STA 4155 Aryan Khandelwal 24061161

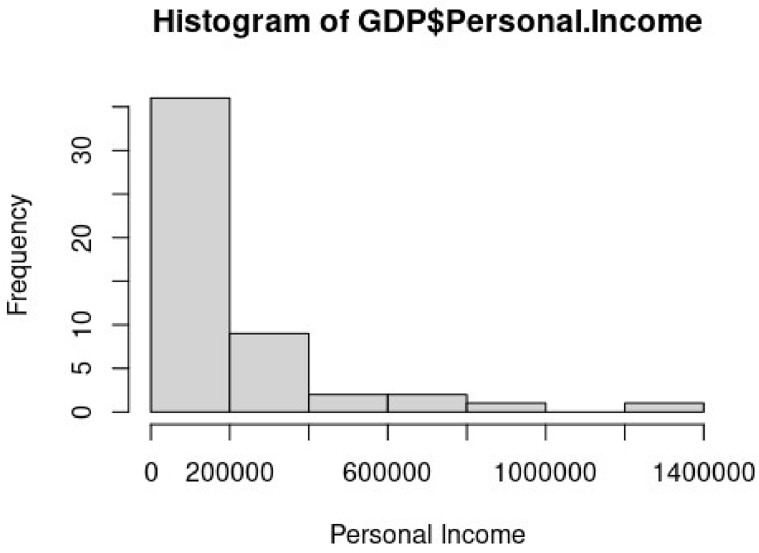
Homework 3

```
library(readr)
GDP <- read.table('GDP.txt', sep = '\t', header = TRUE)
```

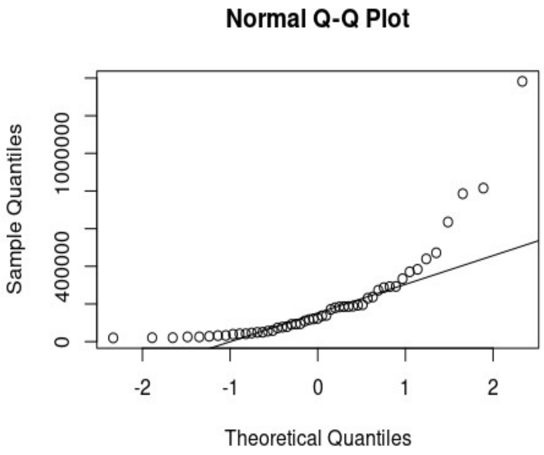
Question 1

In making a normal plot the Normal Population Assumption is violated. The original histogram is skewed to the right and thus transforming it by using a log10, is the most suitable. Because the histogram is skewed to the right we have to go down in the ladders of power. Also, once we transform the new histogram is not skewed in any way. Also, as we see in the last QQ plot made of the transformed version is straighter. This straightness in the qqplot satisfies the normal population assumption.

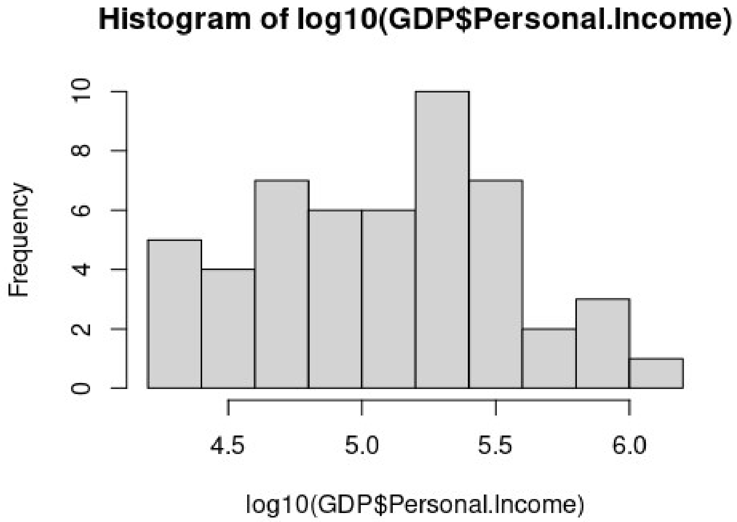
```
hist(GDP$Personal.Income, xlab = "Personal Income")
```



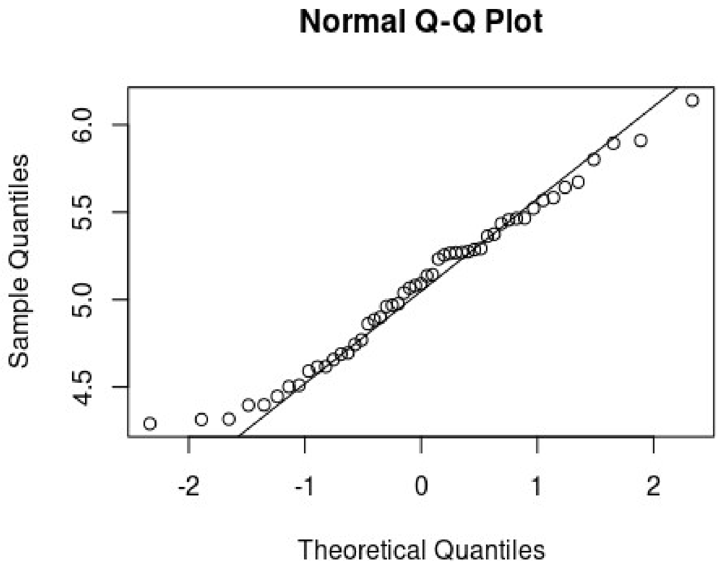
```
qqnorm(GDP$Personal.Income)
qqline(GDP$Personal.Income)
```



```
hist(log10(GDP$Personal.Income))
```



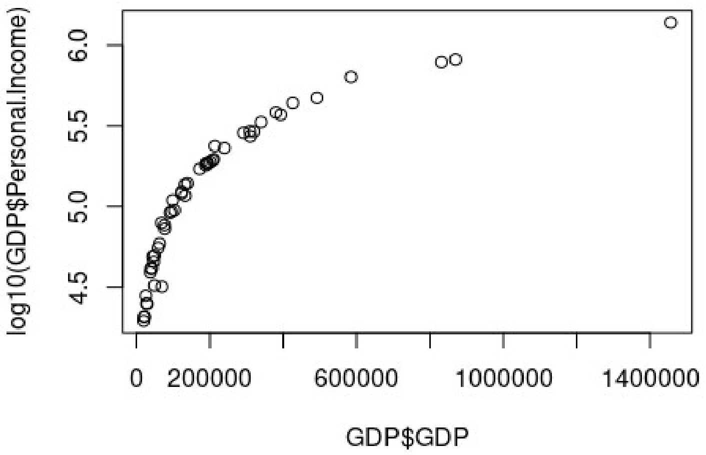
```
qqnorm(log10(GDP$Personal.Income))
qqline((log10(GDP$Personal.Income)))
```



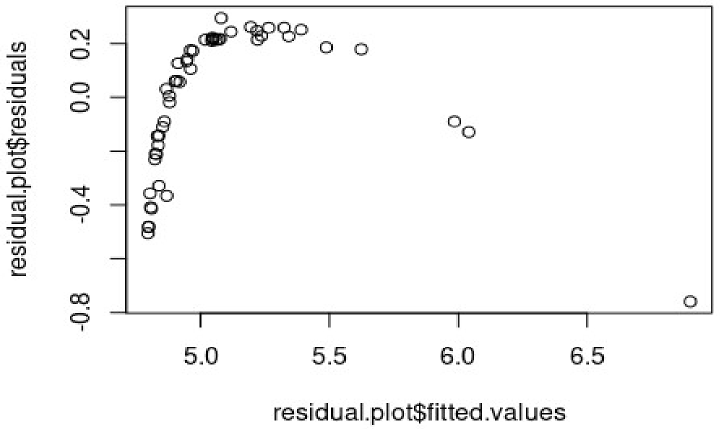
Question 2

The original plot with Personal Income against transformed GDP is not satisfying the linearity condition. Therefore we have to transform the GDP. Due to the shape of the graph, we have to travel down the ladder of power. The correct transformation in this situation will use a log10. After transforming the plot with log10GDP, the plot does satisfy the linearity condition as the scatterplot has a linear pattern.

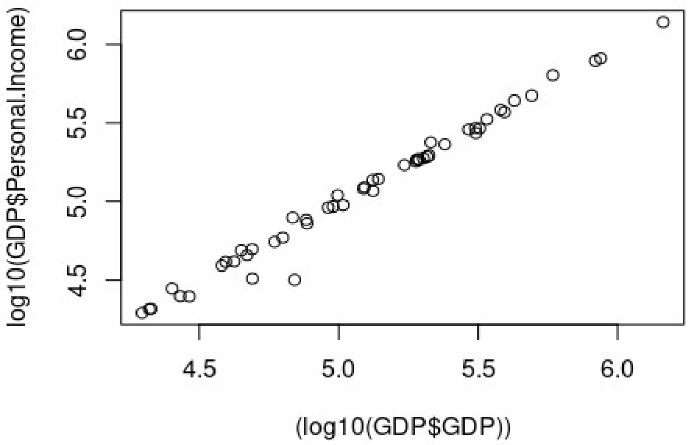
```
residual.plot <- lm(log10(Personal.Income)~GDP,data = GDP)
plot(GDP$GDP,log10(GDP$Personal.Income))
```



```
plot(residual.plot$fitted.values,residual.plot$residuals)
```



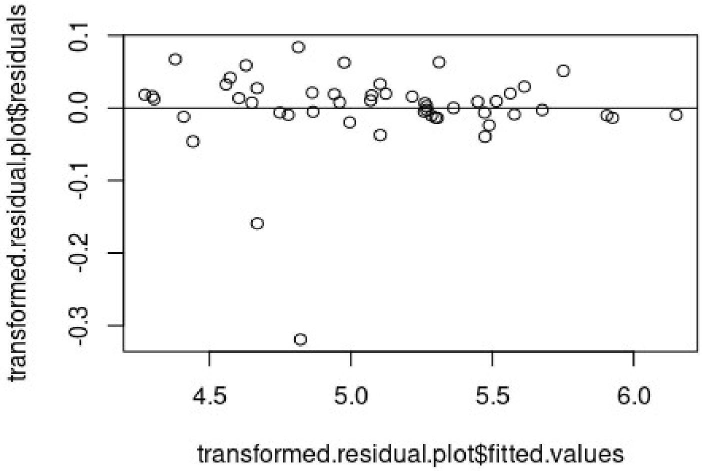
```
transformed.residual.plot <- lm(log10(Personal.Income)~log10(GDP),data = GDP)
plot((log10(GDP$GDP)),log10(GDP$Personal.Income))
```



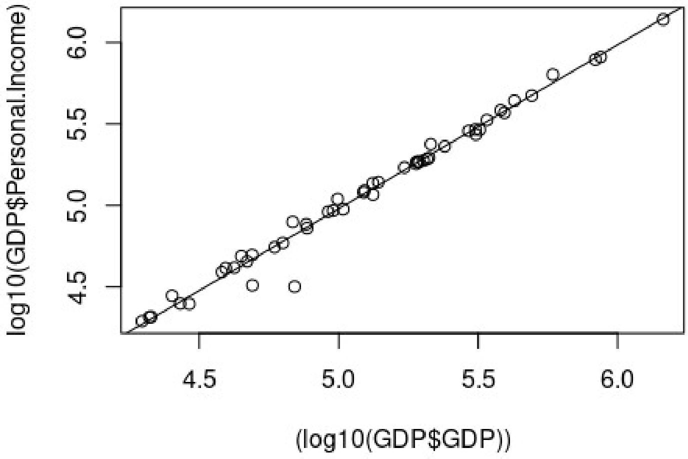
Question 3

The equal variance assumption is satisfied in the linear regression model because the scatterplot is spread nearly constant around the line. Also, as we can see on the residual plot, the points are scattered around 0, without any tendencies to be positive or negative, which also supports that the plot has equal variance.

```
plot(transformed.residual.plot$fitted.values,transformed.residual.plot$residu
als)
abline(a=0,b=0)
```



```
plot((log10(GDP$GDP)),log10(GDP$Personal.Income))
abline(transformed.residual.plot)
```



Question 4

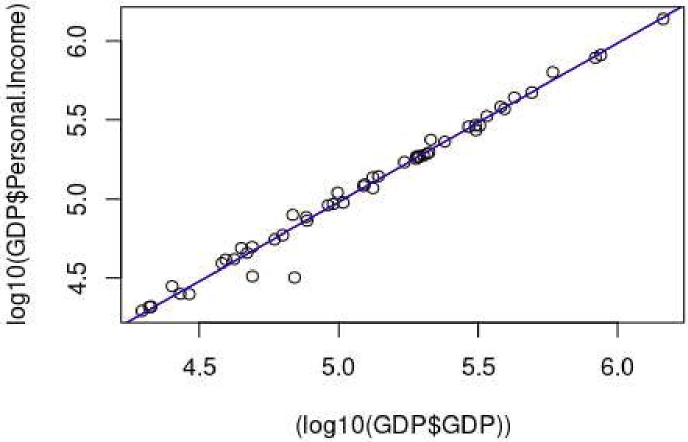
There is an outlier, which is California. It has high leverage, but a low residual. We can see this as it is very far out from the rest of the points, however removing California from the scatterplot does not have any significant effect on the linear regression model. Thus, California has little influence but has high leverage. California has little residuals because it is close to the linear regression line.

```
plot((log10(GDP$GDP)),log10(GDP$Personal.Income))
abline(transformed.residual.plot)
a <-which.min(residual.plot$residuals)
a.new <- GDP[-a,]

b <- GDP[a,]
b

##        State Personal.Income     GDP Population
## 5 California         1382235 1457090   36132147


California.Removed <-lm(log10(Personal.Income)~log10(GDP),data = a.new)
abline(transformed.residual.plot, col = 'red')
abline(California.Removed, col = 'blue')
```



Question 5

The predicted personal income for a GDP of 300,000 is 288544.1. We are 95% confident that the predicted income is between 218840.9 and 380448.5. 95% interval = (218840.9 , 380448.5).

```
summary(transformed.residual.plot)

##
## Call:
## lm(formula = log10(Personal.Income) ~ log10(GDP), data = GDP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31957 -0.00977  0.00719  0.01977  0.08407
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04433    0.09324  -0.475    0.637
## log10(GDP)   1.00501    0.01821  55.180   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05878 on 49 degrees of freedom
## Multiple R-squared:  0.9842, Adjusted R-squared:  0.9838
## F-statistic:  3045 on 1 and 49 DF,  p-value: < 2.2e-16

new.data.new = data.frame(GDP = 300000)
predicted.result <- predict(transformed.residual.plot, newdata = new.data.new
,interval = 'prediction', level = 0.95)
predicted.result
```

```
##        fit      lwr      upr
## 1 5.460212 5.340128 5.580296
```

```
(10^predicted.result)
```

```
##        fit      lwr      upr
## 1 288544.1 218840.9 380448.5
```