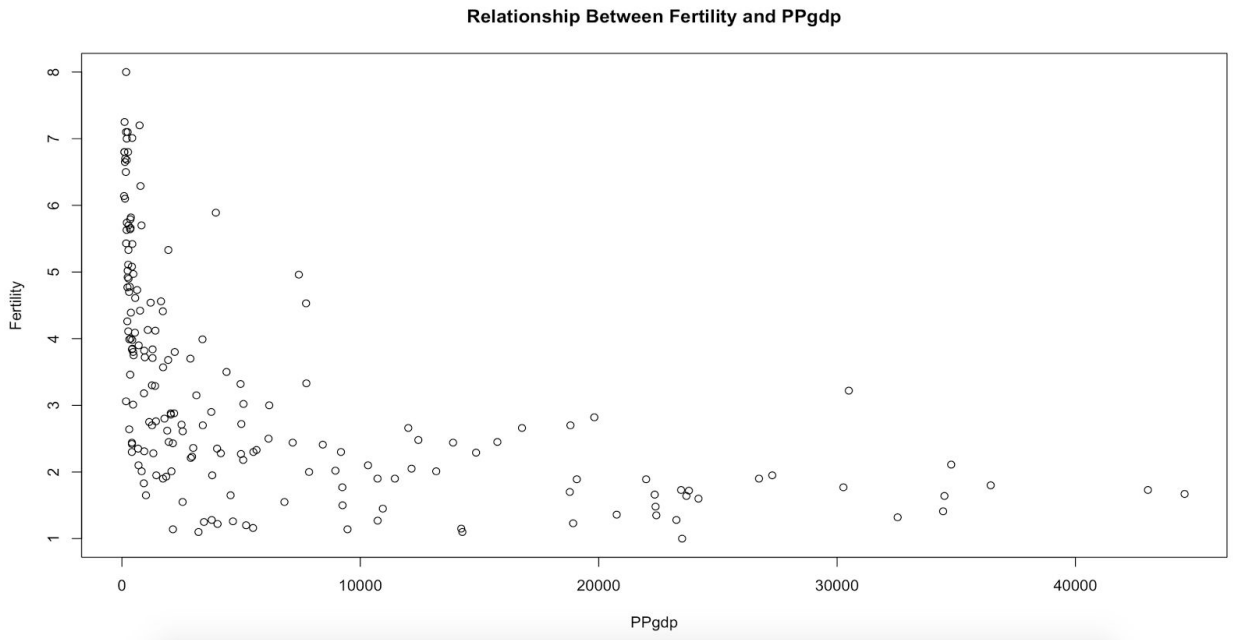
(In red: work that are not from the original R code but are from later on insertion by typing)
(In black: work that are from the original R code)
(In green: coding results from R console)
(In order to clearly show my work part by part, the plots related to the certain piece of code are inserted right after the code pieces, and statements/comments/explanations are followed.)

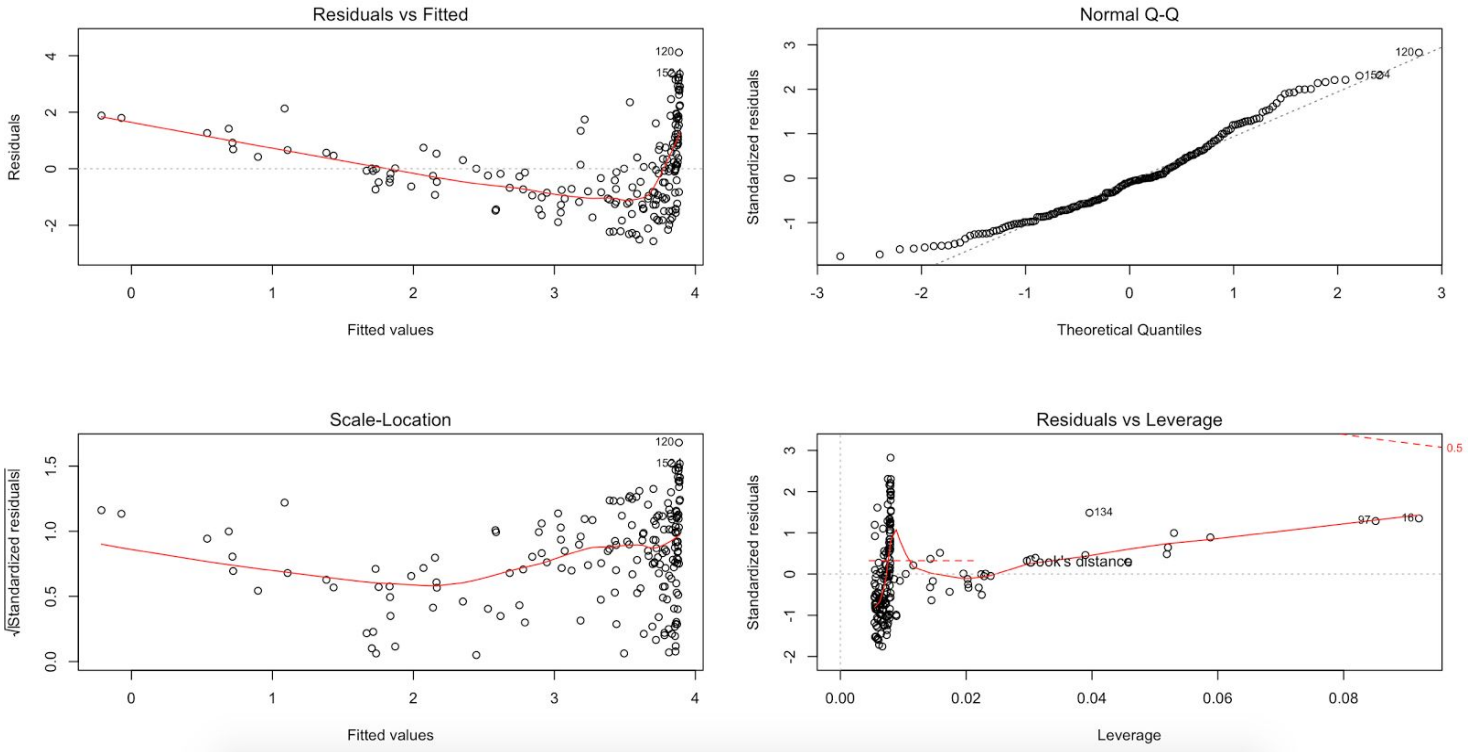
```
# The data in the file "UN.txt" contains PPgdp, the 2001 gross national product per person in $,
# and Fertility, the birth rate per 1000 femals in the population in the year 2000.
# The data are for 184 localities, mostly UN member countries,
# but also other areas such as Hong Kong that are not independent countries.
# In this problem, we study the relationship between Fertility and PPgdp.

# open the data file
dataUN = read.table("~/Desktop/Fall Quarter 2017/STA 108/project1/UN.txt", header=T)

# **********Data visualization and pre-processing**********
# **********                    1                **********
# Draw the scatterplot of Fertility on the vertical axis versus PPgdp on the horizontal axis
# and summarize the information in this graph.
# Does a simple linear regression model seem to be a plausible for a summary of this graph?

# assign values to variables
y = dataUN$Fertility
x = dataUN$PPgdp

# plot the data with a title and variable names
plot(x, y, xlab = "PPgdp", ylab = "Fertility", main = "Relationship Between Fertility and PPgdp")
```



Summary on the scatter plot:

From the scatter plot, first, we can easily see that the relationship between PPgdp (x-axis variables) and fertility (y-axis variables) is not linear, meaning the linearity assumption doesn't hold. On the very beginning of the left hand side of the graph, the corresponding fertility drops really obviously with a great amount as PPgdp increases approximately within the interval [0, 10000], then when PPgdp gets larger, fertility stays approximately within a same range of values. Second, the data are concentrated more on the left hand side of the graph, with couple data fall far away on the right hand side of the graph. Third, as PPgdp increases, the variability of fertility gets smaller and smaller, which means the equal variance assumption for a linear regression model doesn't hold. Therefore, in sum, a simple linear regression model is not a plausible for a summary of this graph.

```
# **********Data visualization and pre-processing**********
# **********                2                **********
# In order to get a better fit, we seek to transform the variables.
# What transformations you would take so that a simple linear regression model is proper?
# State why you choose these transformations. Draw the scatter plot of the transformed variables.
# Comment on the plot.

# fix the data to a model
model=lm(y~x)
plot(model)
```



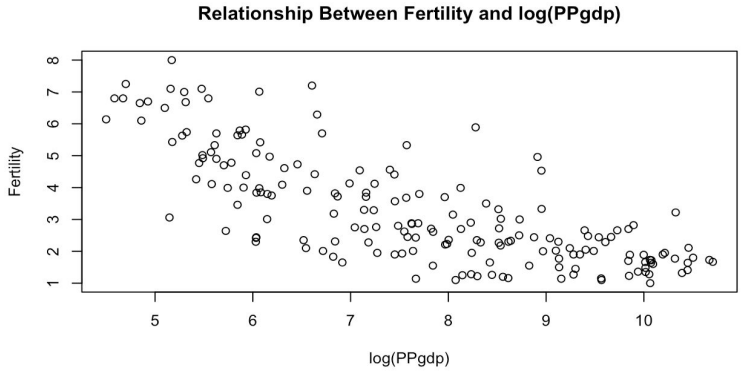
Double check with Diagnostic plot:

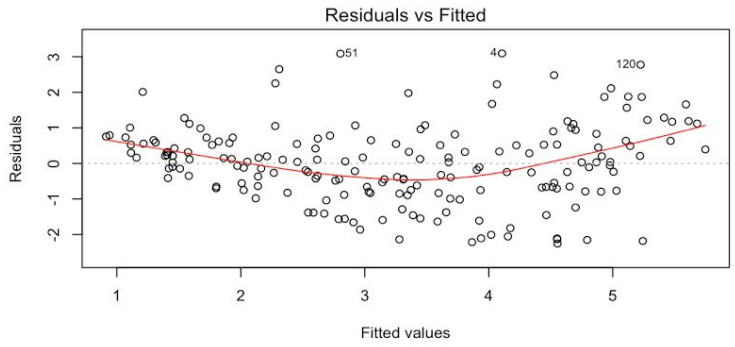
Linearity doesn't hold. Equal variance doesn't hold. Normality approximately looks fine.

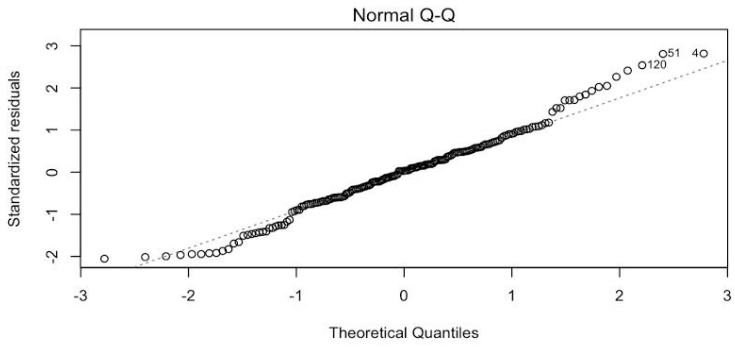
Transformation:

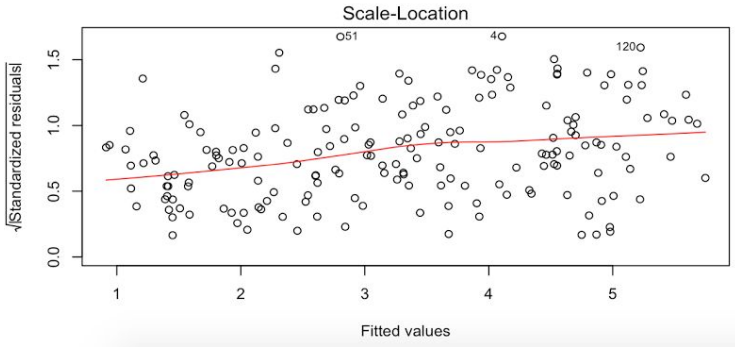
Since the linearity assumption doesn't hold, transform x first. Usually if we see two situations, we try log transformation on x: one is when smaller values(which are close together) are spread further out, the other one is when larger values(which are spread out) are brought closer together. Here we can see in the scatter plot, data points distribute in a pattern that is similar to the first situation, so I would transform x to log(x).

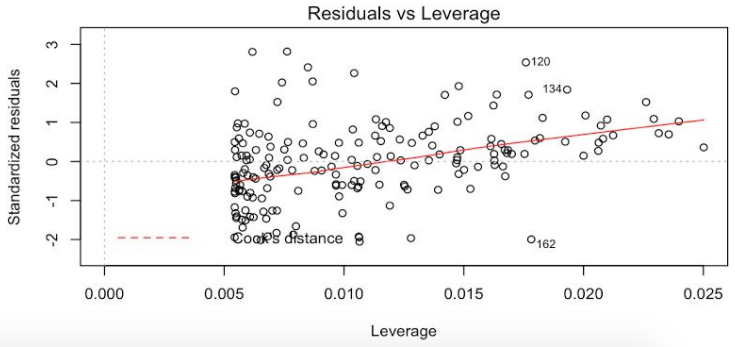
```
xtrans1=log(x)
plot(xtrans1, y, xlab = "log(PPgdp)", ylab = "Fertility", main = "Relationship Between Fertility and log(PPgdp)")
modeltrans1=lm(y~xtrans1)
par(mfrow=c(2,2))
plot(modeltrans1)
```







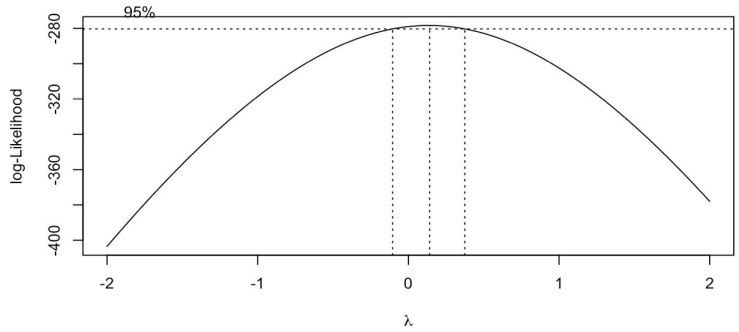




Comments & Transformation:

After transforming x to log(x), we can see from the scatter plot, the data points distribute more evenly along a likely linear pattern, yet the variability of y changes at different log(x); and from Residuals vs Fitted plot, we can see the linearity is fixed more than before, but it still has a curve, meaning that linearity looks better but still need to be fixed. In Scale-Location plot, we see that the reference line has curve and the points are not equally evenly spread along with it, meaning that we have to fix equal variance. In Normal Q-Q plot, both tails are heavier than the normal tails, and the reference line in Residuals vs Leverage is tilted, meaning we need to concern the normality assumption. Since we need to fix the equal variance, and transforming on x won't work for this purpose, I will take transformation on y.

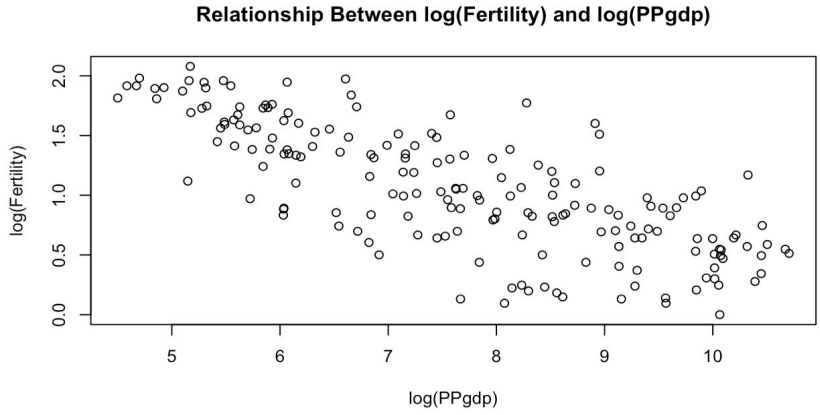
```
# use boxcox() to check what to take on y transformation
boxcox(modeltrans1)
```

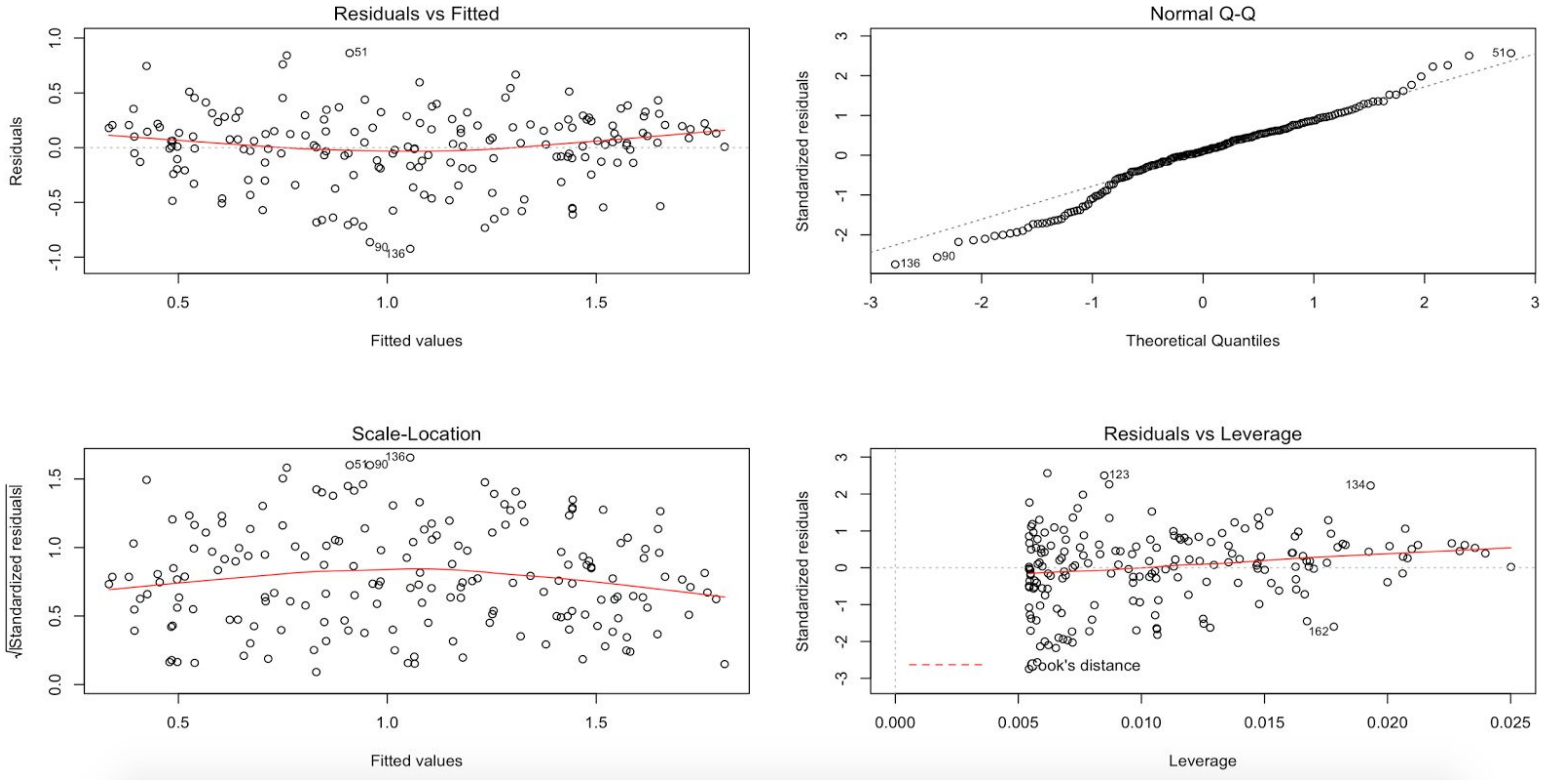


Comments & Transformation:

Based on the boxcox() result, the best value of λ is approximately equal to 0, so I would transform y to log(y).

```
ytrans1 = log(y)
plot(xtrans1, ytrans1, xlab = "log(PPgdp)", ylab = "log(Fertility)", main = "Relationship Between log(Fertility) and
log(PPgdp)")
modeltrans2=lm(ytrans1~xtrans1)
par(mfrow=c(2,2))
plot(modeltrans2)
```





Comments:

From the scatter plot, we can see that data points now spread more evenly in a linear shape. log(y) decreases as log(x) increases from the upper left to the lower right. In the Residual vs Fitted plot, data points are approximately spread along with the flat reference line, which means that the **linearity holds**. Also, in the Scale-Location plot, now the reference line is approximately flat and data points are likely spread along the line with no pattern, meaning the **equal variance holds**. However, from the Normal Q-Q plot, we see that both tails are heavier than the normal tails, and in the Residuals vs Leverage plot, data points seem concentrate around the little-tilted reference line, meaning that after the transformations the normality of new model holds better, but still, we have concerns on the heavier tails. We would say that **normality approximately holds but we keep the concerns**.

```
# **********Model fitting and diagnostic**********
# Fit the simple linear model on the transformed data through three ways.
# Report the least square estimates for the coefficients and R2.
# Add the fitted line to the scatter plot on the transformed data and comment on the fit.
# **********                3a                **********
# Plain coding (not using the 'lm' function or matrix manipulation).

# plain coding
# xbar, ybar
xbar=mean(xtrans1)
ybar=mean(ytrans1)
# SSx, SSxy, b1hat, b0hat, ytrans1hat
SSx=sum((xtrans1-xbar)^2)
SSxy=sum((xtrans1-xbar)*(ytrans1-ybar))
b1hat=SSxy/SSx
b1hat
b0hat=ybar-b1hat*xbar
b0hat
ytrans1hat=b0hat+b1hat*xtrans1
# SSR, SSTO, rsq
SSR=sum((ytrans1hat-ybar)^2)
SSTO=sum((ytrans1-ybar)^2)
rsq=SSR/SSTO
rsq

# add the fitted line
plot(xtrans1, ytrans1, xlab = "Log(PPgdp)", ylab =
"Log(Fertility)", main = "Relationship Between
Log(Fertility) and Log(PPgdp)")
abline(a=b0hat, b=b1hat)
```

```
> # plain coding
> # xbar, ybar
> xbar=mean(xtrans1)
> ybar=mean(ytrans1)
> # SSx, SSxy, b1hat, b0hat, ytrans1hat
> SSx=sum((xtrans1-xbar)^2)
> SSxy=sum((xtrans1-xbar)*(ytrans1-ybar))
> b1hat=SSxy/SSx
> b1hat
[1] -0.2374852
> b0hat=ybar-b1hat*xbar
> b0hat
[1] 2.876072
> ytrans1hat=b0hat+b1hat*xtrans1
> # SSR, SSTO, rsq
> SSR=sum((ytrans1hat-ybar)^2)
> SSTO=sum((ytrans1-ybar)^2)
> rsq=SSR/SSTO
> rsq
[1] 0.5813292
> # add the fitted line
> plot(xtrans1, ytrans1, xlab = "Log(PPgdp)", ylab =
"Log(Fertility)", main = "Relationship Between
Log(Fertility) and Log(PPgdp)")
> abline(a=b0hat, b=b1hat)
```

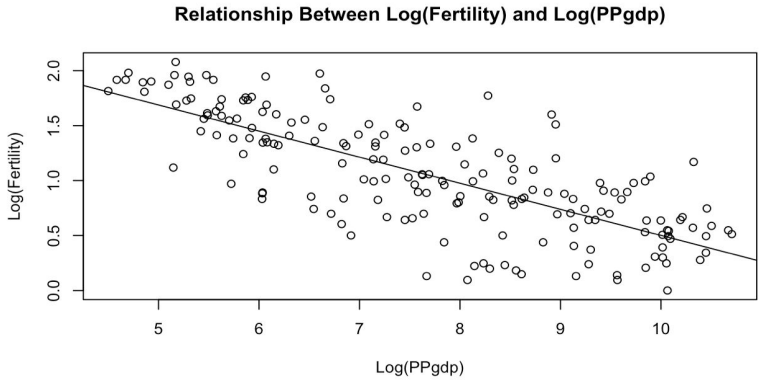
```
# **********Model fitting and diagnostic**********
# **********                3b                **********
# Using the 'lm' function.

#using lm()
modeltrans2=lm(ytrans1~xtrans1)
# get b1hat, b0hat
modeltrans2$coefficients
summary(modeltrans2)$r.squared
# add the fitted line
plot(xtrans1, ytrans1, xlab = "Log(PPgdp)", ylab =
"Log(Fertility)", main = "Relationship Between
Log(Fertility) and Log(PPgdp)")
abline(modeltrans2$coefficients)
```

```
> #using lm()
> modeltrans2=lm(ytrans1~xtrans1)
> # get b1hat, b0hat
> modeltrans2$coefficients
(Intercept)     xtrans1
  2.8760719  -0.2374852
> summary(modeltrans2)$r.squared
[1] 0.5813292
> # add the fitted line
> plot(xtrans1, ytrans1, xlab = "Log(PPgdp)", ylab =
"Log(Fertility)", main = "Relationship Between
Log(Fertility) and Log(PPgdp)")
> abline(modeltrans2$coefficients)
```

```
# **********Model fitting and diagnostic**********
# **********                3c              **********
# Through matrix manipulation.

# xtm: xtrans1 matrix // ytm: ytrans1 matrix
# cbind: combine the columns // rep(1,n): repeat 1 for n
times
xtm = cbind(rep(1,n), xtrans1)
ytm = cbind(ytrans1)

# if a%*%X=b, solve(a,b): solve for X
# if only solve(a): gives the inverse of a
# t(a): gives the transpose of a
# In least square equation, Y=AX+E,
A=[solve(Xt%*%X)]%*%[Xt%*%Y], where Xt is the
transpose of X
# betam: beta matrix // xttm: xtrans1 transpose
xttm = t(xtm)
betam = (solve(xttm%*%xtm))%*%(xttm%*%ytm)
betam

# ythatm: yhat matrix // em: epsilon matrix
ythatm = xtm%*%betam

# find rsq: SSR=sum((ytrans1hat-ybar)^2),
SSTO=sum((ytrans1-ybar)^2), rsq=SSR/SSTO
# rowMeans(a):returns vector of row means
# colMeans(a):returns vector of column means
# rowSums(a): returns vector of row sums
# colSums(a): returns vector of column means
# yhyb: ytrans1hat-ybar // yyb: ytrans1-ybar // ybarm: ybar
matrix
ybarm = colSums(ytm)/n
yhyb = ythatm-ybarm
yyb = ytm-ybarm
SSR = colSums(yhyb*yhyb)
SSTO = colSums(yyb*yyb)
rsq = SSR/SSTO
rsq

# get b0, b1 by betam(row, column) and draw the line
betam[1,1]
betam[2,1]
plot(xtrans1, ytrans1, xlab = "Log(PPgdp)", ylab =
"Log(Fertility)", main = "Relationship Between
Log(Fertility) and Log(PPgdp)")
abline(a=betam[1,1],b=betam[2,1])
```

```
> # xtm: xtrans1 matrix // ytm: ytrans1 matrix
> # cbind: combine the columns // rep(1,n): repeat 1 for n
times
> xtm = cbind(rep(1,n), xtrans1)
> ytm = cbind(ytrans1)
> # if a%*%X=b, solve(a,b): solve for X
> # if only solve(a): gives the inverse of a
> # t(a): gives the transpose of a
> # In least square equation, Y=AX+E,
A=[solve(Xt%*%X)]%*%[Xt%*%Y], where Xt is the
transpose of X
> # betam: beta matrix // xttm: xtrans1 transpose
> xttm = t(xtm)
> betam = (solve(xttm%*%xtm))%*%(xttm%*%ytm)
> betam
             ytrans1
           2.8760719
xtrans1 -0.2374852
> # ythatm: yhat matrix // em: epsilon matrix
> ythatm = xtm%*%betam
> # find rsq: SSR=sum((ytrans1hat-ybar)^2),
SSTO=sum((ytrans1-ybar)^2), rsq=SSR/SSTO
> # rowMeans(a):returns vector of row means
> # colMeans(a):returns vector of column means
> # rowSums(a): returns vector of row sums
> # colSums(a): returns vector of column means
> # yhyb: ytrans1hat-ybar // yyb: ytrans1-ybar // ybarm:
ybar matrix
> ybarm = colSums(ytm)/n
> yhyb = ythatm-ybarm
> yyb = ytm-ybarm
> SSR = colSums(yhyb*yhyb)
> SSTO = colSums(yyb*yyb)
> rsq = SSR/SSTO
> rsq
  ytrans1
0.5813292
> # get b0, b1 by betam(row, column) and draw the line
> betam[1,1]
[1] 2.876072
> betam[2,1]
[1] -0.2374852
> plot(xtrans1, ytrans1, xlab = "Log(PPgdp)", ylab =
"Log(Fertility)", main = "Relationship Between
Log(Fertility) and Log(PPgdp)")
> abline(a=betam[1,1],b=betam[2,1])
```

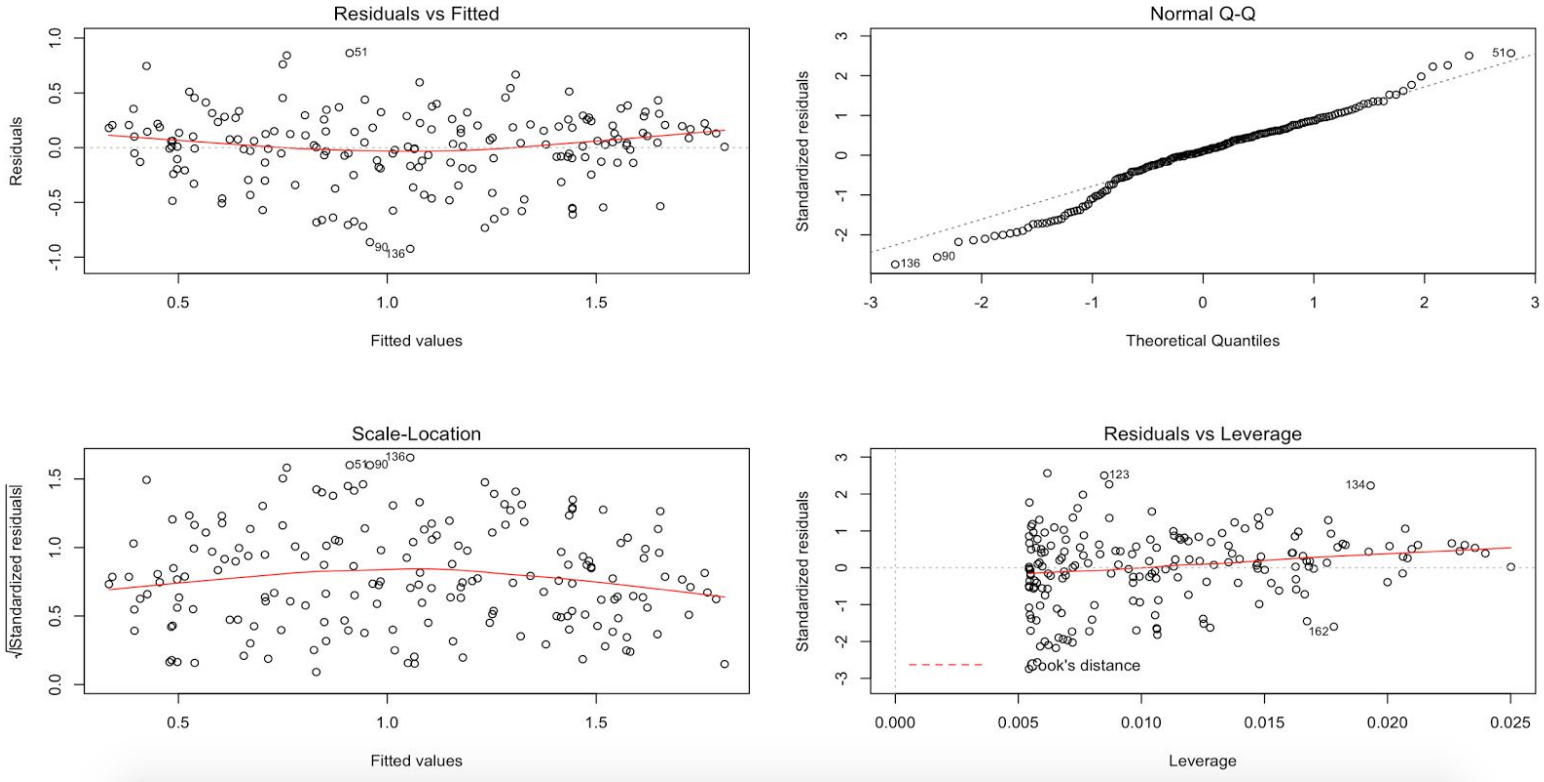
Scatter plot with the fitted linear regression line:



Comment on the fit:

The fitted line approximately fits this model. Data points fall approximately evenly on or around the straight fitted line. As the predictors increases, the variability of the corresponding response mostly remains unchanged.

```
# **********Model fitting and diagnostic**********
# **********                4                **********
# Draw the diagnostic plots and comment.
```



Comments:

From the scatter plot, we can see that data points now spread more evenly in a linear shape. log(y) decreases as log(x) increases from the upper left to the lower right. In the Residual vs Fitted plot, data points are approximately spread along with the flat reference line, which means that the **linearity holds**. Also, in the Scale-Location plot, now the reference line is approximately flat and data points are likely spread along the line with no pattern, meaning the **equal variance holds**. However, from the Normal Q-Q plot, we see that both tails are heavier than the normal tails, and in the Residuals vs Leverage plot, the Cook's distance lines are not shown on the plot, meaning all the data points are within the reasonable normal distributed range; data points seem concentrate around the little-tilted reference line. After the transformation on y the normality of new model holds better, but still, we have concerns on the heavier tails. We would say that **normality approximately holds with concerns**.

**********Making inferences based on the model**********
********** 5 **********
Test whether there is a linear relationship between the transformed variables at 0.05 significance level.
Null hypothesis: H_0: $\beta_1 = 0$
Alternative hypothesis: H_A: $\beta_1 \neq 0$

```
# calculate T*=abs(b1hat/SEb1hat)
# known b1hat, b0hat from previous calculation
b1hat
b0hat
# n, e, SSE, MSE, sigmahat
n=dim(dataUN)[1]
n
e=ytrans1-ytrans1hat
SSE=sum(e^2)
MSE=SSE/(n-2)
sigmahat=sqrt(MSE)

# standard error for b1hat
SEb1hat=sigmahat/sqrt(SSx)
SEb1hat
# standard error for b0hat
SEb0hat=sqrt(MSE*((1/n)+((xbar^2)/SSx)))
SEb0hat
b1hat/SEb1hat
abs(b1hat/SEb1hat)
# check with summary()
summary(modeltrans2)
```

```
> # calculate T*=abs(b1hat/SEb1hat)
> # known b1hat, b0hat from previous calculation
> b1hat
[1] -0.2374852
> b0hat
[1] 2.876072
> # n, e, SSE, MSE, sigmahat
> n=dim(dataUN)[1]
> n
[1] 184
> e=ytrans1-ytrans1hat
> SSE=sum(e^2)
> MSE=SSE/(n-2)
> sigmahat=sqrt(MSE)
> # standard error for b1hat
> SEb1hat=sigmahat/sqrt(SSx)
> SEb1hat
[1] 0.01493916
> # standard error for b0hat
> SEb0hat=sqrt(MSE*((1/n)+((xbar^2)/SSx)))
> SEb0hat
[1] 0.1171479
> b1hat/SEb1hat
[1] -15.89683
> abs(b1hat/SEb1hat)
[1] 15.89683
> # check with summary()
> summary(modeltrans2)

Call:
lm(formula = ytrans1 ~ xtrans1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92398 -0.16996  0.03671  0.20633  0.86331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.87607    0.11715   24.55   <2e-16 ***
xtrans1     -0.23749    0.01494  -15.90   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3377 on 182 degrees of freedom
Multiple R-squared:  0.5813,  Adjusted R-squared:  0.579
F-statistic: 252.7 on 1 and 182 DF,  p-value: < 2.2e-16
```

Test statistic: $T^*=-15.89683$
Its null distribution follows the $T_{n-2} = T_{182}$ distribution.
Since the significance level $\alpha=0.05$, we have:
$1-\alpha/2=0.975$
From the T-table, $T_{182,\,0.975} = 1.9731$

The decision rule is to reject H_0 if $|T^*| > T_{182,\,0.975}$.
Since $T_{182,\,0.975}= 1.9731$, $|T^*|=15.89683 > T_{182,\,0.975}$,
so we reject the null hypothesis and conclude there is a linear association between the transformed variables at the α-level 0.05.

**********Making inferences based on the model**********
********** 6 **********
Provide a 99% confidence interval on the expected Fertility for a region with PPgdp 20,000 US dollars in 2001.
Comment: (x: PPgdp. y:Fertility)
The simple linear regression model is: ytrans1 = b0hat + b1hat * xtrans1 + ε, which is:

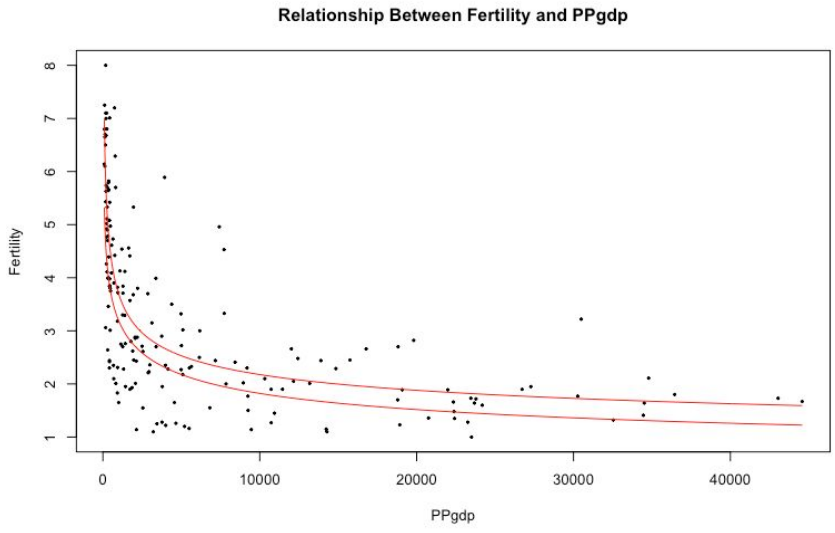
log(y) = b0hat + b1hat * log(x) + ε

```
a=1-0.99
1-a/2
xnew=20000
xtrans1new=log(xnew)
ytrans1new=b0hat+b1hat*xtrans1new


# SEytrans1new
SEytrans1new =
sqrt(MSE*((1/n)+((xtrans1new-xbar)^2/SSx)))
SEytrans1new
# lbytrans1new, ubytrans1new
lbytrans1new=ytrans1new-qt(1-a/2,n-2)*SEytrans1ne
w
lbytrans1new
ubytrans1new=ytrans1new+qt(1-a/2,n-2)*SEytrans1n
ew
ubytrans1new

# calculate the lb, ub for y
lb=exp(lbytrans1new)
lb
ub=exp(ubytrans1new)
ub
```

```
> a=1-0.99
> 1-a/2
[1] 0.995
> xnew=20000
> xtrans1new=log(xnew)
> ytrans1new=b0hat+b1hat*xtrans1new
> # SEytrans1new
> SEytrans1new =
sqrt(MSE*((1/n)+((xtrans1new-xbar)^2/SSx)))
> SEytrans1new
[1] 0.04171823
> # lbytrans1new, ubytrans1new
>
lbytrans1new=ytrans1new-qt(1-a/2,n-2)*SEytrans1ne
w
> lbytrans1new
[1] 0.4155429
>
ubytrans1new=ytrans1new+qt(1-a/2,n-2)*SEytrans1n
ew
> ubytrans1new
[1] 0.6327373
> # calculate the lb, ub for y
> lb=exp(lbytrans1new)
> lb
[1] 1.515193
> ub=exp(ubytrans1new)
> ub
[1] 1.882757
```

Therefore, the 99% confidence interval on the expected Fertility for a region with PPgdp 20,000 US dollars in 2001 is: **[1.515193, 1.882757]**.

```
# **********Making inferences based on the model**********
# **********                    7                    **********
# Provide a 95% confidence band for the relation between the expected Fertility and PPgdp.
# Add the bands to the scatter plot of the original data.

W=sqrt(2*qf(.95, df1=2, df2=n-2))

yhat=function(x){ b0hat+b1hat*x }
sigmahat2=(summary(modeltrans2)$sigma)^2
xbar=mean(xtrans1)
ybar=mean(ytrans1)
SSx=sum((xtrans1-xbar)^2)
seyhat=function(xh){ sqrt(sigmahat2*(1/n + (xh-xbar)^2/SSx)) }

cband=function(xvec){
  d=length(xvec)
  CIs=matrix(0,d,2)
  colnames(CIs)=c("Lower", "Upper")

  for(i in 1:d){  CIs[i,]=c(yhat(x=xvec[i])+c(-1,1)*W*seyhat(xh=xvec[i]))  }
  as.data.frame(CIs)
}

# get the range of x, create the bands
maxx=max(x)
minx=min(x)
range = seq(from=minx, to=maxx, length.out=1000)
bands = cband(log(range))

# draw bands on the original plot
plot(x, y, xlab = "PPgdp", ylab = "Fertility", main = "Relationship Between Fertility and PPgdp", pch=20, cex=.5)
points(range, exp(bands$Lower), col = "red", type = "l")
points(range, exp(bands$Upper), col = "red", type = "l")
```



**********Making inferences based on the model**********
********** 8 **********
Assuming that the same relationship between Fertility and PPgdp holds,
give a 99% prediction interval on Fertility for a region with PPgdp 25,000 US dollars in 2018.
Comment: (x: PPgdp. y:Fertility)
The simple linear regression model is: ytrans1 = b0hat + b1hat * xtrans1 + ε, which is:

log(y) = b0hat + b1hat * log(x) + ε

```
a=1-0.99
1-a/2
xpred=25000
xtrans1pred=log(xpred)
ytrans1pred=b0hat+b1hat*xtrans1pred

# SEytrans1pred
SEytrans1pred =
sqrt(MSE*(1+(1/n)+((xtrans1pred-xbar)^2/SSx)))
# lbytrans1pred, ubytrans1pred
lbytrans1pred=ytrans1pred-qt(1-a/2,n-2)*SEytrans1p
red
lbytrans1pred
ubytrans1pred=ytrans1pred+qt(1-a/2,n-2)*SEytrans1
pred
ubytrans1pred

# calculate the lbpred, ubpred for y
lbpred=exp(lbytrans1pred)
lbpred
ubpred=exp(ubytrans1pred)
ubpred
```

```
> a=1-0.99
> 1-a/2
[1] 0.995
> xpred=25000
> xtrans1pred=log(xpred)
> ytrans1pred=b0hat+b1hat*xtrans1pred
> # SEytrans1pred
> SEytrans1pred =
sqrt(MSE*(1+(1/n)+((xtrans1pred-xbar)^2/SSx)))
> # lbytrans1pred, ubytrans1pred
>
lbytrans1pred=ytrans1pred-qt(1-a/2,n-2)*SEytrans1p
red
> lbytrans1pred
[1] -0.415426
>
ubytrans1pred=ytrans1pred+qt(1-a/2,n-2)*SEytrans1
pred
> ubytrans1pred
[1] 1.35772
> # calculate the lbpred, ubpred for y
> lbpred=exp(lbytrans1pred)
> lbpred
[1] 0.6600591
> ubpred=exp(ubytrans1pred)
> ubpred
[1] 3.887319
```

Therefore, the 99% prediction interval on Fertility for a region with PPgdp 25,000 US dollars in 2018 is:
[0.6600591, 3.887319].

**********Making inferences based on the model**********
********** 9 **********
Based on the diagnostic plots in Part 4, do you have any concern on the above hypothesis testing and inferences?
If so, what are the concerns?

Yes, based on the diagnostic plots in Part 4, I have some concerns. In the Residuals vs Fitted plot, data points distribute along the flat reference line, meaning **linearity assumption** holds. In the Scale-Location plot, data points evenly distribute along the flat reference line with no pattern, meaning **equal variance assumption** holds. However, in the Normal Q-Q plot, both two tails are heavier than the normal tails in the Normal Q-Q plot and in the Residuals vs Leverage plot, data points seem concentrate around the little-tilted reference line, meaning **normality assumption** may hold or may not, depending on the skewed tails. I concern that the both sides heavier tails may affect the data distribution and doesn't lead to a normal distribution for the new model with transformed variables. And therefore, the hypothesis test, confidence interval and prediction interval above may not be accurate, and further transformation or different model fitting may be needed.