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

Answer

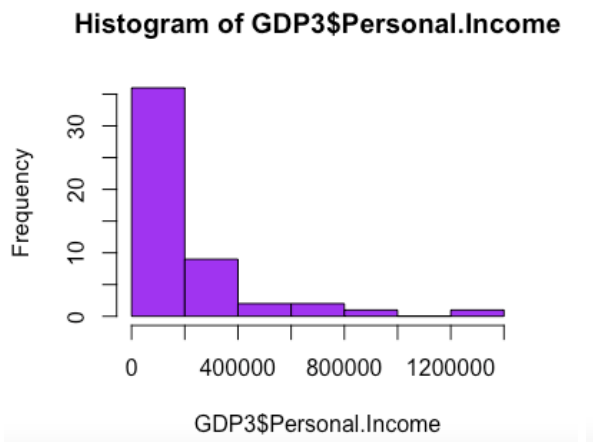
The histogram is skewed to the right and the Q-Q plot is curved upwards. Therefore, we need to go down the ladder to reduce the skewness to the right. We need to transform y by applying square roots or applying log(10). This will make the Q-Q plot and histogram more normal. We would apply the log(10) function in order to make the histogram more normal and obtain linearity in the Q-Q plot.

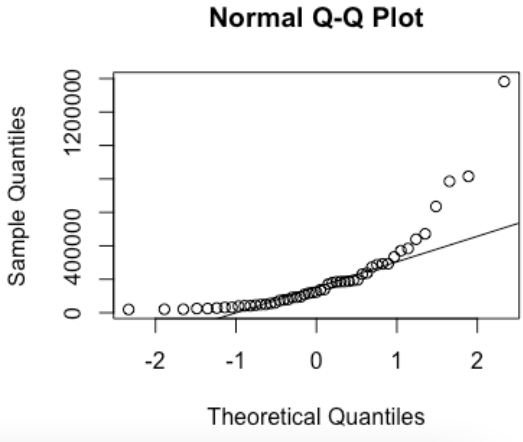
As a result of the log(10) transformation of Personal Income, the histogram has a normal distribution and the linearity in the Q-Q plot is satisfied.

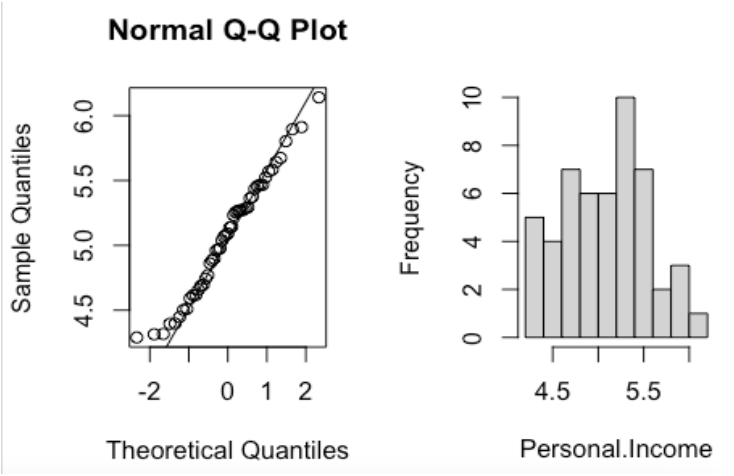
R-code

```
GDP3 <- read.table('GDP.txt', sep = '\t', header = TRUE)
GDP3
hist(GDP3$Personal.Income, xlab='Personal Income', main = "", col="purple") #skew to the right
qqnorm(GDP3$Personal.Income)
qqline(GDP3$Personal.Income) #not linear
hist(log10(GDP3$Personal.Income), xlab= expression(Personal.Income), main = " ")
qqnorm(log10(GDP3$Personal.Income))
qqline(log10(GDP3$Personal.Income))
```

R output







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.

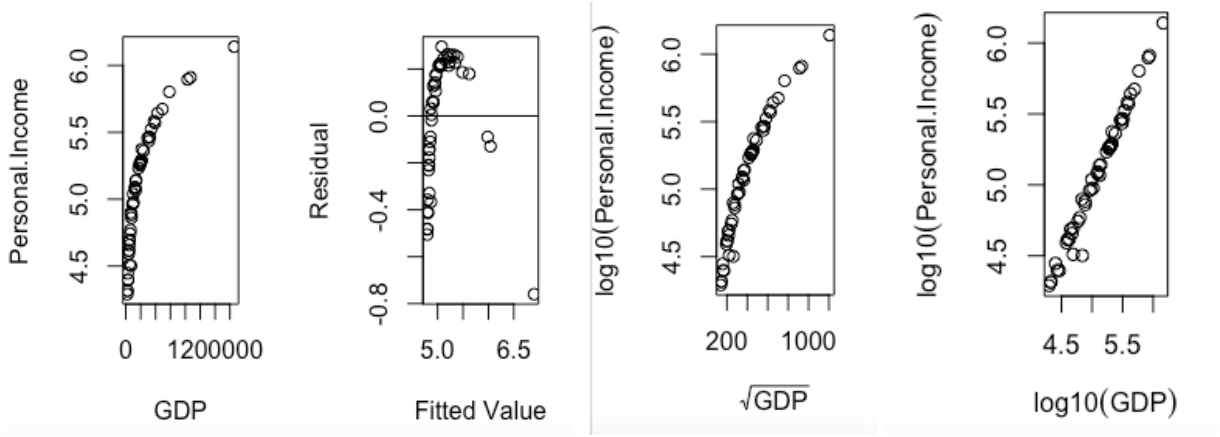
Answer

The linearity assumption isn't satisfied because the graphs is curved downwards, and there's evidence
of unequal spread and a curved pattern in the residual plot. To fix this, we'll try transformations from the ladders of powers on GDP.
After the initial transformation, the plot here is slightly straighter than before, but we would still need to test other transformations for a more linear line.
The log(10) function straightened out the scatterplot well, we'll use the log(10) transformation for GDP.

R-code

```
plot1 <- lm(log10(Personal.Income) ~ GDP, data = GDP3) #residuals
plot(GDP3$GDP, log10(GDP3$Personal.Income), xlab = 'GDP', ylab =
expression('Personal.Income'))
plot(plot1$fitted.values, plot1$residuals, xlab = 'Fitted Value', ylab = 'Residual')
abline(0,0)
plot(sqrt(GDP3$GDP),log10(GDP3$Personal.Income), xlab = expression(sqrt(GDP)),
    ylab = expression(log10(Personal.Income)))
plot(log10(GDP3$GDP), log10(GDP3$Personal.Income), xlab = expression(log10(GDP)),
    ylab = expression(log10(Personal.Income)))
```

R output



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.

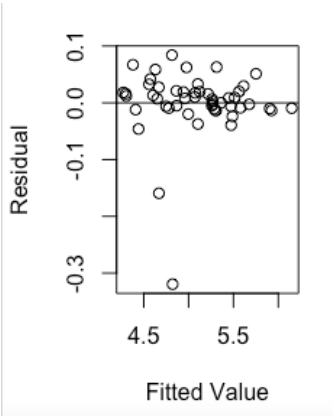
Answer

We'll then explore the outcome of the residual plot based on our recent transformations. There appears to be equal spread about the y=0 line, so the linearity and constant variance assumptions are satisfied.

R-code

```
plot2 <- lm(log10(Personal.Income) ~ log10(GDP), data=GDP3)
plot(plot2$fitted.values, plot2$residuals, xlab = "Fitted Value", ylab = "Residual")
abline(a=0,b=0)
```

R output



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.

Answer

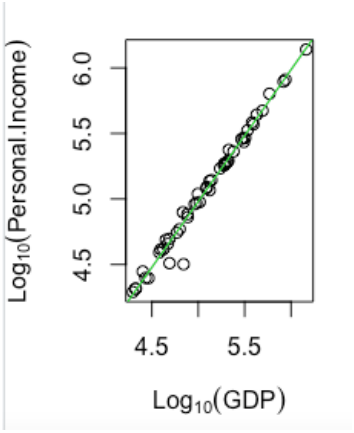
There are unusual observations in the residual plot because there's an outlier.
From the output, the outlier is from the District of Colombia. It has a GDP of 69,470, and a GDP population of 550,521. It has a significantly negative residual, meaning that given GDP is 69,470,
the GDP of that country is much lower than predicted by the linear model.
The point is an outlier because the y-value of the outlier is far from the regression model.
Because the 2 regression lines overlap, it's hard to see any difference.
Therefore, we'll compare the 2 models through the summary function.
Due to the fact that the coefficients and R^2 have similar values in both summary models, we've come to the conclusion that the outlier isn't influential. Thus, we don't need to remove it from the dataset.
The outlier doesn't have high leverage because it has a similar x-value to the average mean of the plot.

R-code

```
findout <- which.min(plot2$residuals)
GDP3[findout, ]
influen.colom <- GDP3[-findout, ]
plot3 <- lm(log10(Personal.Income) ~ log10(GDP), data=GDP3)
plot(log10(GDP3$GDP), log10(GDP3$Personal.Income), xlab = expression(Log[10](GDP)),
    ylab = expression(Log[10](Personal.Income)))
abline(plot2, col = 'purple')
abline(plot3, col = 'green')
summary(plot2)
summary(plot3)
```

R output

```
> findout <- which.min(plot2$residuals)
> GDP3[findout, ]
                State Personal.Income   GDP Population
9 District of Columbia           31779 69470     550521
```




```
Call:
lm(formula = log10(Personal.Income) ~ log10(GDP), data = GDP3)

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(plot3)

Call:
lm(formula = log10(Personal.Income) ~ log10(GDP), data = GDP3)

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
```

5. Suppose there exists a state with 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.

Answer

When a country has a GDP of 300,000, the GDP is predicted to be 278.74 with 95% CI from 195.92 to 396.55.

R-code

```
predict1 <- data.frame(GDP = 300,000)
result <- predict(plot2, newdata=predict1, interval = 'prediction', level=0.95)
result
10^(result) #transform predict1 back to its original scale.
```

R output

```
> predict1 <- data.frame(GDP = 300,000)
> result <- predict(plot2, newdata=predict1, interval = 'prediction', level=0.95)
> result
       fit      lwr      upr
1 2.445192 2.292089 2.598295
> 10^(result) #to reverse a log10 function, do 10^(function)
       fit      lwr      upr
1 278.7354 195.9247 396.5474
```

Appendix

```
GDP3 <- read.table('GDP.txt', sep = '\t', header = TRUE)
GDP3
names(GDP3)
hist(GDP3$Personal.Income, xlab='Personal Income',main = "", col="purple") #skew to the right
qqnorm(GDP3$Personal.Income)
qqline(GDP3$Personal.Income) #not linear
hist(log10(GDP3$Personal.Income), xlab= expression(Personal.Income), main = " ")
qqnorm(log10(GDP3$Personal.Income))
qqline(log10(GDP3$Personal.Income))
plot1 <- lm(log10(Personal.Income) ~ GDP, data = GDP3) #residuals
plot(GDP3$GDP,log10(GDP3$Personal.Income), xlab = 'GDP', ylab =
expression('Personal.Income'))
plot(plot1$fitted.values, plot1$residuals, xlab = 'Fitted Value', ylab = 'Residual')
abline(0,0)
plot(sqrt(GDP3$GDP),log10(GDP3$Personal.Income), xlab = expression(sqrt(GDP)),
    ylab = expression(log10(Personal.Income)))
plot(log10(GDP3$GDP), log10(GDP3$Personal.Income), xlab = expression(log10(GDP)),
    ylab = expression(log10(Personal.Income)))
plot2 <- lm(log10(Personal.Income) ~ log10(GDP), data=GDP3)
plot(plot2$fitted.values, plot2$residuals, xlab = "Fitted Value", ylab = "Residual")
abline(a=0,b=0)
findout <- which.min(plot2$residuals)
GDP3[findout, ]
influen.colom <- GDP3[-findout, ]
plot3 <- lm(log10(Personal.Income) ~ log10(GDP), data=GDP3)
```

```
plot(log10(GDP3$GDP), log10(GDP3$Personal.Income), xlab = expression(Log[10](GDP)),
    ylab = expression(Log[10](Personal.Income)))
abline(plot2, col = 'purple')
abline(plot3, col = 'green')
summary(plot2)
summary(plot3)
predict1 <- data.frame(GDP = 300,000)
result <- predict(plot2, newdata=predict1, interval = 'prediction', level=0.95)
result
10^(result)
```