HW 12

Vishal Arora

November 17, 2019

The attached who.csv dataset contains real-world data from 2008. The variables included follow.

Country: name of the country
LifeExp: average life expectancy for the country in years
InfantSurvival: proportion of those surviving to one year or more
Under5Survival: proportion of those surviving to five years or more
TBFree: proportion of the population without TB.
PropMD: proportion of the population who are MDs
PropRN: proportion of the population who are RNs
PersExp: mean personal expenditures on healthcare in US dollars at average exchange rate
GovtExp: mean government expenditures per capita on healthcare, US dollars at average exchange rate
TotExp: sum of personal and government expenditures.

```
# read data using read.csv function from github.
who_df <- read.csv("https://raw.githubusercontent.com/Vishal0229/Data605/master/Week12/who.csv", header

knitr::kable(head(who_df))
```

Country	LifeExp	InfantSurvival	Under5Survival	TBFree	PropMD	PropRN	PersExp	GovtEx
Afghanistan	42	0.835	0.743	0.99769	0.0002288	0.0005723	20	9
Albania	71	0.985	0.983	0.99974	0.0011431	0.0046144	169	312
Algeria	71	0.967	0.962	0.99944	0.0010605	0.0020914	108	518
Andorra	82	0.997	0.996	0.99983	0.0032973	0.0035000	2589	16972
Angola	41	0.846	0.740	0.99656	0.0000704	0.0011462	36	162
Antigua and Barbuda	73	0.990	0.989	0.99991	0.0001429	0.0027738	503	1254

```
#Summarizing the data before treating for any missing/null values in dataset.
summary(who_df)
```

```
##                 Country       LifeExp      InfantSurvival
##  Afghanistan        :  1   Min.   :40.00   Min.   :0.8350
##  Albania            :  1   1st Qu.:61.25   1st Qu.:0.9433
##  Algeria            :  1   Median :70.00   Median :0.9785
##  Andorra            :  1   Mean   :67.38   Mean   :0.9624
##  Angola             :  1   3rd Qu.:75.00   3rd Qu.:0.9910
##  Antigua and Barbuda:  1   Max.   :83.00   Max.   :0.9980
##  (Other)            :184
##  Under5Survival       TBFree           PropMD             PropRN
##  Min.   :0.7310   Min.   :0.9870   Min.   :0.0000196   Min.   :0.0000883
##  1st Qu.:0.9253   1st Qu.:0.9969   1st Qu.:0.0002444   1st Qu.:0.0008455
##  Median :0.9745   Median :0.9992   Median :0.0010474   Median :0.0027584
##  Mean   :0.9459   Mean   :0.9980   Mean   :0.0017954   Mean   :0.0041336
##  3rd Qu.:0.9900   3rd Qu.:0.9998   3rd Qu.:0.0024584   3rd Qu.:0.0057164
##  Max.   :0.9970   Max.   :1.0000   Max.   :0.0351290   Max.   :0.0708387
##
##     PersExp           GovtExp             TotExp
```

```
##  Min.   :   3.00   Min.   :    10.0   Min.   :    13
##  1st Qu.:  36.25   1st Qu.:   559.5   1st Qu.:   584
##  Median : 199.50   Median :  5385.0   Median :  5541
##  Mean   : 742.00   Mean   : 40953.5   Mean   : 41696
##  3rd Qu.: 515.25   3rd Qu.: 25680.2   3rd Qu.: 26331
##  Max.   :6350.00   Max.   :476420.0   Max.   :482750
##
```

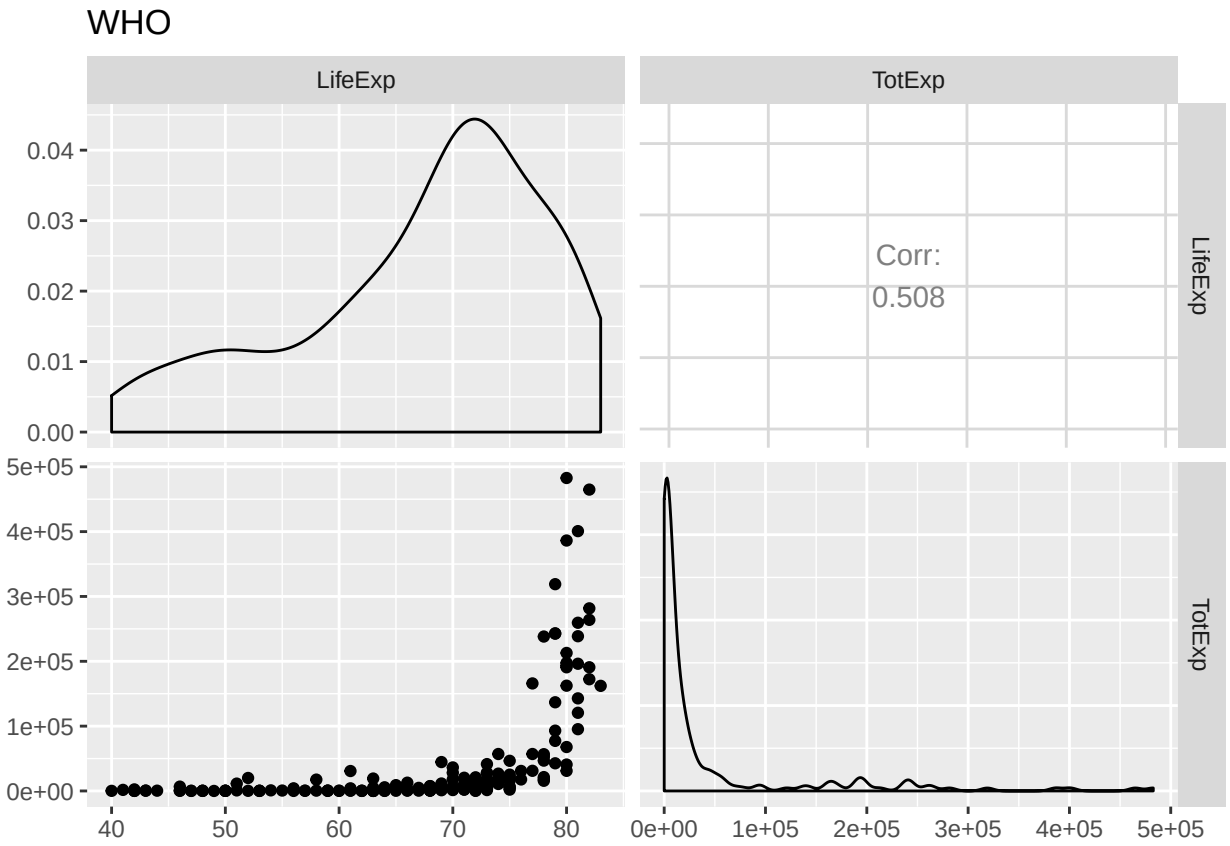
```
## Let's check for any missing values in the data
colSums(is.na(who_df))
```

```
##        Country        LifeExp InfantSurvival Under5Survival         TBFree
##              0              0              0              0              0
##         PropMD         PropRN        PersExp        GovtExp         TotExp
##              0              0              0              0              0
```

It looks that there no NULL values in our datase, hence we are good to use the dataset as it is.

1) Provide a scatterplot of LifeExp~TotExp, and run simple linear regression. Do not transform the variables. Provide and interpret the F statistics, R^2, standard error,and p-values only. Discuss whether the assumptions of simple linear regression met.

```
## using ggpairs function of GGally package
df <- select(who_df,"LifeExp","TotExp")
ggpairs(df,columns=1:2,title="WHO")
```



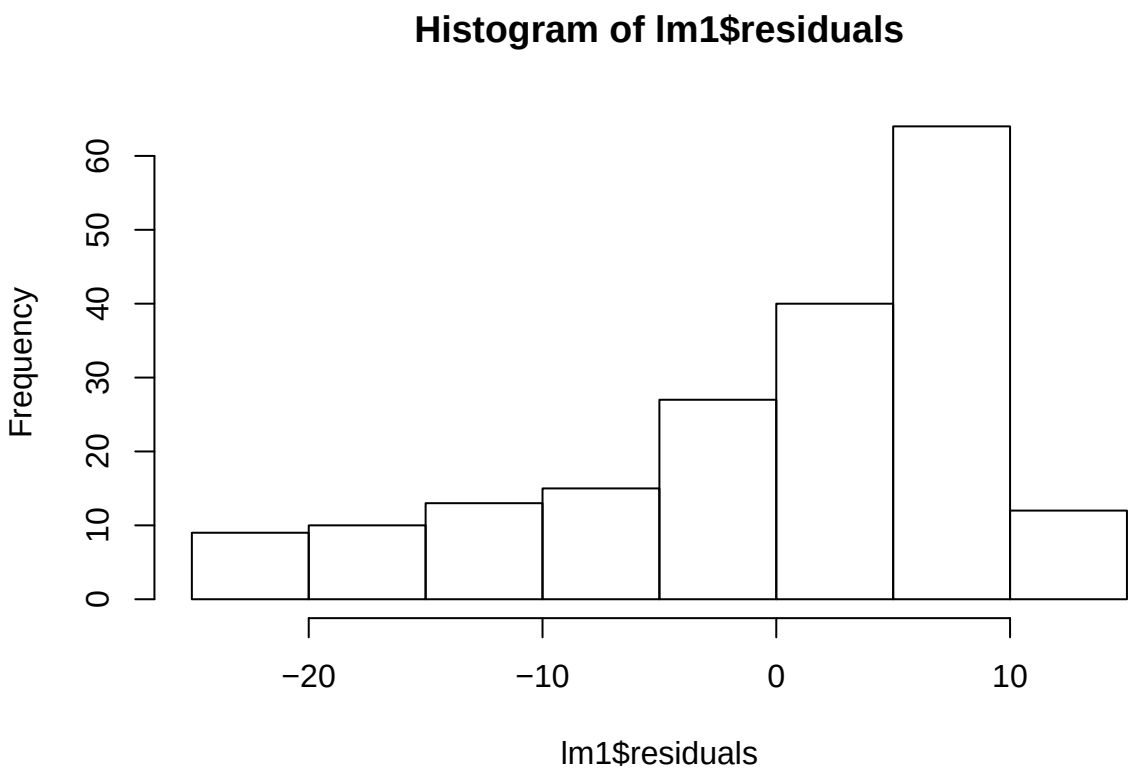
From the above diagram we can see that the correleation between LifeExp & TotExp variables is 0.508 which means that the 2 variables don't have a strong relationship between themselves but it is ok hence we can

check if by adding any other variable the correlation increases. Scatter lot tells us that the relationship is not linear between the 2 variables.

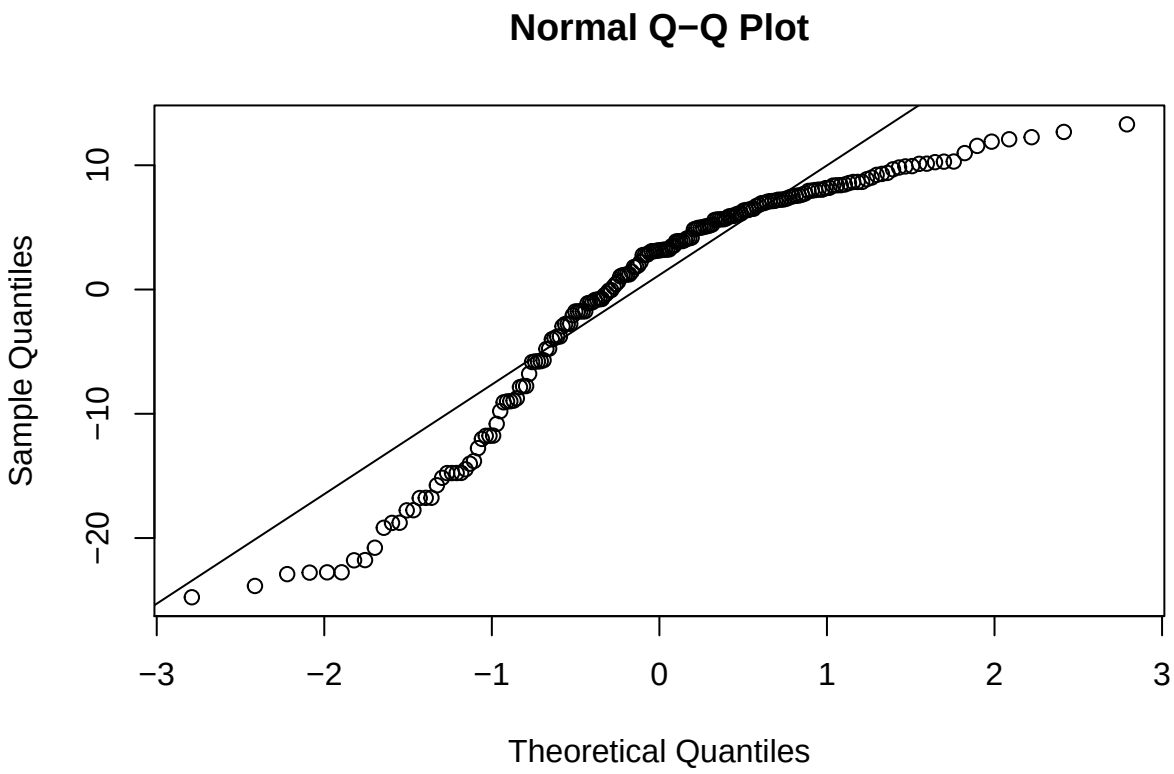
```
lm1 <- lm(LifeExp ~ TotExp,data = df)
summary(lm1)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      6.297e-05  7.795e-06   8.079 7.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.371 on 188 degrees of freedom
## Multiple R-squared:  0.2577, Adjusted R-squared:  0.2537
## F-statistic: 65.26 on 1 and 188 DF,  p-value: 7.714e-14
```

```
#par(mfrow=c(2,2))
#plot(lm1)
hist(lm1$residuals)
```




```
qqnorm(lm1$residuals);
qqline(lm1$residuals)
```



F-Statistic:-
This test statistic tells us if there is a relationship between the dependent and independent variables we are testing. Generally, a large F indicates a stronger relationship.In our case the value is 65.26 which not too good not too bad but signifies that there is a relation between the variables.

R^2 :-
The R2 value is a measure of how close our data are to the linear regression model. R2 values are always between 0 and 1; numbers closer to 1 represent well-fitting models. R2 always increases as more variables are included in the model, and so adjusted R2 is included to account for the number of independent variables used to make the model. In our case the value is only 0.2577 which tells that the models accounts for only 25.77% of varition in the data and it might not be a good fit as alone vriable hence we need to find other variable(s) which in conjunction to the said predictor(TotExp) variable can account of higher number of varaiblity in the data.
Std Error :-
The coefficient standard errors tell us the average variation of the estimated coefficients from the actual average of our response variable. Which in our case is way to high.

p-value is used for rejecting or accpeting the Null hypothesis, if we form a hypothesis

$$H_0 \ : \ LfeExp \ \& \ TotExp \ variables \ are \ not \ related \ to \ each \ other.$$

$$H_A \ : \ LfeExp \ \& \ TotExp \ variables \ have \ some \ relation \ with \ each \ other.$$

P-value:-
The larger the t statistic, the smaller the p-value. Generally, we use 0.05 as the cutoff for significance; when

p-values are smaller than 0.05, we reject H0. In our case p-values are smaller than 0.05, hence we can reject Null hypothesis . Thus there is some relationship between the 2 variables.

Lookging at the histogram & QQ Plot the residuals are clearly not normal or close to normal. Thus Thus the assumption is not met.

To check the heteroscedasticity, there are a couple of tests that comes handy to establish the presence or absence of heteroscedasticity - The Breush-Pagan test and the NCV test.

> heteroskedasticity occurs when the variance for all observations in a data set are not the same. Conversely, when the variance for all observations are equal, we call that homoskedasticity . Null hypothesis that heteroskedasticity is not present (i.e. homoskedastic) against the, Alternative hypothesis that heteroskedasticity is present.

```
lmtest::bptest(lm1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm1
## BP = 2.6239, df = 1, p-value = 0.1053
```

```
car::ncvTest(lm1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.599177, Df = 1, p = 0.10692
```

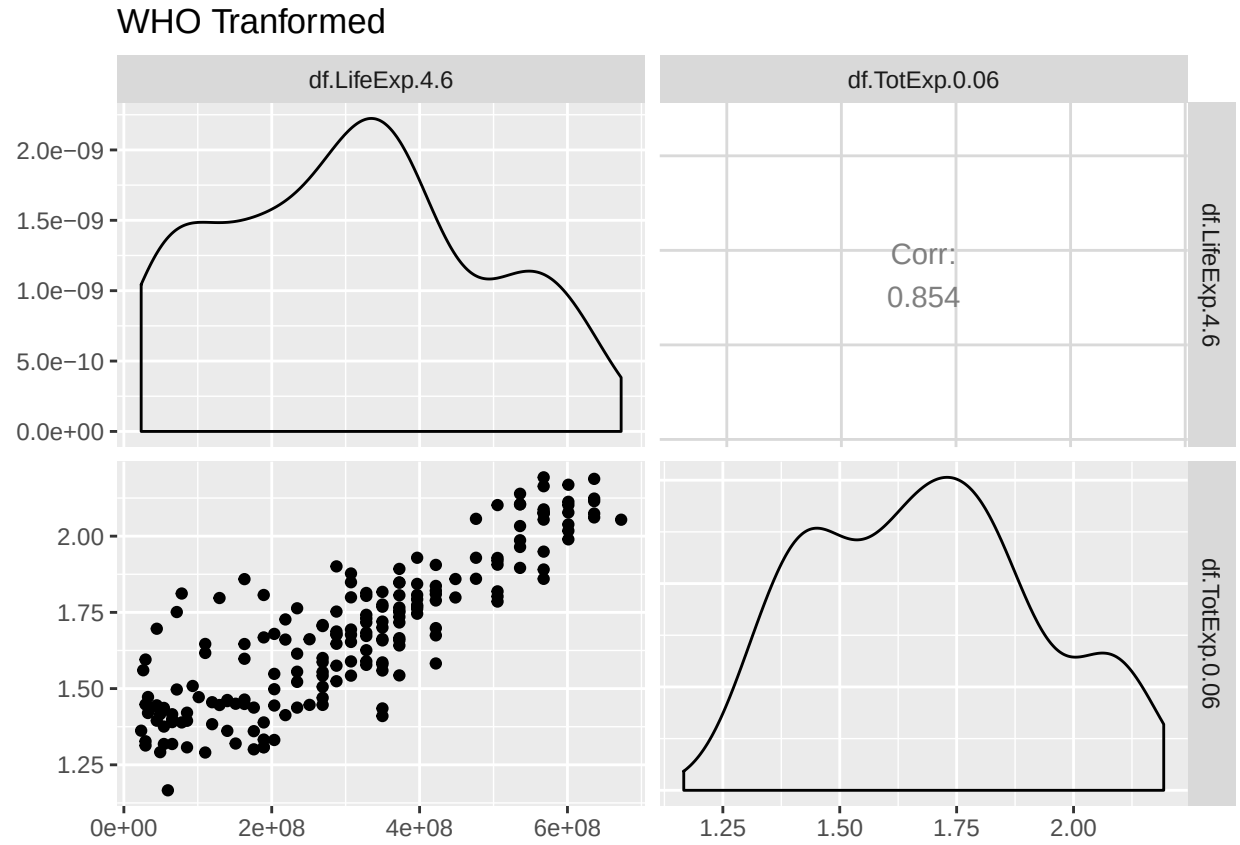
Both these test have a p-value greater that a significance level of 0.05, thus we can reject alternate hypothesis. But looking at the other paramteresthis model can be improved either with the addition of more variables on Indepdendent axis or by trying out tranformation .

Raise life expectancy to the 4.6 power (i.e., LifeExp^4.6). Raise total expenditures to the 0.06 power (nearly a log transform, TotExp^.06). Plot LifeExp^4.6 as a function of TotExp^.06, and r re-run the simple regression model using the transformed variables. Provide and interpret the F statistics, R^2, standard error, and p-values. Which model is "better?"

```
df_trans1 <- data.frame(df$LifeExp^4.6,df$TotExp^0.06)

head(df_trans1)
```

```
##   df.LifeExp.4.6 df.TotExp.0.06
## 1       29305338       1.327251
## 2      327935478       1.625875
## 3      327935478       1.672697
## 4      636126841       2.061481
## 5       26230450       1.560068
## 6      372636298       1.765748
```

```
ggpairs(df_trans1,columns=1:2,title="WHO Tranformed")
```

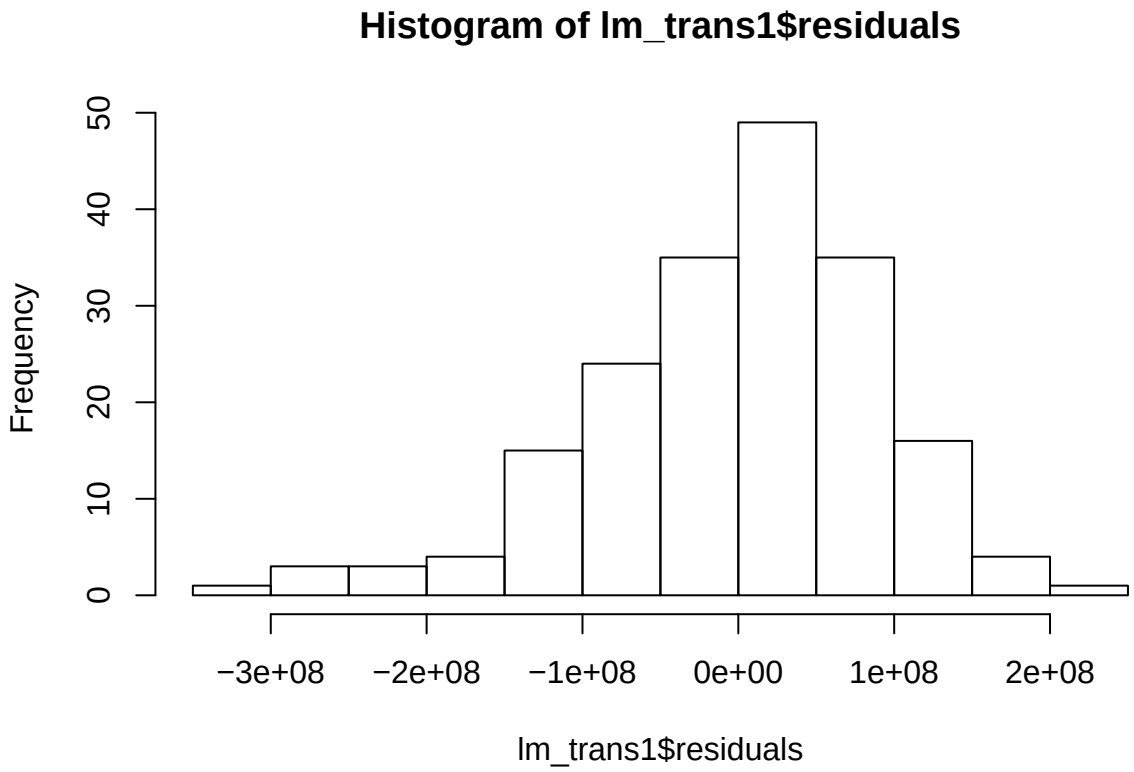


Now we can clearly see that after transfrming the variables the correlation between the variables has increased to 0.854 and also the scatter plot shows a linear relation ship between the 2 tranformed variables.

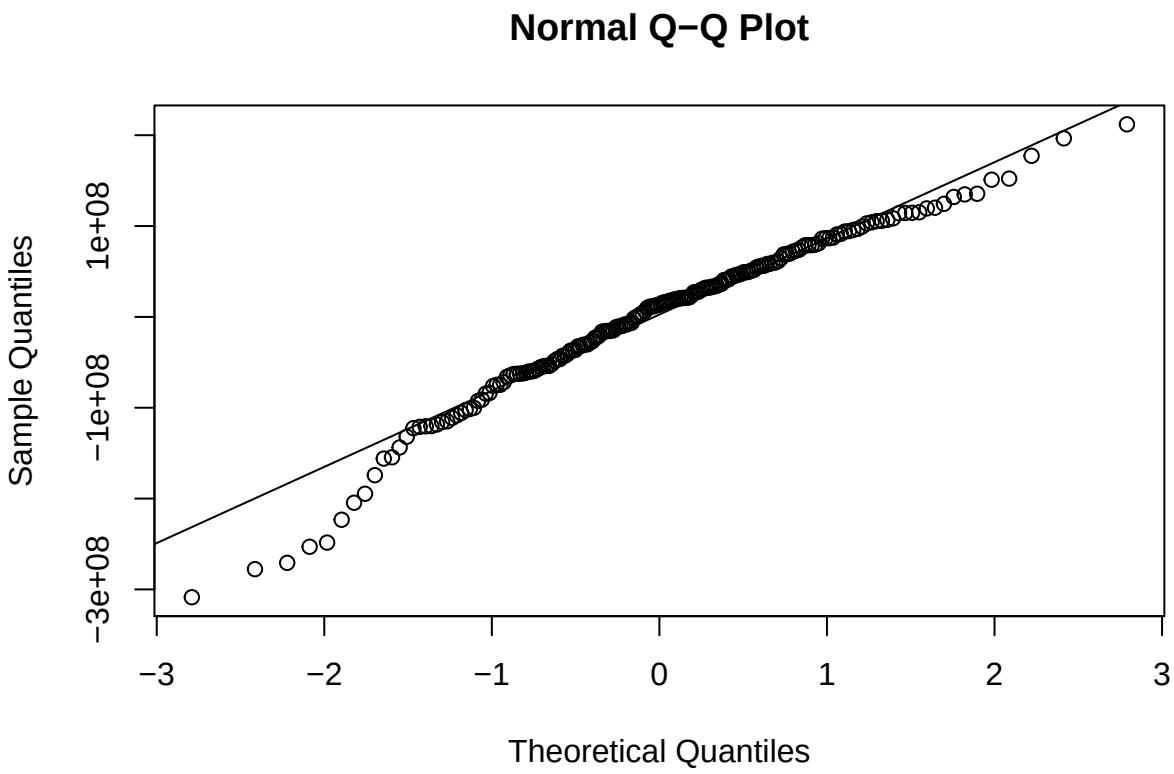
```
lm_trans1 <- lm(df_trans1$df.LifeExp.4.6 ~ df_trans1$df.TotExp.0.06, data=df_trans1)

summary(lm_trans1)
```

```
##
## Call:
## lm(formula = df_trans1$df.LifeExp.4.6 ~ df_trans1$df.TotExp.0.06,
##     data = df_trans1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -736527910   46817945  -15.73   <2e-16 ***
## df_trans1$df.TotExp.0.06    620060216   27518940   22.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 90490000 on 188 degrees of freedom
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.7283
## F-statistic: 507.7 on 1 and 188 DF,  p-value: < 2.2e-16
```

```r
hist(lm_trans1$residuals)
```




```r
qqnorm(lm_trans1$residuals);
qqline(lm_trans1$residuals)
```



F-statistics is 507.7 and adjusted R^2 is 0.7298, P-values both for F-statistics and TotExp_power is less than 0.05. Residual standard error is 90490000 but since variables are rescaled, thus to calculate standard error

```
90490000^(1/4.6)
```

```
## [1] 53.66557
```

the standard error value come out to be 53.66557

Looking at the Histogram , it looks more normal distributed there is slight left skewness which is very minimal in comparson to model1(lm1).

Even the QQ plot shows the same that majority of the dataset in ormally distributed with slight skewness on the left.

Checking the heteroskedasticity for model2 i.e. lm_trans1

```
lmtest::bptest(lm_trans1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_trans1
## BP = 0.28802, df = 1, p-value = 0.5915
```

```
car::ncvTest(lm_trans1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4278017, Df = 1, p = 0.51307
```

In Model 2 also heteroskedasticity is not present as per theBreusch-Pagan & NCV test.

Hence we can say that Model2(lm_trans1) after transformation is a very far improved model in comparison to model1(lm1).

Using the results from 3, forecast life expectancy when TotExp^.06 =1.5. Then forecast life expectancy when TotExp^.06=2.5

```
TransModel3_compute <- function(x)
  {
     # $$y={ \beta  }_{ 0 }\quad +\quad { \beta  }_{ 1 }x\quad +\quad \E .$$

     y <- -736527910 + 620060216 * (x)
      y <- y^(1/4.6)
    print(y)
}


TransModel3_compute(1.5)
```

```
## [1] 63.31153
```

```
TransModel3_compute(2.5)
```

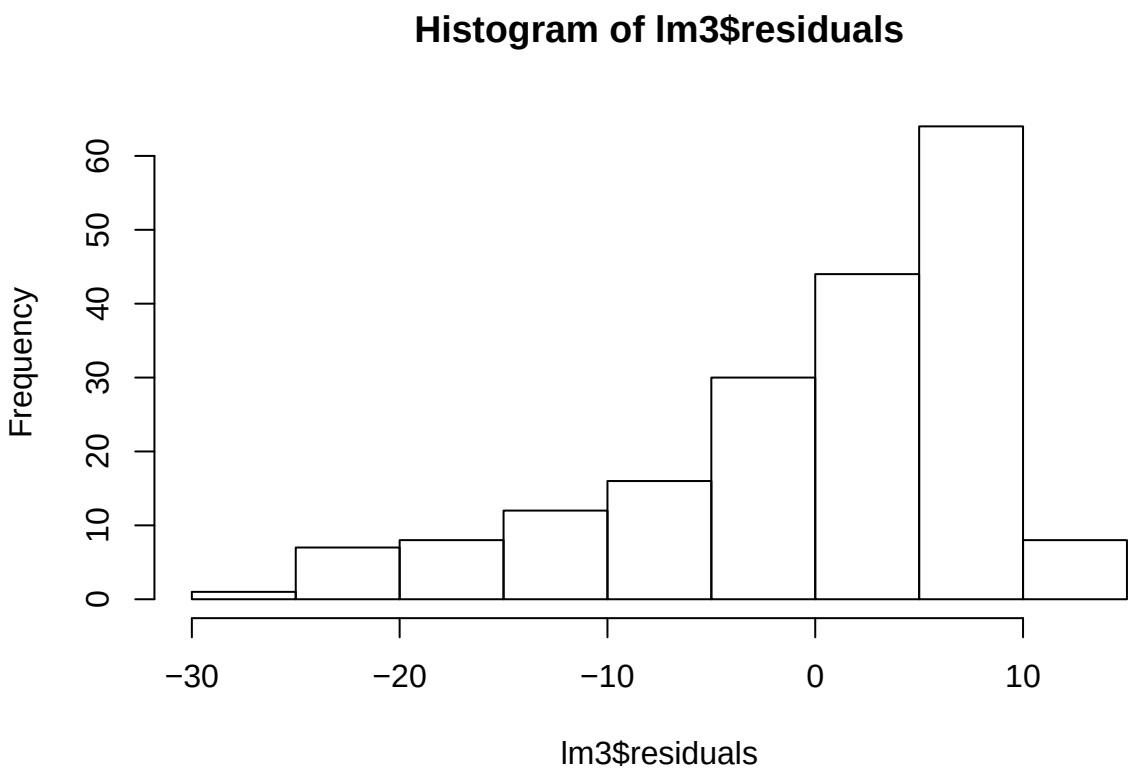
```
## [1] 86.50645
```

Build the following multiple regression model and interpret the F Statistics, R^2, standard error, and p-values. How good is the model? LifeExp = b0+b1 x PropMd + b2 x TotExp +b3 x PropMD x TotExp

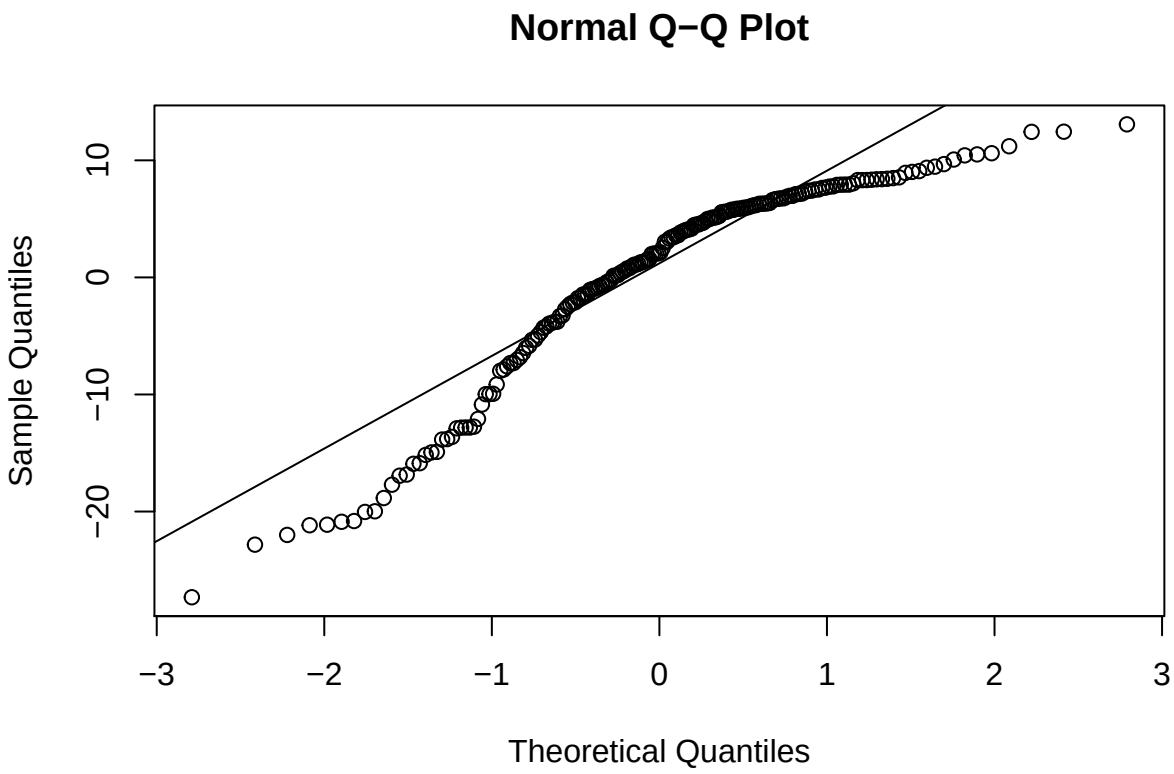
```
lm3 <- lm(LifeExp ~ PropMD+TotExp+(PropMD*TotExp), data=who_df)
summary(lm3)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data = who_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01  78.899  < 2e-16 ***
## PropMD         1.497e+03  2.788e+02   5.371 2.32e-07 ***
## TotExp         7.233e-05  8.982e-06   8.053 9.39e-14 ***
## PropMD:TotExp -6.026e-03  1.472e-03  -4.093 6.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.765 on 186 degrees of freedom
## Multiple R-squared:  0.3574, Adjusted R-squared:  0.3471
## F-statistic: 34.49 on 3 and 186 DF,  p-value: < 2.2e-16
```

```
hist(lm3$residuals);
```




```
qqnorm(lm3$residuals);
qqline(lm3$residuals)
```



The adj R2 accounts for 0.3471 of the variability of the data, which means that only 34% of the variance in the response variable can be explained by the independent variable.Thus means that this model can be improved by either transforming or by finding new predictor variables.

The F-statistic value is quiet less which means that this model is not good for prediction. and p-value indicate that we should reject the null hypothesis (H0), that there isn't a relationship between the variables.

The data does not resemble a normal distribution, as shown in the histogram a hugh left skewness is there and the Q-Q pllots. The residuals do not appear to be centered around 0 from the residual plot.

```
lmtest::bptest(lm_trans1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_trans1
## BP = 0.28802, df = 1, p-value = 0.5915
```

```
car::ncvTest(lm_trans1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4278017, Df = 1, p = 0.51307
```

Rejecting the Null hypothesis, thus there is no heteroskedasticity present in the model.

Forecast LifeExp when PropMD=.03 and TotExp = 14. Does this forecast seem realistic? Why or why not?

```
options(scipen=999)
coef(lm3)
```

```
##      (Intercept)           PropMD           TotExp    PropMD:TotExp
##   62.77270325541 1497.49395251893    0.00007233324   -0.00602568644
```

```
predictMod <- function(x,x1)
  {
  y <- 62.77270325541+1497.49395251893*(x)+(0.00007233324*(x*x1))
  return(y)
}

predictMod(0.03,14)
```

```
## [1] 107.6976
```

This prediction is not a relatistic one, as we know by the initial summary of our WHO dataset the max life expectancy is 83 yrs, whereas as per the new predictions by our Model4 (LifeExp = b0+b1 x PropMd + b2 x TotExp +b3 x PropMD x TotExp) is 107.6976 yrs which is not relaistic. Hence our model is not accurate and needs to be corrected, which can be taken up in next part as how to transform the variables to make the model more effective.

Using log & sqrt to model to see if they improve the ffectiveness.It seems both the tranformation are not accurate hence we might have to use some other techniques for better prediction.

```
lm5 <-  lm(log(LifeExp) ~ log(PropMD)+log(TotExp)+log(PropMD*TotExp), data=who_df)
summary(lm5)
```

```
##
## Call:
## lm(formula = log(LifeExp) ~ log(PropMD) + log(TotExp) + log(PropMD *
##     TotExp), data = who_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34544 -0.03672  0.00996  0.05043  0.21573
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           4.421548   0.089943  49.159 < 0.0000000000000002 ***
## log(PropMD)           0.060750   0.007438   8.168   0.0000000000000458 ***
## log(TotExp)           0.025197   0.004858   5.187   0.0000005530653529 ***
## log(PropMD * TotExp)       NA         NA      NA                    NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1023 on 187 degrees of freedom
## Multiple R-squared:  0.6698, Adjusted R-squared:  0.6663
## F-statistic: 189.7 on 2 and 187 DF,  p-value: < 0.00000000000000022
```

```
lm6 <- lm(sqrt(LifeExp) ~ sqrt(PropMD)+sqrt(TotExp)+sqrt(PropMD*TotExp), data=who_df)
summary(lm6)
```

```
##
```

```
## Call:
## lm(formula = sqrt(LifeExp) ~ sqrt(PropMD) + sqrt(TotExp) + sqrt(PropMD *
##     TotExp), data = who_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20039 -0.23263  0.06744  0.31077  0.72879
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)              7.251198   0.069764 103.939 < 0.0000000000000002
## sqrt(PropMD)            20.929658   2.084802  10.039 < 0.0000000000000002
## sqrt(TotExp)             0.004152   0.000440   9.436 < 0.0000000000000002
## sqrt(PropMD * TotExp) -0.050778   0.007351  -6.907      0.0000000000759
##
## (Intercept)             ***
## sqrt(PropMD)            ***
## sqrt(TotExp)            ***
## sqrt(PropMD * TotExp) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4371 on 186 degrees of freedom
## Multiple R-squared:  0.6052, Adjusted R-squared:  0.5988
## F-statistic: 95.04 on 3 and 186 DF,  p-value: < 0.00000000000000022
```

```
# using Model5 (i.e. lm5 ) which uses log to predict the values
options(scipen=999)
coef(lm5)
```

```
##          (Intercept)          log(PropMD)          log(TotExp)
##           4.42154762           0.06074968           0.02519734
## log(PropMD * TotExp)
##                   NA
```

```
logPredictMod <- function(x,x1)
  {
  y <- 4.42154762+0.06074968*(x)
  return(y)
}

logPredictMod(0.03,14)
```

```
## [1] 4.42337
```