**STAT332: Generalised Linear Models**

**Autumn 2023**

**Assignment 3**

**Ruben Traicevski 6790021**

**Q1**

**a)**

Ridge Regression estimator is given by the following,

$$\widehat{\beta_{ridge}} = argmin_{\beta}\{||Y - X\beta||_2^2 + \lambda||\beta||\}$$

An expression for $\widehat{\beta_{ridge}}$ , taken from Week 5 Lecture Notes, is given by,

$$\widehat{\beta_{ridge}} = (X^TX + \lambda I)^{-1}X^TY$$

To obtain this result, we take an expression for $RSS_\lambda$ and derive the minima,

$$RSS_\lambda = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

$$= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta + \lambda\beta^T\beta, \qquad as\ transpose\ respects\ addition$$

Then we will take the partial derivative w.r.t $\beta$ and set to zero

$$\frac{\partial}{\partial\beta}RSS_\lambda = \frac{\partial}{\partial\beta}Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta + \lambda\beta^T\beta$$

$$= -2X^TY + 2X^TX\beta + 2\lambda\beta$$

Setting this to zero and rearranging for $\beta$,

$$0 = -2X^TY + 2X^TX\beta + 2\lambda\beta$$

$$X^TX\beta + \lambda\beta = X^TY$$

$$(X^TX + \lambda I)\beta = X^TY, \qquad by\ the\ Identity\ Matrix, I\beta = \beta$$

$$\widehat{\beta_{ridge}} = (X^TX + \lambda I)^{-1}X^TY$$

$$For\ \widehat{\beta_{ridge}}\ to\ be\ unbiased, we\ require\ that,$$

$$E\left[\hat{\beta}\right] - \beta = 0$$

$$E\left[\hat{\beta}\right] - \beta = E[(X^TX + \lambda I)^{-1}X^TY - \beta$$

$$= (X^TX + \lambda I)^{-1}X^TE[Y] - \beta$$

$$= (X^TX + \lambda I)^{-1}X^TX\beta - \beta, \qquad as\ E[Y] = X\beta$$

$$= (X^TX + \lambda I)^{-1}(X^TX - (X^TX + \lambda I))\beta$$

$$= -\lambda(X^TX + \lambda I)^{-1}\beta$$

Thus, we see that $E[\hat{\beta}] - \beta \neq 0$ for any $\lambda > 0$

This means for it to be unbiased; we would require $\lambda = 0$

b)

For Generalised Least Squares, we begin with a standard linear regression model and assumptions.

$$1. Y = X\beta + \varepsilon$$

$$2. E[\varepsilon \mid X] = 0, \qquad Cov[\varepsilon \mid X] = \Sigma$$

$\Sigma$ is given to be known and nonsingular covariance matrix

$$X\ must\ be\ full\ column\ rank\ so\ that\ the\ columns\ are\ linearly\ independent,$$

$$to\ ensure\ X\ is\ invertible$$

For an AR(1) model we define that

$$e_i \sim N(pe_{i-1}, \sigma^2), \qquad i > 1$$

$$e_1 \sim N(0, \sigma^2)$$

$$\rightarrow e \sim N(0, \sigma^2\Sigma), \qquad errors\ become\ correlated$$

Thus, we now have that.

$$e_t = pe_{t-1} + \varepsilon_t$$

$$\rightarrow E[e_t] = pE[e_{t-1}] + E[\varepsilon_t]$$

$$V[e_t] = E[e_t^2] - \mu^2 = \frac{\sigma_\varepsilon^2}{1 - p^2}$$

$$1 - p^2 > 0, \qquad |p| < 1\ for\ a\ finite\ variance$$

c)

$$Given\ that$$

$$V[e] = \sigma^2 LL^T$$

Then by GLS, we can equate,

$$V[e] = \sigma^2 LL^T = V[e] = \sigma^2\Sigma, \qquad \rightarrow\ LL^T = \Sigma$$

We are also given that,

$$z = L^{-1}y, \qquad M = L^{-1}X$$

This will be used to define a new system,

$$Let\ z = M\beta + d$$

Thus,

$$\hat{\beta} = (M^T M)^{-1} M^T z = (X^T (L^{-1})^T L^{-1} X)^{-1} (X^T (L^{-1})^T L^{-1} Y)$$

$$By\ the\ Tranpose, \qquad (A^T)^{-1} = (A^{-1})^T, \qquad assuming\ A\ is\ invertible$$

$$By\ symmetry\ of\ Transpose, \qquad (AA^T)^T = AA^T$$

$$\rightarrow \quad (L^{-1})^T L^{-1} = (LL^T)^{-1} = \Sigma^{-1}$$

Therefore, we see that,

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} (X^T \Sigma^{-1} Y) = \hat{\beta}_{GLS}$$

d)

For an AR(1) model,

$$Var[e] = \sigma^2 \Sigma$$

$\Sigma$ is an n x n positive definite matrix that is known

This means we will take,

$$Var[Y] = \sigma^2 \Sigma$$

Thus,

$$Var[\widehat{\beta_{OLS}}] = Var[(X^T X)^{-1} X^T Y], \qquad By\ definition\ of\ the\ OLS\ estimator$$

$$= (X^T X)^{-1} X^T Var[Y] X (X^T X)^{-1}$$

$$\sigma^2 (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$$

Q2

a)

| $\beta$ | Estimate | Pr $(>\|t\|)$ |
|---|---|---|
| $\beta_0$ | 25.4929 | 0.001259 < 0.05 |
| $\beta_1$ | 0.2269 | 0.388631 > 0.05 |
| $\beta_2$ | -1.4585 | 0.000178 < 0.05 |

| $F-statistic$ | $R-squared$ | $p-value$ |
|---|---|---|
| 26.32 | 0.8143 ~ 81% | 4.096e-05 << 0.05 |

- The estimated coefficients indicate Capital has a positively correlated relation to Real Value Added while Labour has a negatively correlated relation to Real Value Added. Both $\beta_2$, and $\beta_0$ have p-value < 0.05 indicating Labour has a statistically significant relationship while $\beta_1$, has a p-value > 0.05 indicating it does not have a statistically significant relationship.
- The F-statistic is 26.32 > critical value of 3.9 for a 0.05 significance level, thus, rejecting the null hypothesis which holds that our all our coefficient estimates are equal to zero. This is reinforced by a p – value of 4.096e-05 << 0.05, meaning we can confidently assume that at least one independent variable is correlated with the dependent variable.
- The R-Squared value is ~ 81% which is very large, meaning the model accounts for 81% of the variation of the dependent variable caused by the independent variables.

b)

| Shapiro Wilk Test | Durbin Watson Test |
|---|---|
| p-value = 0.3766 > 0.05 | p-value = 0.01549 > 0.05 |
| W = 0.93954 | DW = 1.3049 |

- The Shapiro Wilk Test indicates that the null hypothesis which states residuals are normally distributed is not rejected at the 0.05 significance level.
- The Durbin Watson Test for one step correlation indicates that the null hypothesis which states that the residuals are uncorrelated is not rejected at the 0.05 significance level.
- E[residuals] = -7.979e-17 ~ 0

Then we can further justify the model by presenting the plots below

- The Residual against Fitted is showing the observations are spread randomly about the horizontal axis. We can conclude that the residuals are i.i.d and have a constant variance.
- Residuals against Leverage is indicating some collinearity which can be seen also in the ACF plots for the covariates. This isn't of much concern as there isn't perfect multicollinearity and linear regression being a robust method.
- The Normal Q-Q plot is showing that the residuals are normally distributed. I.e, Taking a log transformation makes it lognormally distributed, meaning normally distributed residuals is preserved.
- The ACF plots of both covariates is indicating some multicollinearity at lower levels of lag, however they are not greatly outside the 0.05 band indicating it can be disregarded.

We can conclude that the log transformation model is an adequate model. Possible improvements include perhaps eliminating the covariate 'Capital' considering the regression model was indicating to not statistically significant with the dependent variable. This could also be done in conjunction with taking a square root transformation of the data as intuitively this will inflate smaller residuals at 'small' X values away from the fitted line while pushing large residuals at 'large' X values towards the fitted line, as the Residuals vs Fitted plot shows some residuals far from the fitted line while some are relatively much closer

```
## Create Data frame with data from excel file

obsDF <- data.frame(Y=Q2$`Real Value Added`,X1=Q2$Capital,X2=Q2$Labour)

## Apply log transformation over the model to fit regression model

fit <- lm(log(Y) ~ log(X1) + log(X2), data = obsDF)

summary(fit)

plot(fit)

res = residuals(fit)

shapiro.test(res)

autocorrelation <- acf(Q2$Capital, lag.max=40, plot=TRUE)

autocorrelation <- acf(Q2$Labour, lag.max=40, plot=TRUE)
```

install.packages("lmtest")

library(lmtest)

dwtest(formula = fit)

sum(res)

c)

## 95% Confidence Interval

confint(fit)

| | | |
|---|---|---|
| $\beta_1$ | -0.3256987 | 0.7794063 |
| $\beta_2$ | -2.0541609 | -0.8627954 |

d)

capital = 300000

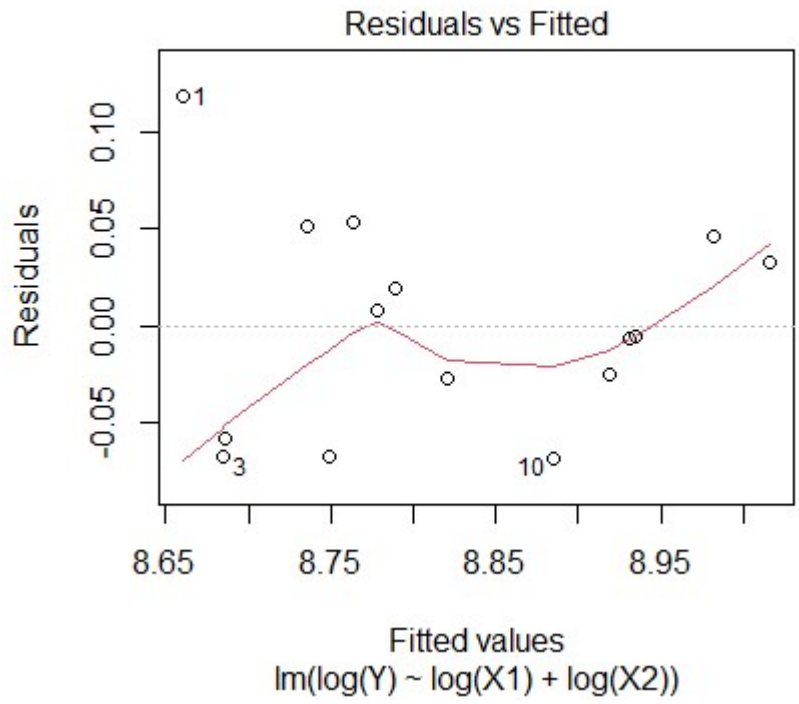
labour = 675000

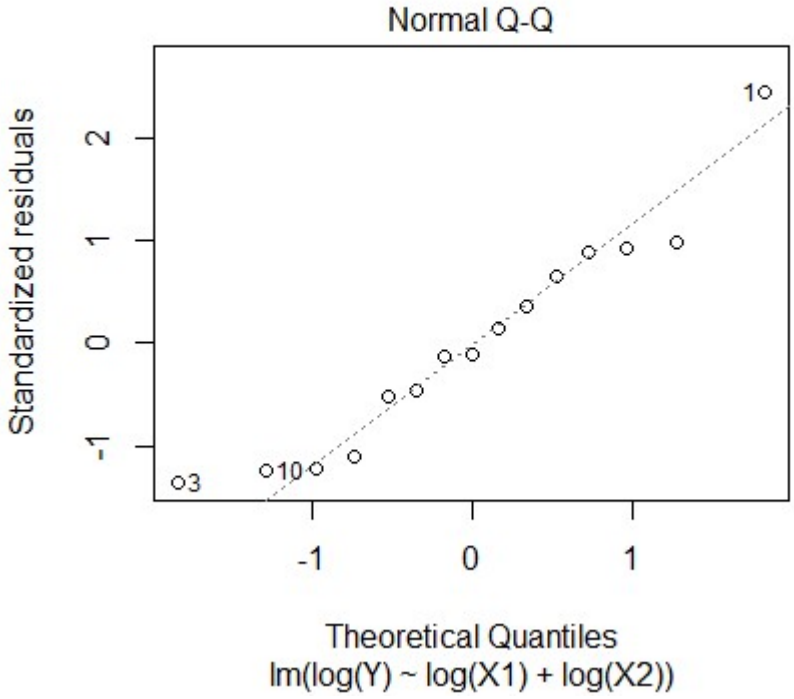
DataPoint = data.frame(X1 = capital, X2 = labour )

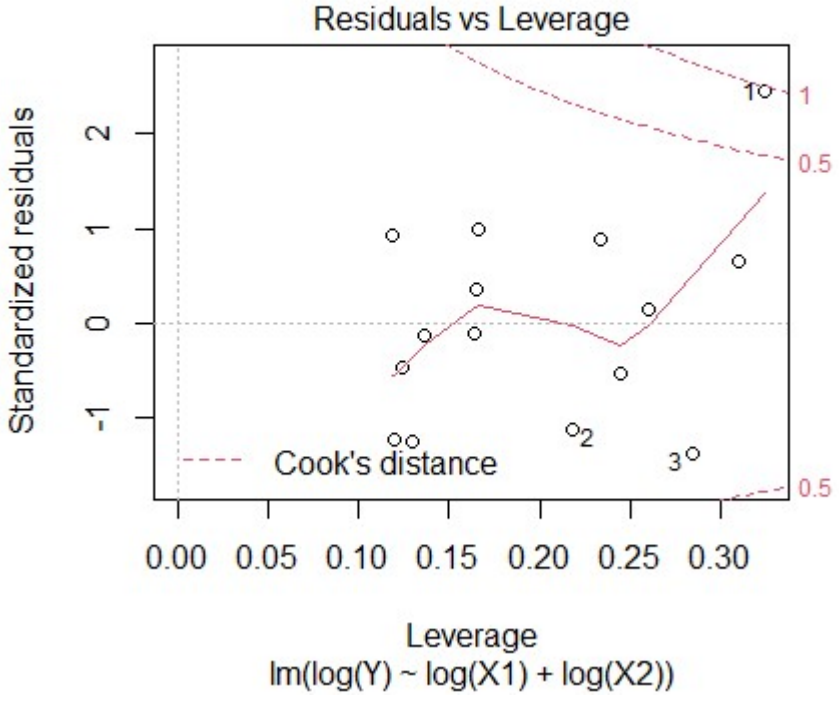
exp(predict(fit, DataPoint, interval="predict")) ## Need exponential to undo log

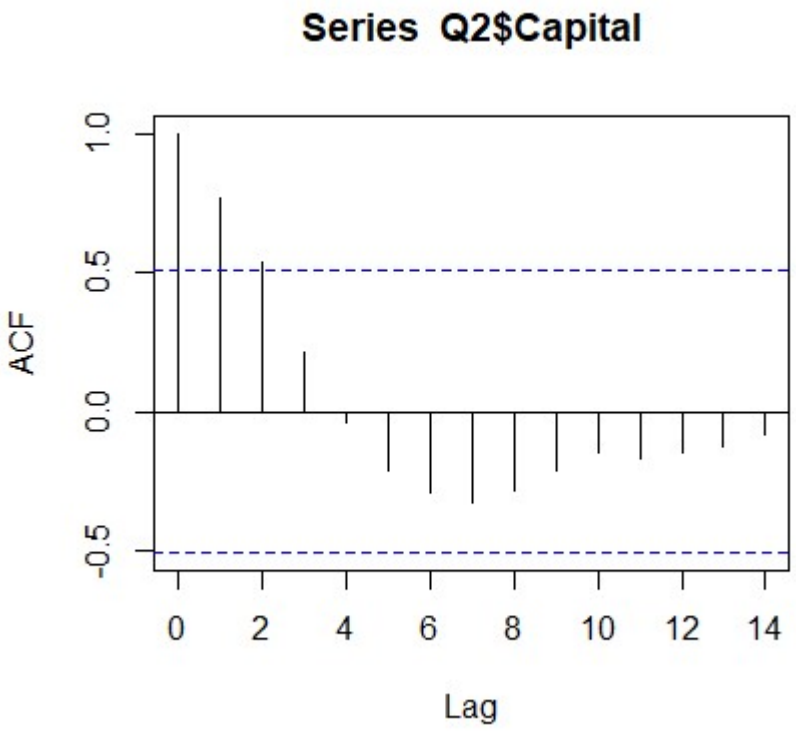
Predicted Value = 6486.535

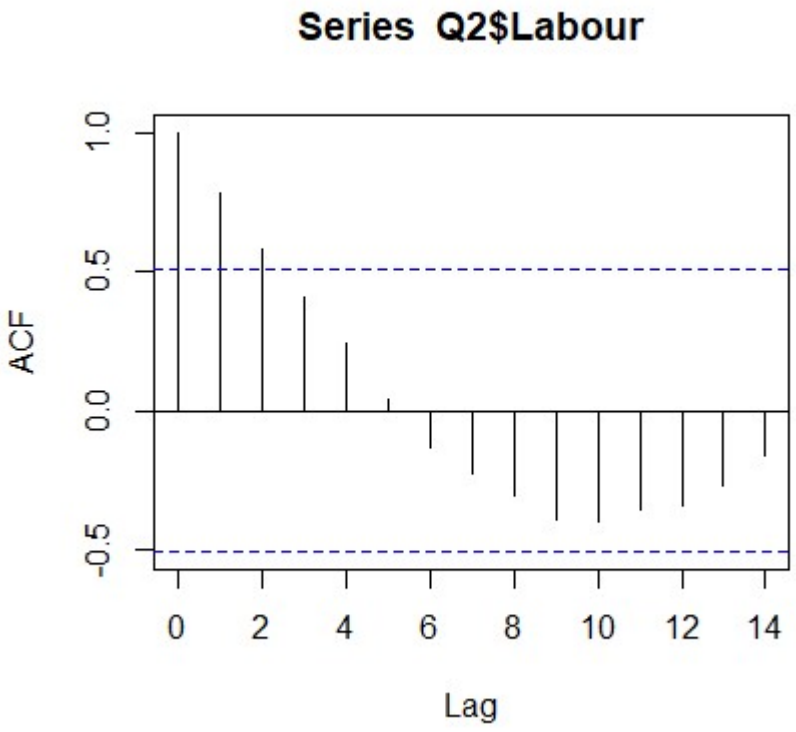
This is an acceptable result indicating the adequacy of the model as the predicated value is in agreement with the observations found the data set.




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(Y) ~ log(X1) + log(X2))

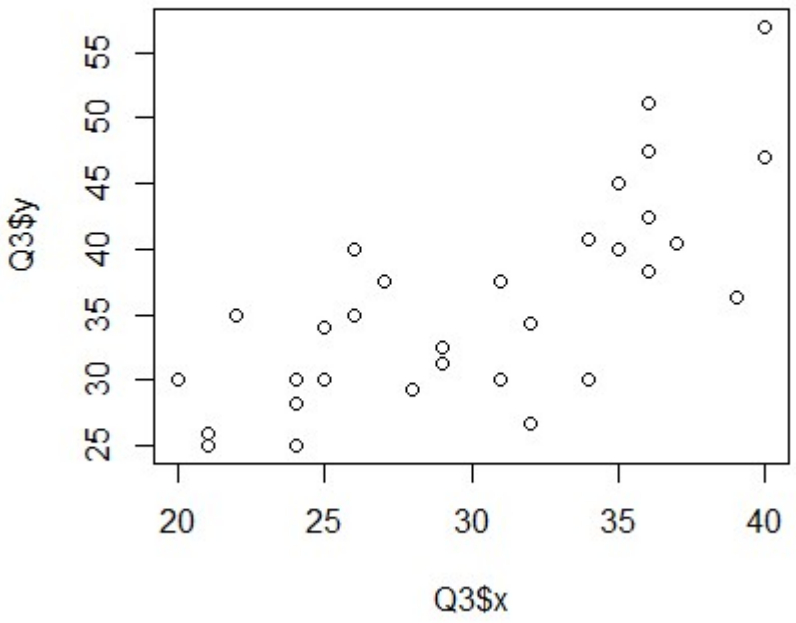

Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(log(Y) ~ log(X1) + log(X2))


Series Q2$Capital
ACF
1.0
0.5
0.0
-0.5
0
2
4
6
8
10
12
14
Lag


Series Q2$Labour
ACF
1.0
0.5
0.0
-0.5
0
2
4
6
8
10
12
14
Lag

Q3

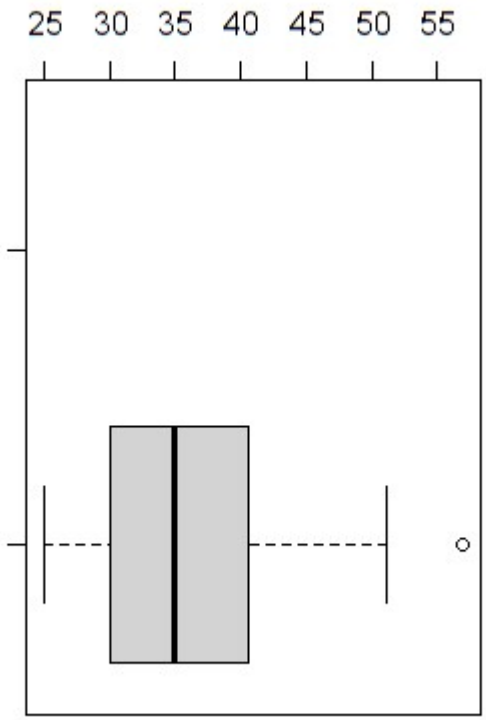
a)

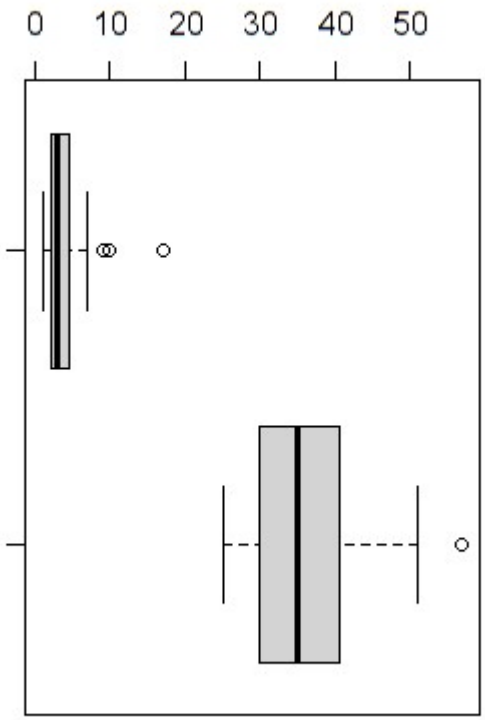



Plotting Travel Time, x, against Average of Travel Times, y, shows a moderately strong positive correlation between the two.

B)

```
obsDF2$mergeDest <- factor(obsDF2$Destination, levels = 1:7)

obsDF2$mergeDest[20] <- 6 #Destination 6

obsDF2$mergeDest[11] <- 6 #Destination 10

obsDF2$mergeDest[29] <- 6 #Destination 17

obsDF2$mergeDest[28] <- 5 #Destination 9

obsDF2$mergeDest[22] <- 5 #Destination 5

obsDF2$mergeDest[17] <- 5 #Destination 5

boxplot(Q3$destination,Q3$y, data=obsDF2)

boxplot(obsDF2$mergeDest,Q3$y, data=obsDF2)
```


25
30
35
40
45
50
55


0
10
20
30
40
50

c)

## Most common models are simple linear regression and logarithmic model

## Logarithmic models are good to handle non-linear relationships between

# independent and dependent variables, i.e preserves the linear model

## Plotting x,y we can see that there is significant correlation between

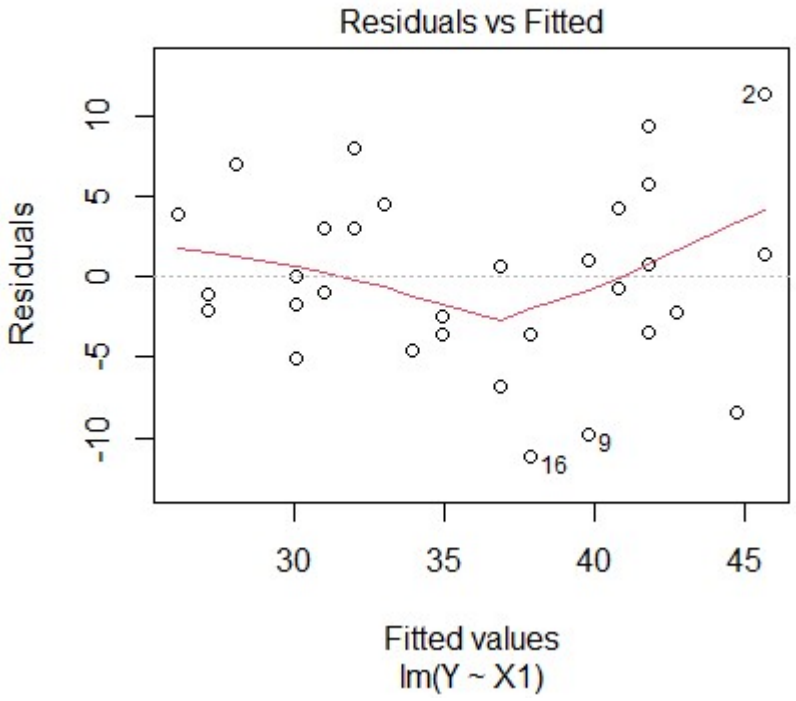
# the two. The relation ship looks linear or may be Y ~ X^2, meaning

# a square root transformation is a good candidate.

- We will see from below with the Residuals vs Fitted that all three models are valid. The residuals are distributed randomly about the horizontal axis, indicating that they are i.i.d. It is clear there is no pattern in how they are spread meaning the assumptions for a valid are met.

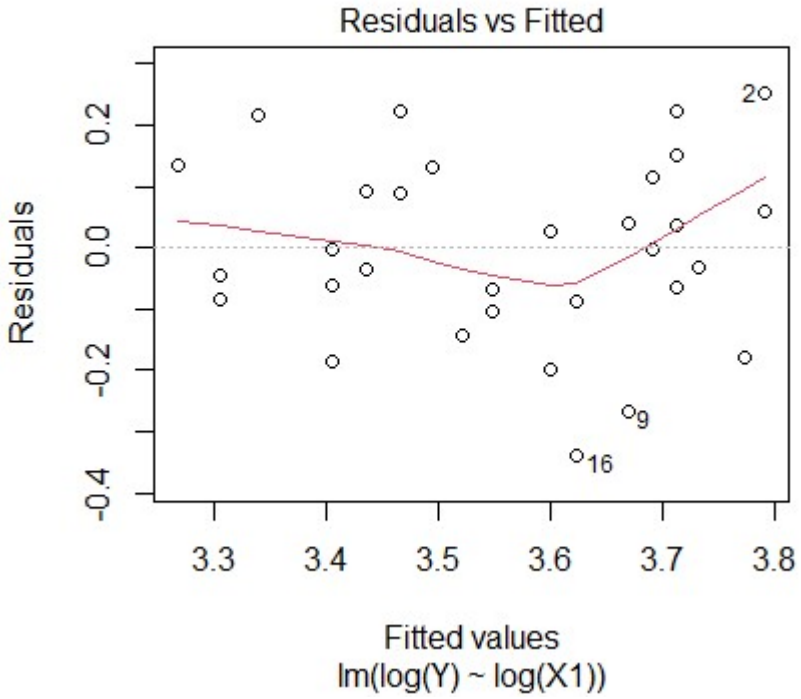
## M1, Simple Linear Regression Model

```
m1<-lm(Y ~ X1, data=obsDF2)
summary(m1)
plot(m1)
```



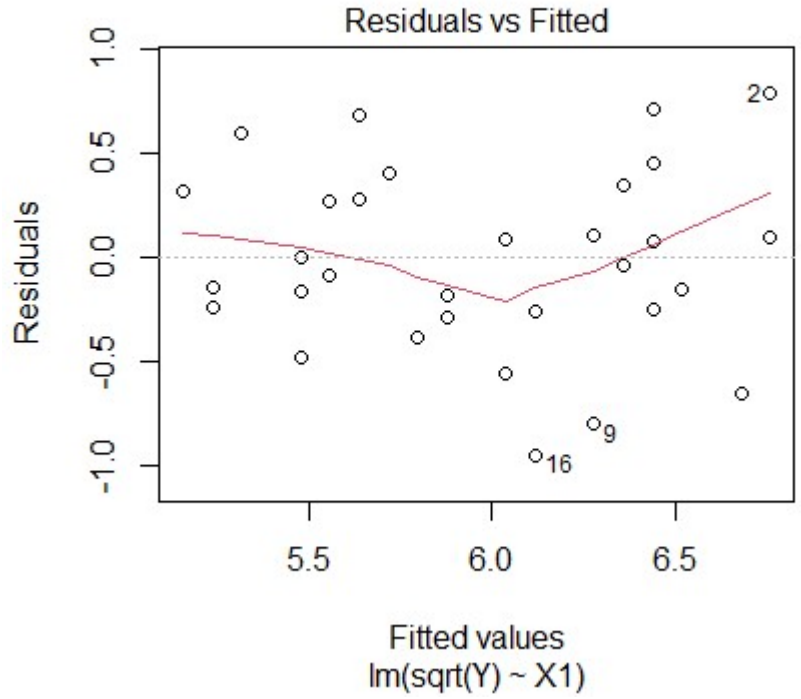
## M2, Logarithmic Model

```
m2<-lm(log(Y)~log(X1), data=obsDF2)
summary(m2)
plot(m2)
```



## M3, Square Root Model

```
m3<-lm(sqrt(Y)~X1)
summary(m3)
plot(m3)
```




d)

AIC ( Akaike Information Criterion)

| M1 | M2 | M3 |
|---|---|---|
| 110.6806 | -119.6116 | -50.08639 |

- According to the AIC comparison, Model 2 should be selected as it yielded the lowest AIC value indicating the least amount of information lost in relation to true model.

Mallows' Cp ( Comparing m2 and m3 to m1)

| M2 | M3 |
|---|---|

| | |
|---|---|
| -27.97753 | -27.80265 |

- Both models are roughly the same and both values > # of Predictor Variables + 1, meaning both models are unbiased.

BIC ( Bayesian Information Criterion)

| M1 | M2 | M3 |
|---|---|---|
| 207.8899 | -22.40237 | 47.12289 |

- According to BIC, model two is the best candidate as it as the lowest value out of the three meaning it has the best likelihood to predict values