y_lambda <- as.numeric(powerTransform(fit_transformed)$lambda)
```

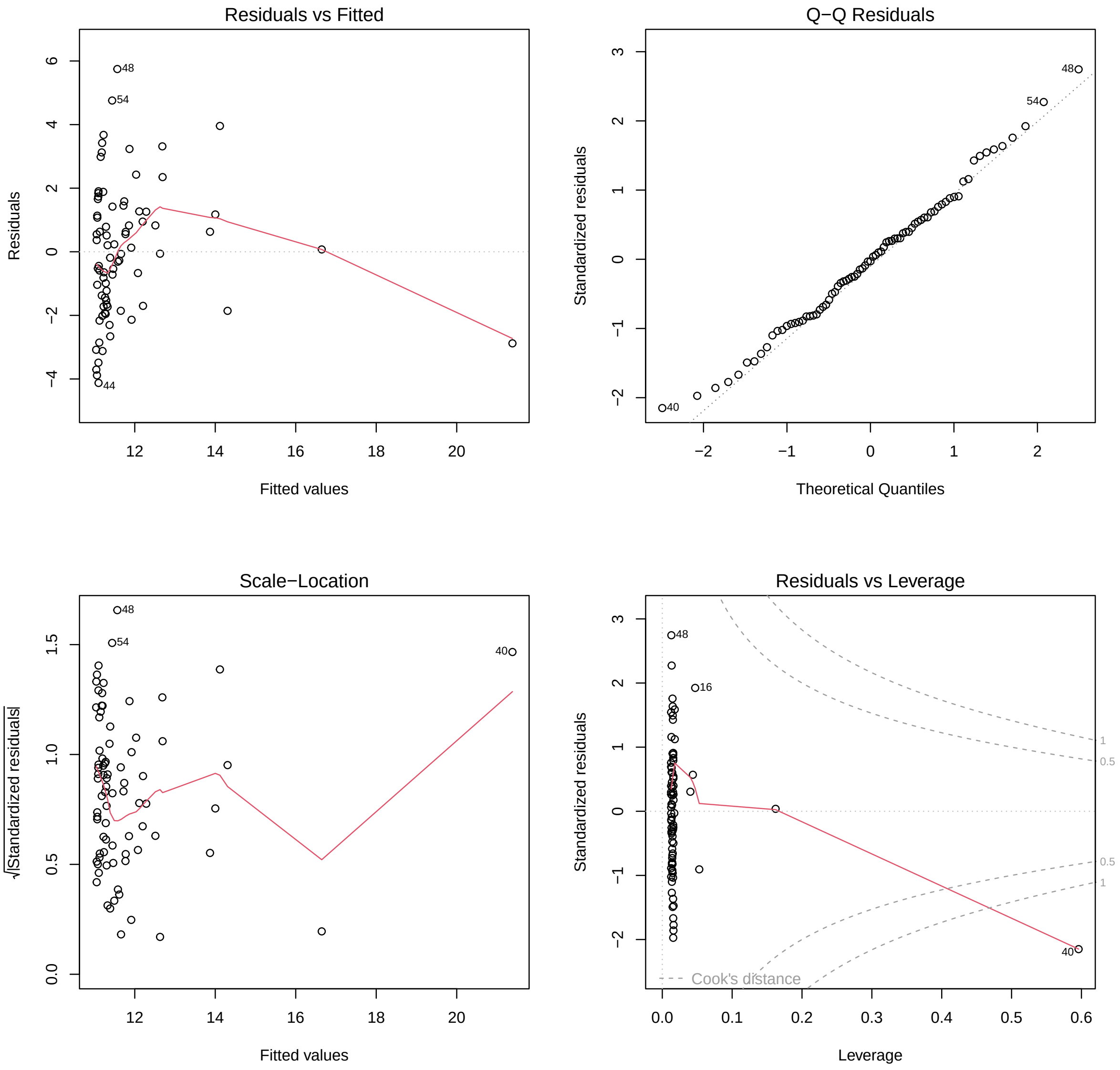
Using powerTransform we selected the highest log likelihood value corresponding with λ. The selected λ for Assets is 0.09050312.

```r
Assets_transformed <- (Assets^y_lambda-1)/y_lambda
```

d)

```r
Sales_transformed <- psi_chosen
fully_transformed_fit <- lm(Assets_transformed~Sales_transformed)
```

```r
par(mfrow = c(2,2))
plot(fully_transformed_fit)
```



The most striking difference is how much closer the residuals are to a normal distribution. The QQ plot shows the the points fitting the normal QQ line a lot better. The Residuals vs Fitted plot shows the points more evenly distributed around 0. In the residuals vs Leverage plot we can again that the points are more evenly dispersed along the -2 to 2 deviations. The outlier values are also now only at about 2.5 standard deviations away, significantly better than the 4 standard deviations away in model 1. The one weakness is the the bad leverage point indexed at 40. That would need to be examined further.

e)

Model 2 is superior. Model 1 broke the assumption of normally distributed errors but model 2's residuals are rather close to being normally distributed as discussed in part d. However model 2 does have a clear weakness. Point 40 is a very influential point as it is an outlier and a leverage point, it should be examined before we perform any inference or prediction.

f)

```
log_sales <- log(Sales)
log_assets <- log(Assets)
log_fit <- lm(log_assets~log_sales)
summary(log_fit)
```

```
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 β_1 is 0.587. Our interpretation is that if Sales increases by 1% we expect Assets to increase by .587%.

g)

The 95% confidence interval of the regression line given $6,571$ million sales is the below, where we transform the endpoints and add the correction factor.

$$g^{-1}(\hat{y}^\star + \hat{\sigma}^2/2) \pm g^{-1}(t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{SXX}})$$

Where $\hat{\sigma}^2 = MS_{Error} = \sum_{i=1}^n \hat{e}^2/n - 2$ and $g^{-1}(x) = e^x$.

```
sigma_hat_sq <- sum(log_fit$residuals^2)/(length(log_sales)-2)
sigma_hat_sq
```

```
[1] 0.9860089
```

```
sigma_hat <- sqrt(sigma_hat_sq)
n <- length(log_sales)
t_stat <- qt(1-0.05/2,n - 2)
x_bar <- mean(log_sales)
x_star <- 6571
SXX <- sum((log_sales - x_bar)^2)
y_hat <- predict(log_fit,data.frame(log_sales = c(log(6571))))
```

```
lower <- exp(y_hat + sigma_hat_sq/2) - exp(t_stat*sigma_hat*sqrt(1/n + (log(x_star)-x_bar)
upper <- exp(y_hat + sigma_hat_sq/2) + exp(t_stat*sigma_hat*sqrt(1/n + (log(x_star)-x_bar)
```

The confidence interval $E[Y|X = 6571]$ is:

(9408.61231, 9411.35174)

For companies with sales of 6571 million we are 95% confident the mean of their assets are between 9408.61231 and 9411.35174.