Marcus Crowder
STA-3155
R HW #3

1. Make a histogram and Q-Q plot of P ersonal Income to check the Normal Population Assumption. If the assumption is violated, propose an appropriate transformation from the Ladder of Powers for P ersonal Income, and justify it.

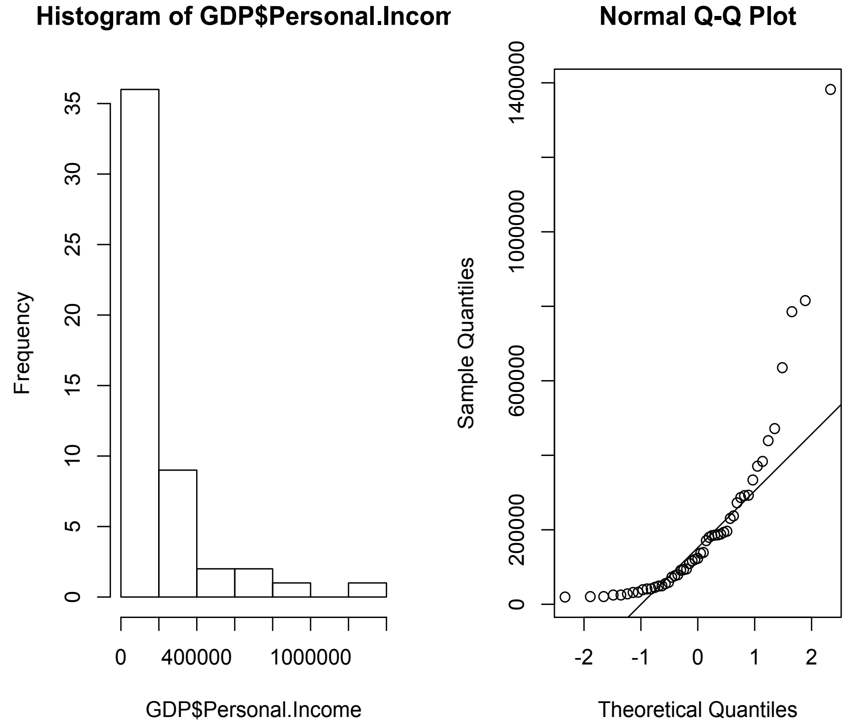
The first thing I did was load the data in with the R command:

```
> GDP <- read.table('~/Downloads/GDP.txt', sep = '\t', header = TRUE)
```

Then I went on to create the histogram and QQ plot for personal income

```
> par(mfrow = c(1,2))
> hist(GDP$Personal.Income)
> qqnorm(GDP$Personal.Income)
> qqline(GDP$Personal.Income)
```

Just by looking at the graphs it is possible to see that the data does not fit. On the histogram the data is skewed to the right. A similar scenario occurs on the Q-Q plot which means it is time to check if any transformations from the ladder of powers may work.

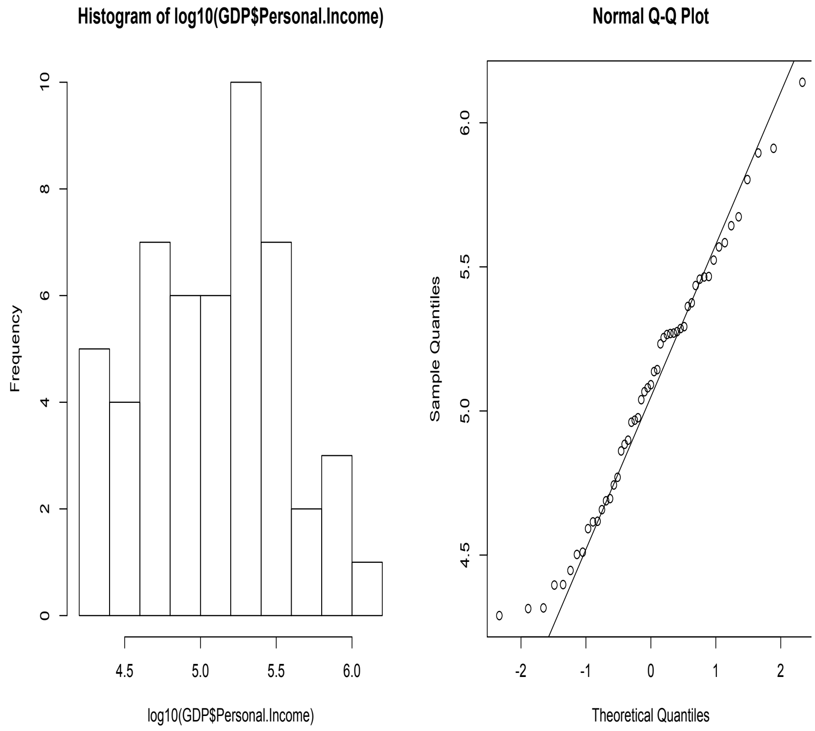


I decided to go with the Power "0" or log10 which states: "For data that are non-negative especially for those that grow by percentage increases, e.g., salaries". Regarding personal income, I consider it a salary and it was also the best graph in terms of skewness and fitting Q-Q line.

The histogram now appears to have less skewness towards the right or atleast not as noticeable as the previous histogram. In addition the looks more normal when looking at the Q-Q plot as the points do not wander far from the line.

R commands used:

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



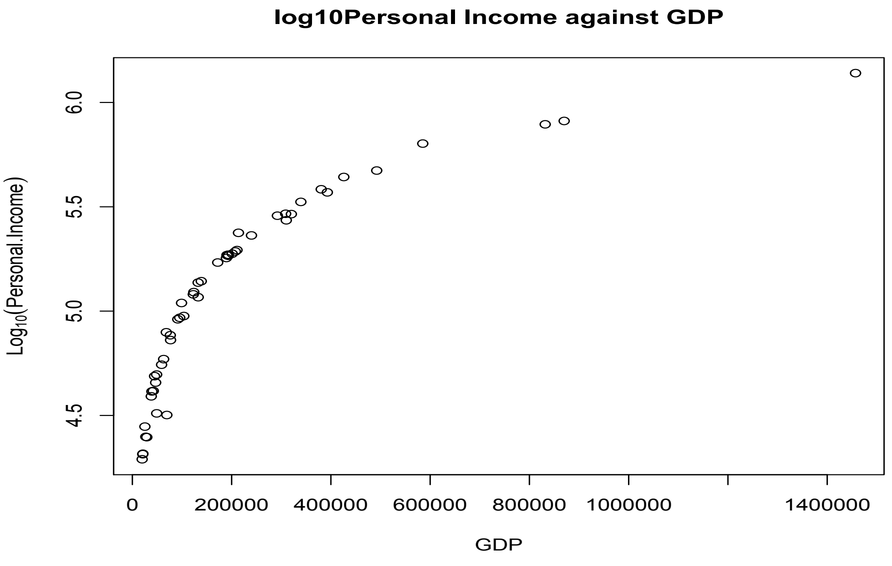
2. Make a scatterplot of the transformed Personal Income against GDP, and check the linearity assumption. If the assumption is not satisfied, propose an appropriate transformation from the Ladder of Powers on GDP, and justify it.

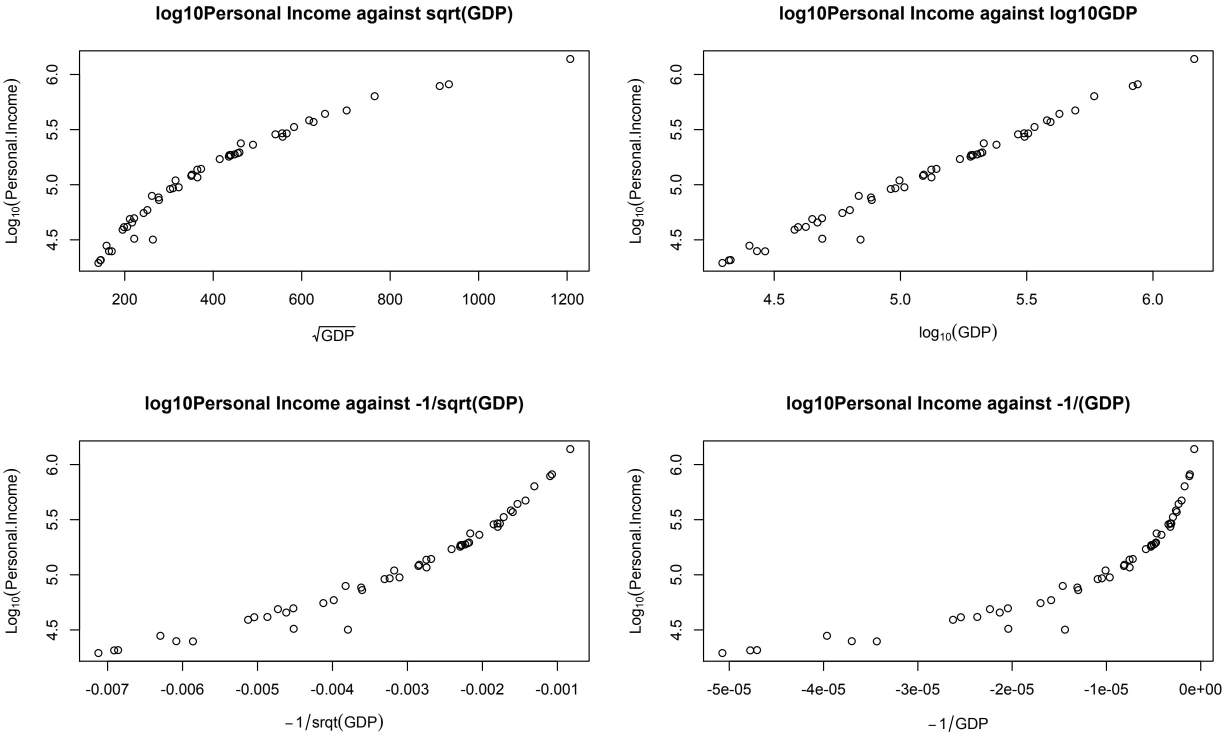
After encountering some technical errors with regard to commas and capitalization in R I was able to plot the Transformed Personal Income Against GDP

Command:

```
> plot(GDP$GDP, log10(GDP$Personal.Income), main = "log10Personal Income against GDP",
xlab="GDP", ylab =expression(Log[10](Personal.Income)))
```

However there seems to be a problem with this plot. It does not fit the linearity assumption. There seems to be a bend or a curve occurring at the top of the scatterplot. It is now up to us to transform the GDP until the assumption is satisfied.





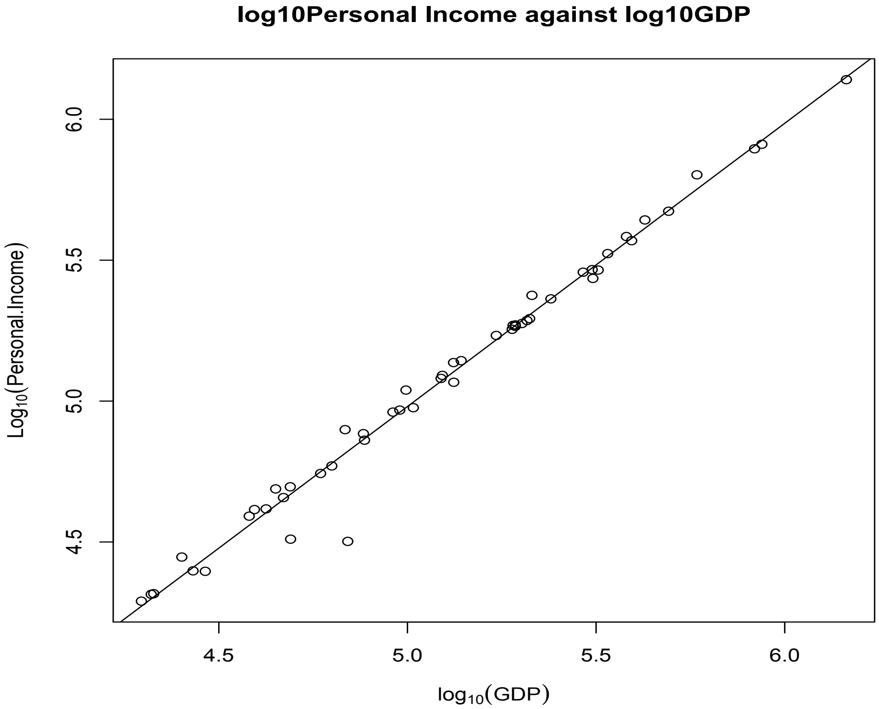
Which led me to try transforming GDP by sqrt(GDP), by log10(GDP), by -1/sqrt(GDP) and by -1/GPD. By looking at the graph we can see that the line is most linear with the log10 transformation of GDP.

Commands used to produce Graph:

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

> plot(sqrt(GDP$GDP), log10(GDP$Personal.Income), main = "log10Personal Income against
sqrt(GDP)", xlab=expression(sqrt(GDP)), ylab =expression(Log[10](Personal.Income)))

> plot(log10(GDP$GDP), log10(GDP$Personal.Income), main = "log10Personal Income against
log10GDP", xlab=expression(log[10](GDP)), ylab =expression(Log[10](Personal.Income)))

> plot(-1/sqrt(GDP$GDP), log10(GDP$Personal.Income), main = "log10Personal Income
against -1/sqrt(GDP)", xlab=expression(-1/srqt(GDP)), ylab
=expression(Log[10](Personal.Income)))

> plot(-1/GDP$GDP, log10(GDP$Personal.Income), main = "log10Personal Income against -
1/(GDP)", xlab=expression(-1/GDP), ylab =expression(Log[10](Personal.Income)))
```

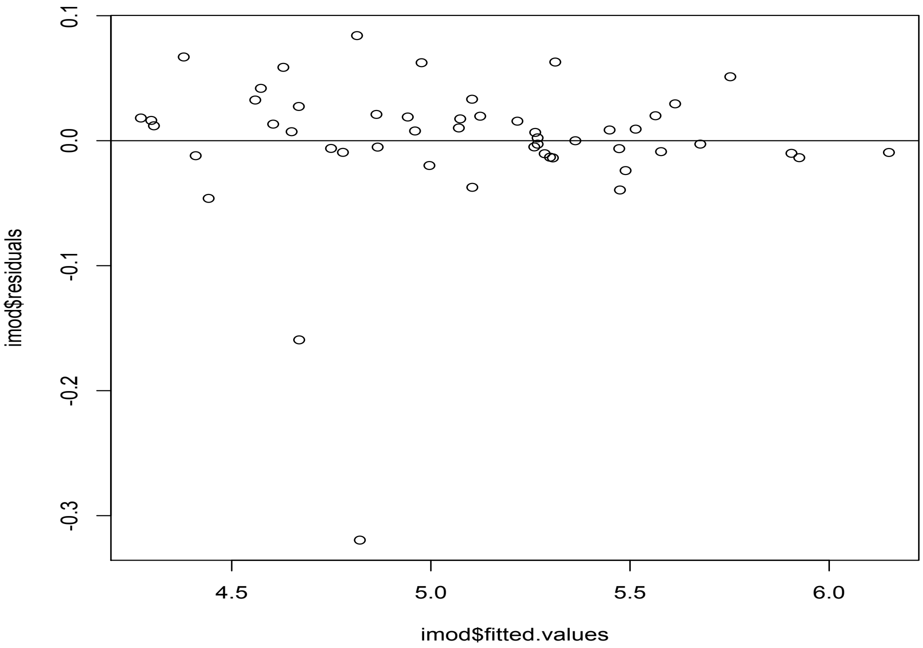
3. Fit a linear regression model using the transformed GDP as x-variable and the transformed Personal Income as y-variable, based on the transformations determined in the previous questions. Produce a residual plot against fitted values. Is the equal-variance assumption satisfied in the fitted model? Explain.

This is the plot of log10(Personal Income) Against the plot of log10(GDP) with a fitted linear regression line through it. The data points look great with little residuals or outliers. Points tend to be close to the model.

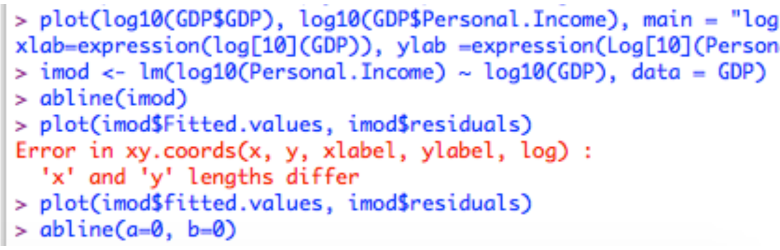


This is the fitted linear regression model (imod) of log10(PersonalIncome) and log10(GDP) where the residual values are plotted against the fitted values.

The Equal Variance assumption appears to be satisfied because there seems to be equal spread around the line 0,0.



R commands used.



```
> plot(log10(GDP$GDP), log10(GDP$Personal.Income), main = "log10Personal Income against log10GDP",
xlab=expression(log[10](GDP)), ylab =expression(Log[10](Personal.Income)))
> imod <- lm(log10(Personal.Income) ~ log10(GDP), data = GDP)
> abline(imod)
> plot(imod$Fitted.values, imod$residuals)
Error in xy.coords(x, y, xlabel, ylabel, log) :
  'x' and 'y' lengths differ
> plot(imod$fitted.values, imod$residuals)
> abline(a=0, b=0)
```

```
> plot(log10(GDP$GDP), log10(GDP$Personal.Income), main = "log10Personal Income against
log10GDP", xlab=expression(log[10](GDP)), ylab =expression(Log[10](Personal.Income)))
> imod <- lm(log10(Personal.Income) ~ log10(GDP), data = GDP)
> abline(imod)
```

```
> plot(imod$fitted.values, imod$residuals)
> abline(a=0, b=0)
```

4. Is there any unusual observation according to the linear model you fitted above? If so, find the states they come from. Are they outliers? Do they have high leverages? Are they influential points? Explain.

There seem to be two unusual observations surrounding 2 points with high residuals when compared to other points. The points appear low on the residual plot graph. To find these 2 values we notice that the points are below -0.1. Then input these R commands.

```
> outlier <- (imod$residual < -0.1)
> GDP[outlier,]
               State Personal.Income   GDP Population
8           Delaware           32359 49001     843524
9 District of Columbia         31779 69470     550521
```

Answer: Considering these points appear to be far from the regression line of y values and appears to have a high residual we can claim these points are outliers.
The points come from the State of Delaware and District of Columbia.

To find out if the point has a high leverage I calculated the log10 of the points compared with the mean log10 of all the points. And got these results:

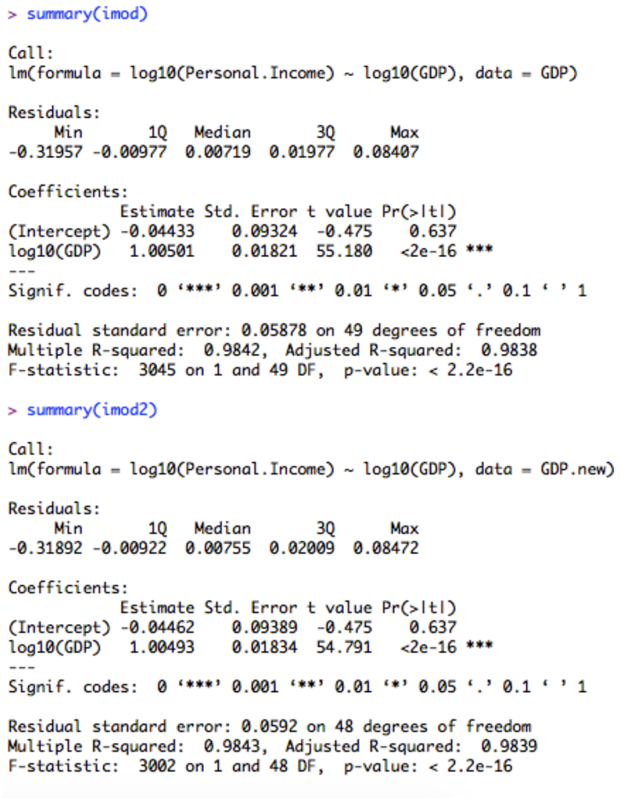
```
> log10(GDP$GDP[8])
[1] 4.690205
> log10(GDP$GDP[9])
[1] 4.841797
```

```
> mean(log10(GDP$GDP))
[1] 5.099385
```

The points of the outliers do not same to be too far away from the mean x value of all the points so we can say that it does not have a high leverage.

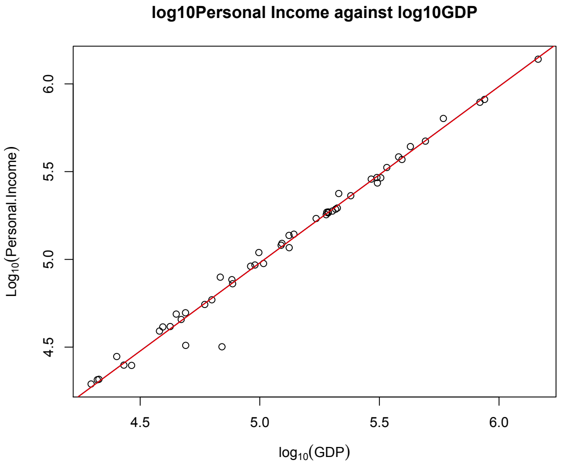
By using the summary() function I was able to tell that the points were not that influential. If we take away the points there is a negligible change in the mean of log10(GDP)   1.00501 vs 1.00493
The Y intercepts do not seem to differ by much either.

Also there were no major differences in the linear regression line when the points are taking out of the plot.



```
> summary(imod)

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

> summary(imod2)

Call:
lm(formula = log10(Personal.Income) ~ log10(GDP), data = GDP.new)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.31892 -0.00922  0.00755  0.02009  0.08472 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.04462    0.09389  -0.475    0.637    
log10(GDP)   1.00493    0.01834  54.791   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0592 on 48 degrees of freedom
Multiple R-squared:  0.9843,	Adjusted R-squared:  0.9839 
F-statistic:  3002 on 1 and 48 DF,  p-value: < 2.2e-16
```

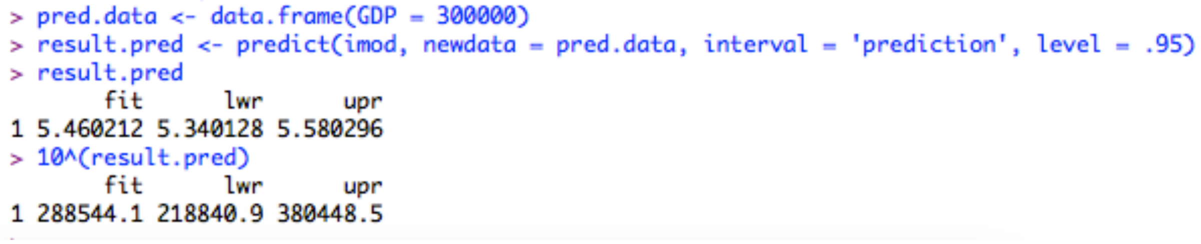
```
> plot(log10(GDP$GDP), log10(GDP$Personal.Income), main = "log10Personal Income against log10GDP", xlab=expression(log[10](GDP)), ylab =expression(Log[10](Personal.Income)))
> abline(imod, col = 'blue')
> abline(imod2, col = 'red')
```



5. Suppose there exists a State51 and its GDP = 300, 000. Can you predict the Personal Income for that state using the model from Question (3)? Please also show its 95% prediction interval.

After checking to find out if we can use the data without extrapolation.
Turns out we can.

```
> which.max(GDP$GDP)
[1] 5
> GDP[5,]
       State Personal.Income     GDP Population
5 California         1382235 1457090   36132147

> which.min(GDP$GDP)
[1] 51
> GDP[51,]
     State Personal.Income   GDP Population
51 Wyoming           19501 19713     509294
```



```
> pred.data <- data.frame(GDP = 300000)
> result.pred <- predict(imod, newdata = pred.data, interval = 'prediction', level = .95)
> result.pred
       fit      lwr      upr
1 5.460212 5.340128 5.580296
> 10^(result.pred)
       fit      lwr      upr
1 288544.1 218840.9 380448.5
```

Answer: We can predict the Personal Income of a city with GDP = 300,000 because it fits our model. And the prediction Interval (after applying power of 10 to counter log10 transformation) is
A fit of 288544.1 while lower tail is 218840.9 and upper tail is at 380448.5