Drivers of life expectancy: a study of public policy and international health

Owen Callen, Vladislav Kozlovsky, and Amy Wen

Stat 109 Final Project May 6, 2019

1 Introduction

1.1 Motivation

We were motivated by a desire to find out what measures could be made from a public health perspective to improve health and well-being in a given population. We were able to find some datasets from the World Health Organization (WHO) and the UN's World Happiness Report (WHR) that contain attributes for various countries that we suspect have an influence on life expectancy, such as GDP, disease burden, and social support. Ideally, we would like to identify predictive attributes for which we could recommend actionable items that would potentially have the most impact. For example, if percentage of DTP3 immunization coverage among one-year-olds is highly predictive of life expectancy, then increasing DTP3 immunization coverage would be a good measure to undertake (although we are aware that correlation does not necessarily imply causation).

1.2 Research Question

Our specific research question is whether or not there exists a relationship between life expectancy and various indicators of physical, economic, and emotional well-being, and the relative importance of any such relationships uncovered.

1.3 Hypotheses

We hypothesize from the list of attributes we are considering that among the top predictors for life expectancy are:

1. **Status:** It is expected that a developed country would generally have a higher life expectancy than a developing country.
2. **Adult Mortality:** A high rate of adult mortality should directly translate to a lower life expectancy.
3. **Immunization:** Rates of immunizations in generally are likely to be highly correlated with each other, but at least one of the immunization categories (HepB, Polio, and Diphtheria) should be predictive of life expectancy, with a higher rate of immunization corresponding with a higher life expectancy.

In addition, we are also looking to test if failing to include social or emotional markers from WHR data results in hidden variable bias. If we determine that point estimates for health and economic indicators are significantly different with and without controlling for social and emotional factors, this would suggest future studies into human well-being and life expectancy would need to control for these factors to obtain valid results.

1.4 Methods and Data Sources

1.4.1 WHO Data

Data was originally obtained from a Kaggle dataset, which combined life expectancy and health data from the WHO Global Health Observatory with economic data from the UN Human Development Reports. As we explain later, PercentExpenditure and BMI were manually replaced due to apparent corruption in the data.

Our dependent variable:

- **LifeExpectancy:** Life Expectancy in years.

21 predictive variables:

- General variables:

1. **Country:** Country
2. **Year:** Year
3. **Status:** Developed or Developing

- Variables related to death/mortality:

4. **AdultMortality:** Probability of dying between 15 and 60 years per 1000 population
5. **InfantDeaths:** Number of Infant Deaths per 1000 population
6. **UnderFiveDeaths:** Number of under-five deaths per 1000 population
7. **HIVAIDS:** Deaths per 1000 live births due to HIV/AIDS (0-4 years)

- Variables related to disease:

8. **Measles:** Number of reported measles cases per 1000 population

- Variables related to vaccinations:

9. **HepB:** Hepatitis B immunization coverage among 1-year-olds (%)
10. **Polio:** Polio (Pol3) immunization coverage among 1-year-olds (%)
11. **Diphtheria:** Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

- Variables related to country resources:

12. **PercentExpenditure:** Expenditure on health as a percentage of GDP per capita (%)
13. **TotalExpenditure:** General government expenditure on health as a percentage of total government expenditure (%)
14. **IncomeComposition:** Human Development Index as a measure of standard of living/income composition of resources (index ranging from 0 to 1)
15. **GDP:** Gross Domestic Product per capita (in USD)

- Variables related to health/nutrition:

16. **Alcohol:** Liters of pure alcohol consumed per capita (15+)
17. **BMI:** Average Body Mass Index of entire population
18. **Thin59:** Prevalence of thinness among children ages 5 to 9 (%)
19. **Thin1019:** Prevalence of thinness among children and adolescents ages 10 to 19 (%)

- Other variables:

20. **Population:** Population of the country
21. **Schooling:** Average expected number of years of schooling

1.4.2 World Happiness Report Data

We noticed a lack of social and emotional indicators in the WHO data, and (as stated above) we wanted to discover if there was a problem with hidden variable bias if we failed to consider these kinds of indicators into our analysis, so we looked for data that might alow us to control for emotional and social factors. We decided to incorporate data from the 2015 World Happiness Report by the United Nations Sustainable Development Solutions Network. The indicators we included:

- **Healthy.life.expectancy.at.birth:** Life Expectancy in years (measured slightly differently than in WHO, more on this in EDA:Merged Data).

22. **Social.support:** Percentage of people who responded positively to the question: "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?". -2015 World Happiness Report, page 22

23. **Freedom.to.make.life.choices:** Percentage of people who responded positively to the question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?". -2015 World Happiness Report, page 22

24. **Generosity:** "Generosity is the residual of regressing the national average of GWP responses to the question "Have you donated money to a charity in the past month?" on GDP per capita." -2015 World Happiness Report, page 22

25. **Perceptions.of.corruption:** "Perceptions of corruption are the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?" " -2015 World Happiness Report, page 22

26. **Positive.affect:** "Positive affect is defined as the average of previous-day affect measures for happiness, laughter and enjoyment for GWP waves 3-7" -2015 World Happiness Report, page 22

27. **Negative.affect:** "Negative affect is defined as the average of previous-day affect measures for worry, sadness and anger for all waves." -2015 World Happiness Report, page 22

28. & 29. **Democratic.Quality, Delivery.Quality:** "[...]a number of studies have divided the World Bank's six main indicators of governmental quality into two groups, with the four indicators for effectiveness, rule of law, quality of regulation, and control of corruption combined to form an index of the quality of delivery, and the two indicators for voice and accountability and for political stability and absence of violence combined to form an index of the democratic quality of government." World Happiness Report

1.4.3 Assumptions, Limitations, and P-values

Except where noted we are depending on the accuracy of the data provided from the WHO and WHR. Much of this data is survey data that may not be reflective of the entire population. There are also quite a number of countries with missing data for certain measures and years, so in the end we limited our final analysis to 103 observations for the year 2011.

Throughout this project we are testing almost 30 variables with only 103 rows of data. Additionally, we are doing multiple hypothesis testing at once. Using the Bonferroni correction of $\alpha * = 0.05/n$, where n is the number of hypothesis tests being performed simultaneously, we estimated that we wanted to go down to an α of 0.01 to test for statistical significance for our backwards stepwise regression rather than $\alpha = 0.05$ by assuming we get a final model that has around 5 variables.

1.4.4 Process

In doing our analysis we generally followed the steps below:

1. Clean data and perform Exploratory Data Analysis.
2. Merge WHO and WHR datasets.
3. Transform variables where appropriate (most often using Box-Cox).
4. Use stepwise regression for variable selection/model building (first without WHR, then with WHR data).
5. Check model validity by verifying all linear model assumptions.
6. Check back on original hypotheses, to see which variables have a statistically significant effect on life expectancy.
7. Use the partial F-test, and examine confidence intervals for coefficient estimates to determine if there is evidence of hidden variable bias when not considering social and emotional factors.

2 Results

2.1 Summary of Results

2.1.1 WHO Only Model

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	56.53875	3.11419	18.155	< 2e-16 ***
adult.mortality.sqrt	-0.34323	0.09851	-3.484	0.000739 ***
hivaids.1oversqrt	3.70428	0.49332	7.509	2.79e-11 ***
thin.10.19.power0.15	-2.05804	0.53362	-3.857	0.000206 ***
schooling	0.92519	0.19518	4.740	7.24e-06 ***

All modeling assumptions held for this model.

2.1.2 WHO and WHR

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	52.5070	3.5138	14.943	< 2e-16 ***
adult.mortality.sqrt	-0.3008	0.0873	-3.445	0.000847 ***
infant.deaths.shift1.log	-0.9627	0.2488	-3.868	0.000200 ***
hivaids.1oversqrt	4.4092	0.4028	10.946	< 2e-16 ***
positive.affect	13.7794	3.3446	4.120	8.04e-05 ***
negative.affect	13.9296	4.9942	2.789	0.006373 **
delivery.quality	2.6422	0.5497	4.806	5.68e-06 ***

All modeling assumptions held for this model.

The only variable we hypothesized would be a key predictor that ended up in our final model was adult mortality. However, the group of predictors we did end up with still make intuitive sense.

In our partial F test, we discovered that we could not simply drop all emotional and social markers ($p << .01$).

While our models both with and without emotional factors do contain some of same predictors, we also observe some notable changes:

1. The original model used the prevalence of thinness in individuals aged 10-19, while the new model does not.
2. The original model used the number of years of schooling, while the new model does not.
3. The original model did not find infant deaths to be statistically significant (using our reduced

alpha value), but it does appear to be an important factor in the new model.

4. In terms of social and emotional variables, the delivery quality of government services, reporting feeling positive emotions, and reporting feeling negative emotions (this was marginal) were all statistically significant when predicting life expectancy.

We did have some surprises when it comes to interpretation of variables. In the model that included emotional markers, both positive and negative emotions were positively correlated with longer life expectancy. This was initially surprising to our team, however, this may indicate that a lack of emotions is associated with shorter life expectancy. It should be noted that as the prevalence of HIV/AIDS goes up, we expect life expectancy to go down. At the same time, our transformed variable hivaids.1oversqrt goes down, so the positive coefficient on hivaids.1oversqrt is not surprising.

As for the point estimates for the indicators that we did keep between the 2 models, confidence intervals for the point estimates for both models contained the point estimates for the other models, so we did not observe significant evidence of hidden variable bias in the variables that we retained. However, the fact that we did not end up with the same set of variables indicates that failing to control for emotional markers in health research may result in radically different findings. We would recommend that future health research control for both social and emotional factors.

## Full Results

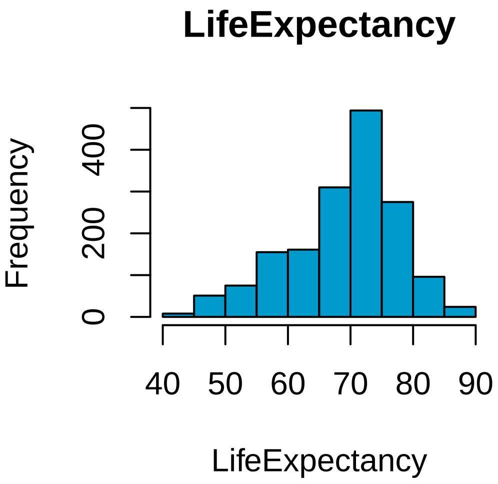
2.1.3 Exploratory Data Analysis

We examined the data from the two datasets we wanted to combine. One contained health and economic well-being indicators from the WHO Global Health Observatory and the UN Human Development Reports (shortened to WHO data), while another contained emotional well-being indicators from the World Happiness Report (WHR data).

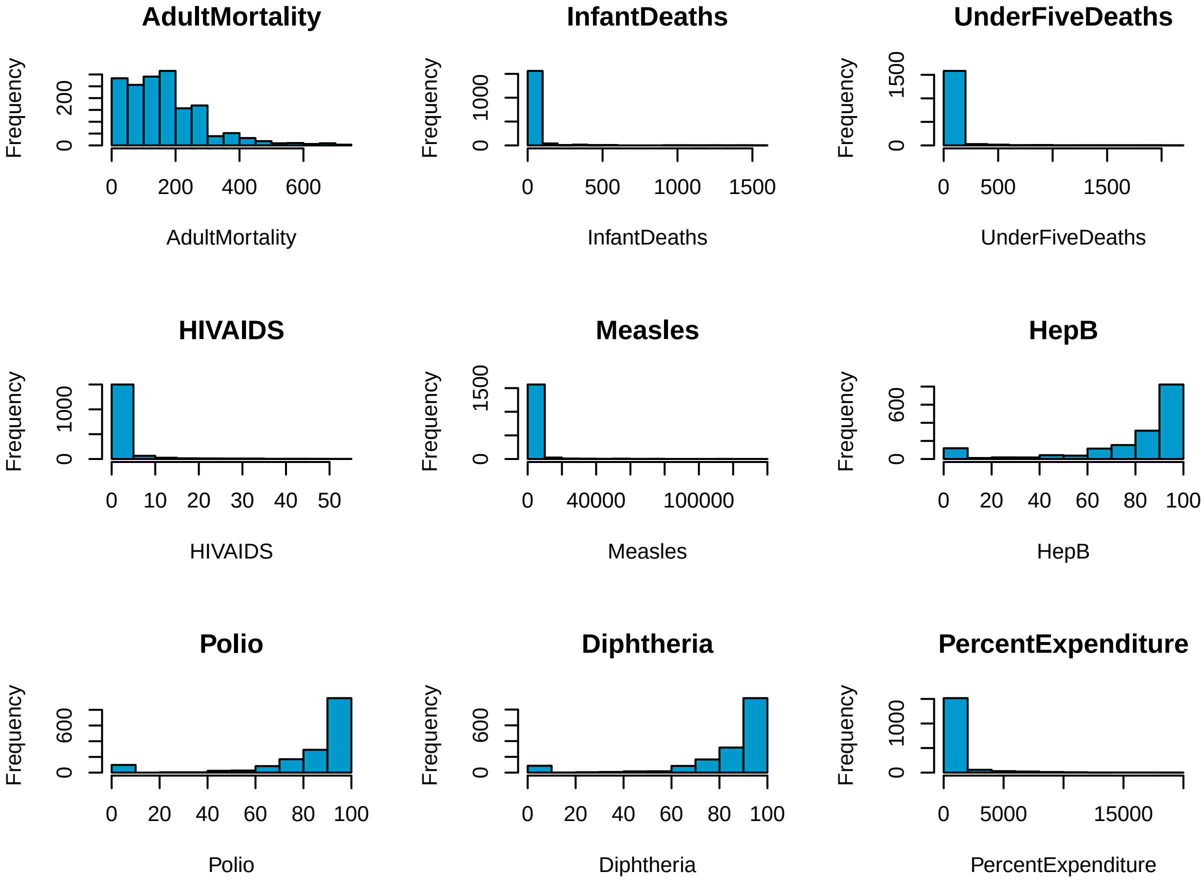
2.1.3.1 WHO Data

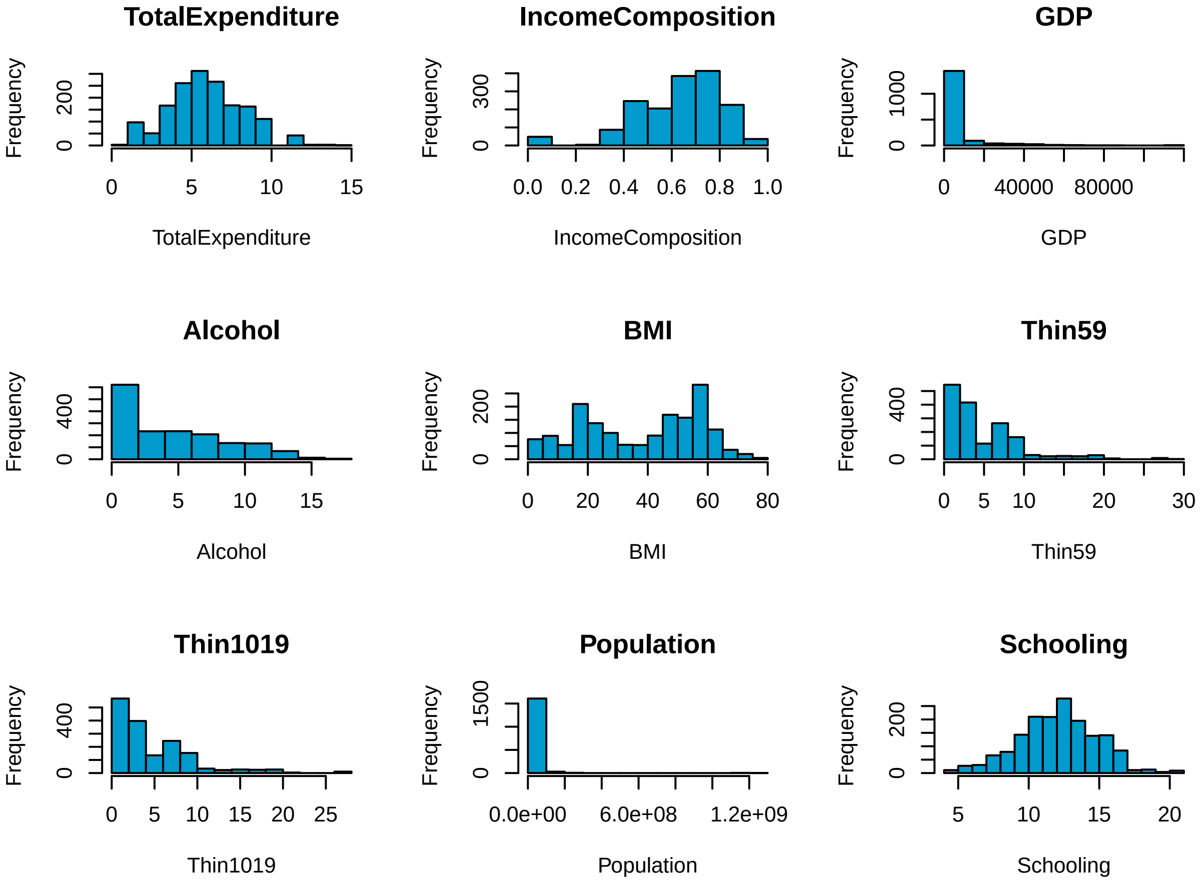
We first looked at the WHO data and cleaned up and reorganized the predictor variables so that they were grouped by common themes (deaths, diseases, vaccinations, economic resources, health/nutrition, and population/schooling) (See Appendix A). Histograms of the outcome variable (life expectancy) and the various predictor variables are shown below. Note that many of the variables are highly skewed, so we will need to transform them later. There are also some strange values for PercentExpenditure and BMI that will needed to be investigated later since PercentExpenditure should not be greater than 100% and BMI should not reach values of 80.

```
# Histogram of outcome (life expectancy)
hist(lifeData.new$LifeExpectancy, xlab=names(lifeData.new)[1], ylab="Frequency",
       main=names(lifeData.new)[1], col = "deepskyblue3")
```



```
# Histogram of predictive variables
oldpar = par(mfrow = c(3, 3))
for (i in 5:22) {
  hist(lifeData.new[ , i], xlab=names(lifeData.new)[i], ylab="Frequency",
       main=names(lifeData.new)[i], col = "deepskyblue3")
}
```


AdultMortality
Frequency
0
200
0
200
400
600
AdultMortality
InfantDeaths
Frequency
0
1000
0
500
1000
1500
InfantDeaths
UnderFiveDeaths
Frequency
0
1500
0
500
1500
UnderFiveDeaths
HIVAIDS
Frequency
0
1000
0
10
20
30
40
50
HIVAIDS
Measles
Frequency
0
1500
0
40000
100000
Measles
HepB
Frequency
0
600
0
20
40
60
80
100
HepB
Polio
Frequency
0
600
0
20
40
60
80
100
Polio
Diphtheria
Frequency
0
600
0
20
40
60
80
100
Diphtheria
PercentExpenditure
Frequency
0
1000
0
5000
15000
PercentExpenditure


TotalExpenditure
Frequency
0
200
0
5
10
15
TotalExpenditure
IncomeComposition
Frequency
0
300
0.0
0.2
0.4
0.6
0.8
1.0
IncomeComposition
GDP
Frequency
0
1000
0
40000
80000
GDP
Alcohol
Frequency
0
400
0
5
10
15
Alcohol
BMI
Frequency
0
200
0
20
40
60
80
BMI
Thin59
Frequency
0
400
0
5
10
20
30
Thin59
Thin1019
Frequency
0
400
0
5
10
15
20
25
Thin1019
Population
Frequency
0
1500
0.0e+00
6.0e+08
1.2e+09
Population
Schooling
Frequency
0
200
5
10
15
20
Schooling

```
par(oldpar)
```

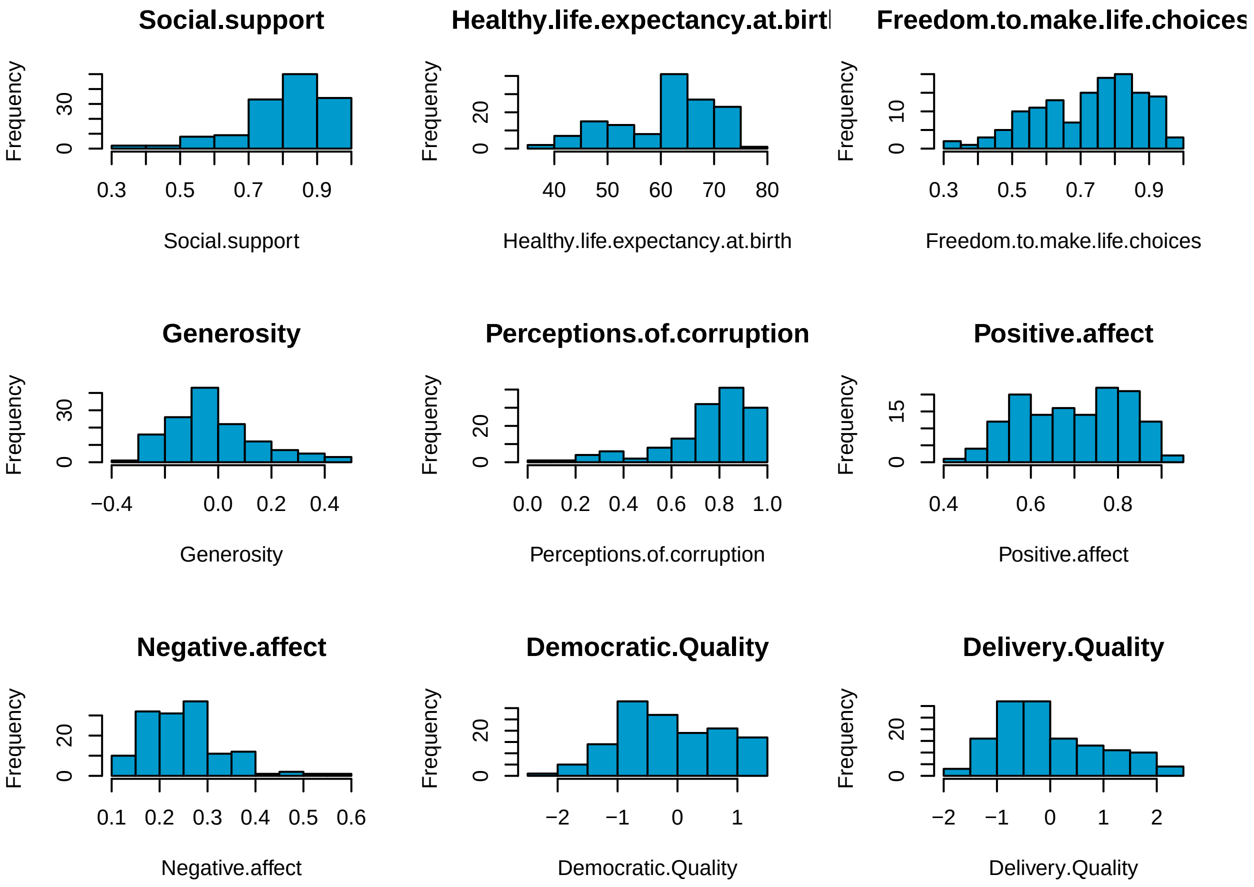
2.1.3.2 WHR Data

We next examined the WHR data to see what information was available. Many of the years did not have data from all the countries, so we decided to limit our analysis to 2011, which had at least some data available for the most countries (See Appendix B).

Some of the predictor variables from the WHR had a lot of missing information. For example, survey data on people's perception on how much most people could be trusted had up to 138 out of 164 entries missing. Since we did not have data to incorporate the columns with large amounts of missing data, we had to get rid of important markers that indicate levels of social trust (See Appendix B). From the remaining data, we had to also drop some rows missing data for Perceptions.of.corruption, which is less than 5% of our data.

After cleaning the data, we examined histograms of the predictive variables to get an idea of their distribution. Overall, they look pretty good compared to the WHO data, with much less skew.

```r
# Histogram of predictive variables
oldpar = par(mfrow = c(3, 3))
for (i in c(2:10)) {
  hist(whr.2011.clean[ , i], xlab=names(whr.2011.clean)[i], ylab="Frequency",
       main=names(whr.2011.clean)[i], col = "deepskyblue3")
}
```

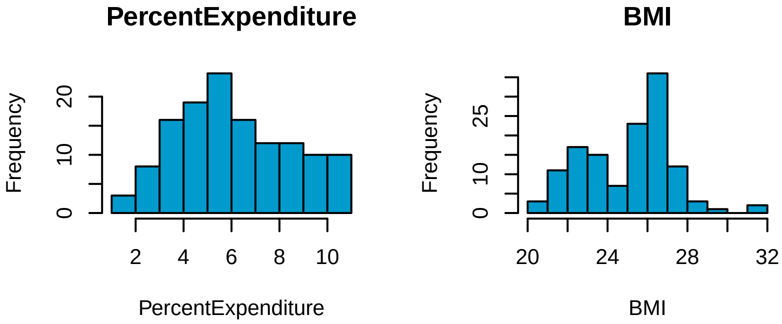


```r
par(oldpar)
```

2.1.3.3 Merged Data

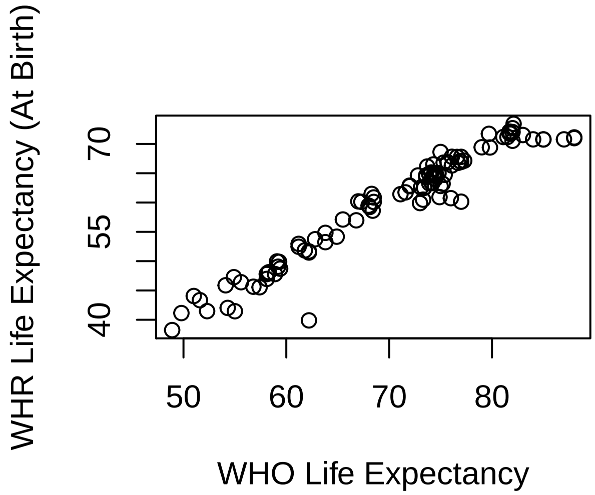
We finally want to combine the two datasets to consolidate all the predictors into one dataframe. As mentioned before, the WHO dataset had some anomalies to it. Since we decided to look only at 2011 based on the data available from the WHR, we went back to the primary sources (WHO Global Health Observatory and UN Human Development Reports) and obtained the data specifically for 2011 for PercentExpenditure and BMI with similar cleaning and reordering as before (See Appendix C). As can be observed below, PercentExpenditure now varies only from 0 to 20% and BMI only varies from 20 to 30, which are reasonable and make better sense than exceeding 100% and going up to 80, respectively.

```r
# Histogram of corrected predictive variables
oldpar = par(mfrow = c(1, 3))
for (i in c(13,18)) {
  hist(lifeData.2011[ , i], xlab=names(lifeData.2011)[i], ylab="Frequency",
       main=names(lifeData.2011)[i], col = "deepskyblue3")
}
par(oldpar)
```



Now satisfied with our two datasets, we merged them by country to obtain a final clean dataset. Since both datasets had variables corresponding to life expectancy in them, we did a final check to ensure that they agree with each other. We found that although they were not perfect matches, they were close enough that we were comfortable with the data (slope close to 1 and R^2 also close to 1). They were not expected to be perfect matches because one is an indicator of average life expectancy that year while the other is an indicator of expected life expectancy at birth.

```r
# Merge WHO and WHR data by country
whr.and.who.2011.clean = na.omit(merge(lifeData.2011, whr.2011, by = "Country"))

# Check if Life Expectancy matches between WHO and WHR
plot(whr.and.who.2011.clean$LifeExpectancy, whr.and.who.2011.clean$Healthy.life.expectancy.at.
```



```
summary(lm(LifeExpectancy~Healthy.life.expectancy.at.birth, data = whr.and.who.2011.clean))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ Healthy.life.expectancy.at.birth,
##     data = whr.and.who.2011.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8484 -1.3242 -0.3613  0.3950 11.8566
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       10.7547     1.5699   6.851 5.88e-10 ***
## Healthy.life.expectancy.at.birth   0.9923     0.0258  38.461  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.436 on 101 degrees of freedom
## Multiple R-squared:  0.9361, Adjusted R-squared:  0.9355
## F-statistic:  1479 on 1 and 101 DF,  p-value: < 2.2e-16
```

### Determining Variable Transformations

2.1.3.4 WHO Data

We examined the WHO 2011 data separately from the merged WHO and WHR 2011 data so that we can get an idea of whether there is value added from including the emotional well-being indicators. Status was the only categorical variable we had, which we will use as a dummy variable (Developed or not). From a table of the Status variable, we find that the majority of the countries have a "Developing" status (85%), which is as we would expect.

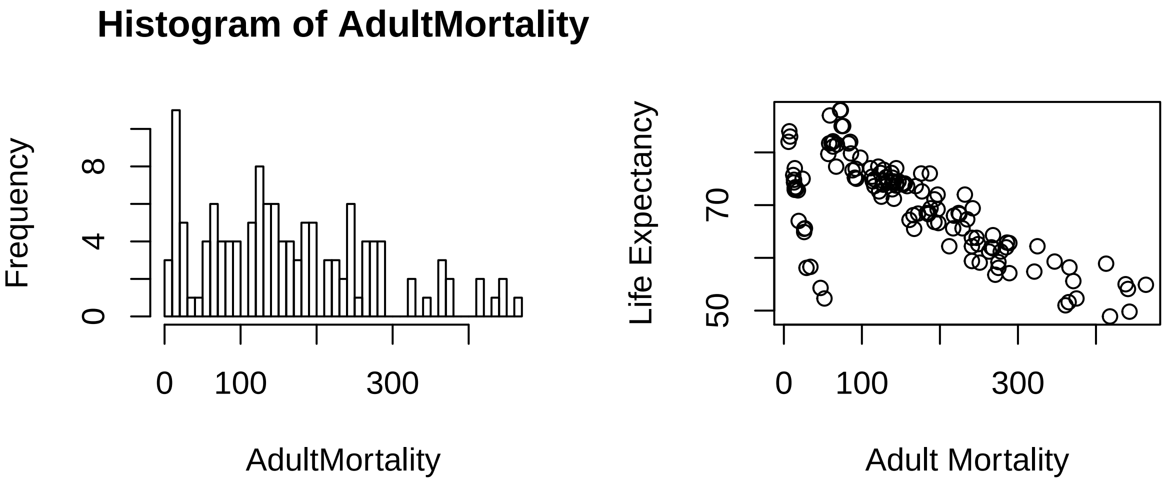
```
# Table of Status => mostly developing countries
with(lifeData.2011, prop.table(table(Status))*100)
```

```
## Status
##  Developed Developing
##   14.61538   85.38462
```

```
# Our "clean, transformed" data frame
lifeData.df = data.frame(developed = (lifeData.2011$Status == "Developed"))
```

Continuing on to the numerical variables, we looked at the variables individually to determine what kinds of transformations are needed to be made to the predictor variables to result in more normally distributed residuals for fits. We first did a quick preliminary run using the powerTransform function to give us an idea of the types of transformations that may be good to perform, but analyzed each individually for different transformations before deciding on the best (See Appendix D). An example with Adult Mortality is shown below. The histogram of AdultMortality shows that it is right-skewed. If we consider a linear model based only on Adult Mortality, we find that there is a signifcant and high negative correlation between Adult Mortality and Life Expectancy. We therefore definitely want to keep this predictor in the model. We considered log and square root of the variable as potential transformations and in the end settled on square root, which resulted in pretty close to a normal distribution of the predictor variable, as can be observed from the histogram of the distribution as well as the normal qqplot being close to linear.

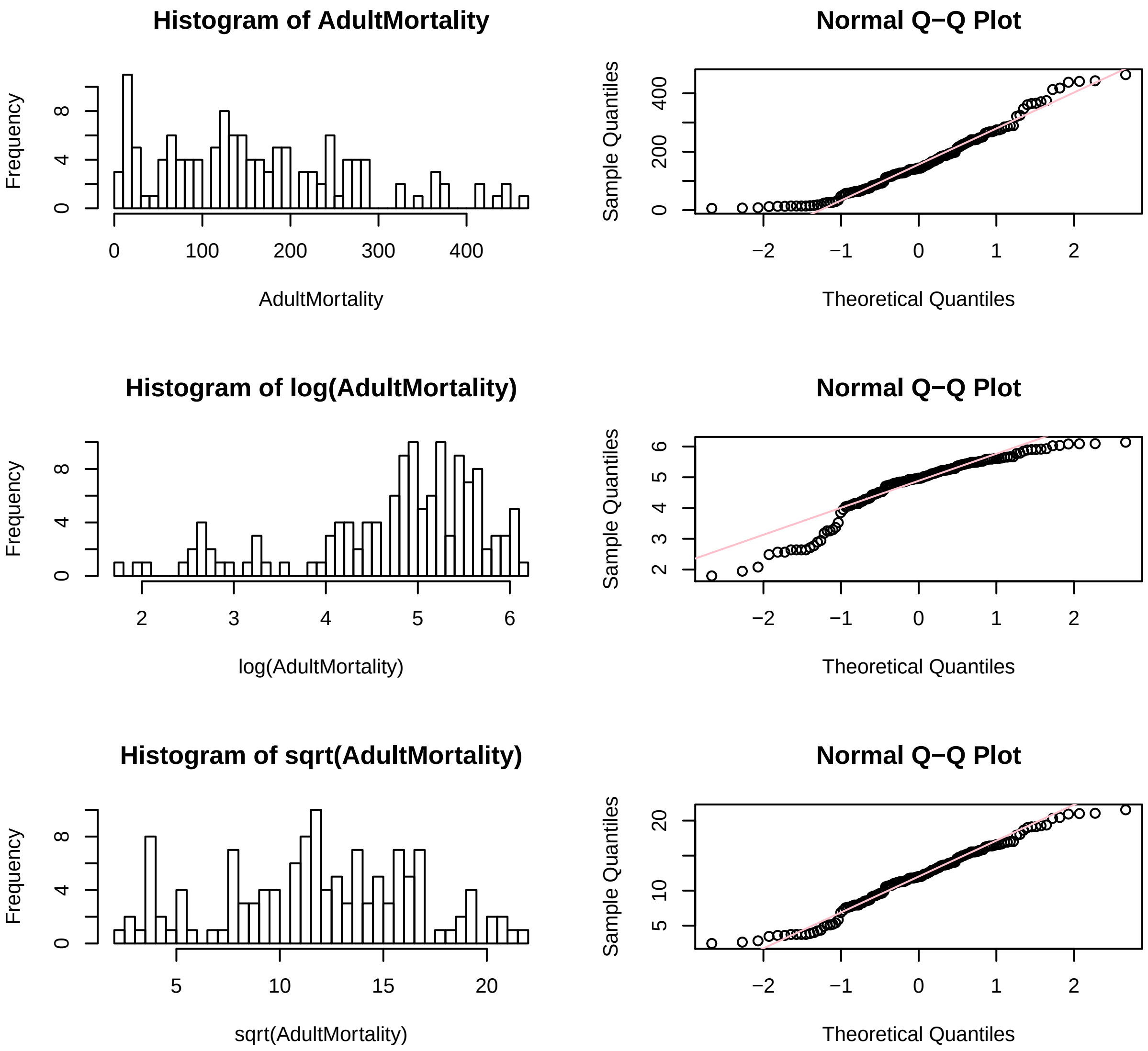
```
# AdultMortality
par(mfrow = c(1,2))
with(lifeData.2011,hist(AdultMortality,50))
plot(lifeData.2011$AdultMortality, lifeData.2011$LifeExpectancy,
     xlab="Adult Mortality", ylab="Life Expectancy")
```



```
summary(lm(LifeExpectancy~AdultMortality, data=lifeData.2011)) # Significant
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ AdultMortality, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.090  -2.980   1.053   3.639  12.840
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    79.435068   0.950677   83.56   <2e-16 ***
## AdultMortality -0.058564   0.004798  -12.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.128 on 128 degrees of freedom
## Multiple R-squared:  0.5379, Adjusted R-squared:  0.5343
## F-statistic:   149 on 1 and 128 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(3,2))
with(lifeData.2011,hist(AdultMortality,50))
with(lifeData.2011,qqnorm(AdultMortality))
with(lifeData.2011,qqline(AdultMortality,col="pink"))
with(lifeData.2011,hist(log(AdultMortality),50))
with(lifeData.2011,qqnorm(log(AdultMortality)))
with(lifeData.2011,qqline(log(AdultMortality),col="pink"))
with(lifeData.2011,hist(sqrt(AdultMortality),50))
with(lifeData.2011,qqnorm(sqrt(AdultMortality)))
with(lifeData.2011,qqline(sqrt(AdultMortality),col="pink"))
```



```
# use sqrt
```

In the end, the transformations we decided on can be found in Table 1.

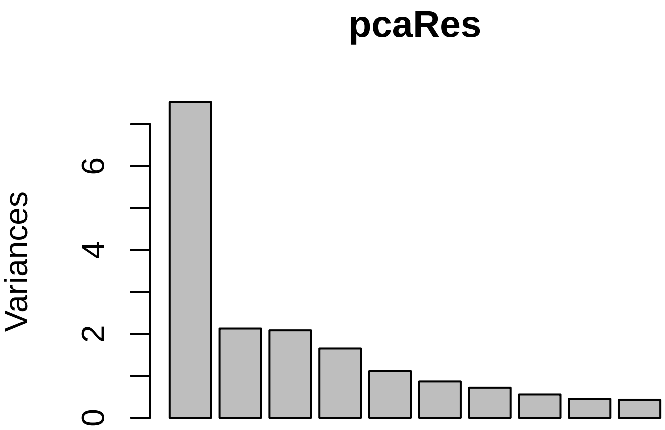
Variable	powerTransform	Final Decision
AdultMortality	0.72	0.5
InfantDeaths	0.14	0
Alcohol	0.5	0.5
PercentExpenditure	0.5	0.5
HepB	3.94	4
Measles	0.08	0
BMI	1	1
UnderFiveDeaths	0.15	0
Polio	4.08	4
TotalExpenditure	1	1
Diphtheria	4.64	5
HIVAIDS	-0.50	-0.5
GDP	0.12	0
Population	0.07	0
Thin1019	0.13	0.15
Thin59	0.14	0.15
IncomeComposition	1	1
Schooling	1.48	1

Table 1: Table of transformation recommended by powerTransform and our final decisions on transformations.

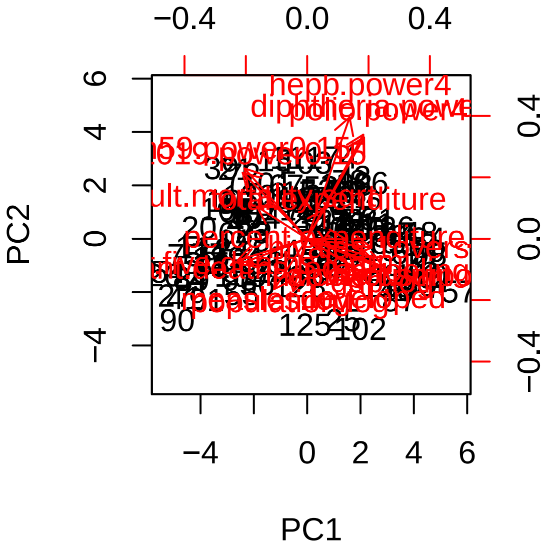
2.1.3.5 Principal Component Analysis

To get an idea of potentially important predictors, we performed Principal Component Analysis on the explanatory variables. On the scaled data, the 1st PC explains about 40% of variance and the next two PCs explain roughly comparable amount of variance – 11%. Multiple attributes in the dataset contribute comparably to the loadings: "positive" measures such as income, schooling, bmi, gdp, polio vaccination are positively correlated and "negative" measures such as infant deaths are negatively correlated with the first principal component. The second principal component is positively correlated with hepb, diphtheria, polio, thin59, thin1019 and negatively correlated with measles, population, and developed, to name a few.

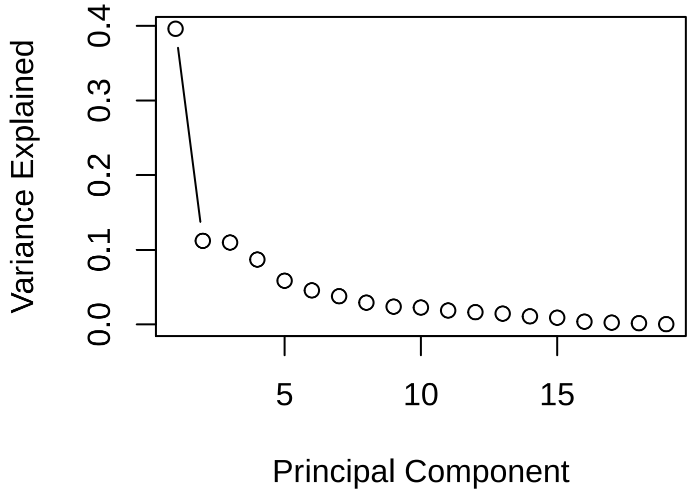
```r
# PCA on scaled data
pcaRes <- prcomp(lifeData.df,scale=TRUE)

par(mfrow = c(1,1))
plot(pcaRes)
```



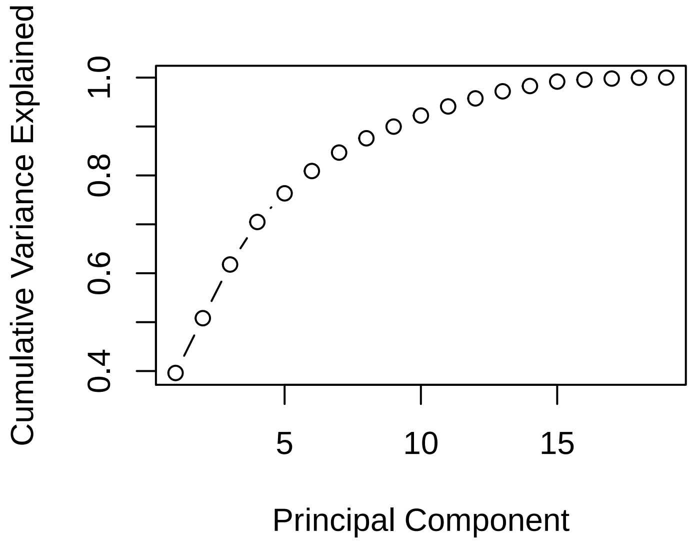
```
biplot(pcaRes, scale=0)
```



```
# print(pcaRes)
# summary(pcaRes)

std_dev <- pcaRes$sdev
pr_var <- std_dev^2
varExplained <- pr_var/sum(pr_var)
# varExplained

plot(varExplained, xlab = "Principal Component",
             ylab = "Variance Explained",
             type = "b")
```



```
plot(cumsum(varExplained), xlab = "Principal Component",
             ylab = "Cumulative Variance Explained",
             type = "b")
```



```
print("Ten largest by absolute value loadings, PC1:")
```

```
## [1] "Ten largest by absolute value loadings, PC1:"
```

```
print(pcaRes$rotation[order(abs(pcaRes$rotation[,1]),decreasing=TRUE)[1:10],1])
```

```
##    income.composition             schooling under.five.deaths.log
##             0.3225624             0.3099321            -0.2837303
##     infant.deaths.log      thin1019.power0.15      hivaids.1oversqrt
##            -0.2759220            -0.2629572             0.2623888
##      thin59.power0.15                   bmi               gdp.log
##            -0.2557721             0.2535997             0.2442099
##          polio.power4
##             0.2317297
```

```r
print("Ten largest by absolute value loadings, PC2:")
```

```
## [1] "Ten largest by absolute value loadings, PC2:"
```

```r
print(pcaRes$rotation[order(abs(pcaRes$rotation[,2]),decreasing=TRUE)[1:10],2])
```

```
##          hepb.power4    diphtheria.power5         polio.power4
##            0.4942682            0.4237781            0.4125311
##     thin59.power0.15   thin1019.power0.15       population.log
##            0.2864844            0.2690440           -0.2133754
##          measles.log            developed              gdp.log
##           -0.2053695           -0.1995342           -0.1483268
## adult.mortality.sqrt
##            0.1312273
```

2.1.3.6 WHO and WHR Data Combined

We followed the exact same procedure as before to decide how to transform the predictor variables from the WHR (See Appendix E). In the end, we decided that the variables were okay as they were and did not end up transforming them.

2.1.4 Regression

2.1.4.1 WHO Data

Trying linear regression with all the transformed variables from the WHO data alone, several variables already start to jump out, particularly income composition and adult mortality, but after examining VIFs, we see that some variables need to be eliminated in order to resolve multicollinearity issues (some VIFs as high as 70-80). We eliminated variables one by one until we were satisfied with the VIFs, during which process we lost under.five.deaths, thin1019, and income.composition to end up with all VIFs under 10 (See Appendix F).

```r
# Linear regression with all transformed variables
fit.2011 = lm(life.expectancy~.,data=lifeData.df)
summary(fit.2011)
```

```
##
## Call:
## lm(formula = life.expectancy ~ ., data = lifeData.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1827 -1.6842 -0.1311  1.9456  6.3600
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.475e+01  6.269e+00  10.328  < 2e-16 ***
## developedTRUE             4.225e-03  1.071e+00   0.004 0.996859
## adult.mortality.sqrt     -3.395e-01  7.642e-02  -4.443 2.12e-05 ***
## infant.deaths.log         1.998e-01  1.186e+00   0.169 0.866479
## under.five.deaths.log    -3.924e-01  1.211e+00  -0.324 0.746451
## hivaids.1oversqrt         1.638e+00  4.492e-01   3.647 0.000407 ***
## measles.log              -1.569e-01  1.159e-01  -1.355 0.178340
## hepb.power4               8.126e-09  1.884e-08   0.431 0.667131
## polio.power4             -2.507e-08  2.329e-08  -1.077 0.284044
## diphtheria.power5         1.011e-10  2.794e-10   0.362 0.718265
## percent.expenditure.sqrt  2.098e+00  8.063e-01   2.602 0.010544 *
## total.expenditure         6.083e-02  1.379e-01   0.441 0.659936
## income.composition        5.979e+01  6.745e+00   8.865 1.54e-14 ***
## gdp.log                  -6.529e-02  2.294e-01  -0.285 0.776450
## alcohol.sqrt             -6.272e-01  4.145e-01  -1.513 0.133161
## bmi                      -7.309e-01  1.913e-01  -3.821 0.000220 ***
## thin59.power0.15         -4.912e+00  5.768e+00  -0.852 0.396265
## thin1019.power0.15        2.226e+00  5.914e+00   0.376 0.707324
## population.log           -7.733e-02  1.278e-01  -0.605 0.546527
## schooling                -1.085e+00  2.623e-01  -4.135 6.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.024 on 110 degrees of freedom
## Multiple R-squared:  0.9033, Adjusted R-squared:  0.8866
## F-statistic: 54.09 on 19 and 110 DF,  p-value: < 2.2e-16
```

```r
# Examine VIFs
vif(fit.2011)
```

```
##                developed     adult.mortality.sqrt        infant.deaths.log
##                 2.034642                 1.848613                71.304094
##    under.five.deaths.log        hivaids.1oversqrt              measles.log
##                80.467108                 3.535992                 2.186941
##              hepb.power4             polio.power4        diphtheria.power5
##                 4.448693                 6.831936                 9.076448
## percent.expenditure.sqrt        total.expenditure       income.composition
##                 2.210152                 1.634975                15.528108
##                  gdp.log             alcohol.sqrt                      bmi
##                 2.172079                 2.523439                 2.433351
##         thin59.power0.15       thin1019.power0.15           population.log
##                15.978128                16.196183                 1.687115
##                schooling
##                 7.799522
```

```r
# Eliminate under.five.deaths, thin1019, and income.composition (see Appendix F)
lifeDataFit.df = lifeData.df[,-c(5,13,18)]
fit.2011 = lm(life.expectancy~.,data=lifeDataFit.df)
summary(fit.2011)
```

```
##
## Call:
```

```
## lm(formula = life.expectancy ~ ., data = lifeDataFit.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0545  -2.0402   0.3009   2.0541   8.5166
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.339e+01  7.985e+00   7.939 1.64e-12 ***
## developedTRUE             3.232e-01  1.386e+00   0.233 0.816014
## adult.mortality.sqrt     -3.822e-01  9.891e-02  -3.864 0.000187 ***
## infant.deaths.log        -6.129e-01  3.301e-01  -1.857 0.065950 .
## hivaids.1oversqrt         3.621e+00  4.919e-01   7.361 3.16e-11 ***
## measles.log              -1.009e-01  1.470e-01  -0.687 0.493549
## hepb.power4              -3.741e-08  2.345e-08  -1.595 0.113403
## polio.power4              1.066e-08  2.978e-08   0.358 0.721051
## diphtheria.power5         2.612e-10  3.616e-10   0.722 0.471589
## percent.expenditure.sqrt  5.141e-01  1.020e+00   0.504 0.615233
## total.expenditure        -4.473e-02  1.757e-01  -0.255 0.799502
## gdp.log                   4.498e-01  2.872e-01   1.566 0.120114
## alcohol.sqrt              7.367e-01  4.991e-01   1.476 0.142719
## bmi                      -2.891e-01  2.397e-01  -1.206 0.230273
## thin59.power0.15         -2.648e+00  2.650e+00  -0.999 0.319843
## population.log            1.140e-01  1.622e-01   0.703 0.483659
## schooling                 5.788e-01  2.403e-01   2.408 0.017641 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.926 on 113 degrees of freedom
## Multiple R-squared:  0.8326, Adjusted R-squared:  0.8088
## F-statistic: 35.12 on 16 and 113 DF,  p-value: < 2.2e-16
```

```
vif(fit.2011)
```

```
##                developed     adult.mortality.sqrt        infant.deaths.log
##                 2.021228                 1.837123                 3.278827
##        hivaids.1oversqrt              measles.log              hepb.power4
##                 2.514208                 2.086873                 4.085496
##             polio.power4        diphtheria.power5 percent.expenditure.sqrt
##                 6.624424                 9.018869                 2.098054
##        total.expenditure                  gdp.log             alcohol.sqrt
##                 1.574999                 2.020354                 2.169624
##                      bmi         thin59.power0.15           population.log
##                 2.266102                 2.000102                 1.611443
##                schooling
##                 3.882540
```

With the remaining variables, we performed forward and backward stepwise regression in order to eliminate insignificant variables, then further refined the model by eliminating variables one by one that were not statistically significant (below our α cut-off of 0.01). The final model consists of four predictors: adult.mortality, hivaids, alcohol, and schooling.

```
# Forward and backward stepwise regression
step <- stepAIC(fit.2011, direction="both", trace=0)
# step$anova # Display results

#summary(step)

fit.final2=lm(life.expectancy ~ adult.mortality.sqrt + infant.deaths.log +
    hivaids.1oversqrt + hepb.power4 + gdp.log + alcohol.sqrt + schooling,
    data = lifeDataFit.df)
# summary(fit.final2)

fit.final3=lm(life.expectancy ~ adult.mortality.sqrt + infant.deaths.log +
    hivaids.1oversqrt + gdp.log + alcohol.sqrt + schooling, data = lifeDataFit.df)
# summary(fit.final3)

fit.final4=lm(life.expectancy ~ adult.mortality.sqrt + infant.deaths.log +
    hivaids.1oversqrt + alcohol.sqrt + schooling, data = lifeDataFit.df)
# summary(fit.final4)

fit.final5=lm(life.expectancy ~ adult.mortality.sqrt + hivaids.1oversqrt +
                alcohol.sqrt + schooling, data = lifeDataFit.df)

# Checking modeling assumptions
summary(fit.final5)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + hivaids.1oversqrt +
##     alcohol.sqrt + schooling, data = lifeDataFit.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3384  -2.2897   0.1218   2.1222  10.0013
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          53.01323    2.61247  20.292  < 2e-16 ***
## adult.mortality.sqrt -0.42053    0.09458  -4.446 1.91e-05 ***
## hivaids.1oversqrt     3.91077    0.46021   8.498 4.88e-14 ***
## alcohol.sqrt          1.19455    0.44630   2.677  0.00844 **
## schooling             0.87309    0.21125   4.133 6.51e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.005 on 125 degrees of freedom
## Multiple R-squared:  0.8073, Adjusted R-squared:  0.8012
## F-statistic: 130.9 on 4 and 125 DF,  p-value: < 2.2e-16
```

```
ncvTest(fit.final5)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.3600022, Df = 1, p = 0.5485
```

```
shapiro.test(residuals(fit.final5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit.final5)
## W = 0.98477, p-value = 0.1555
```

```
vif(fit.final5)
```

```
## adult.mortality.sqrt    hivaids.1oversqrt         alcohol.sqrt
##             1.614630             2.115903             1.667817
##            schooling
##             2.884187
```

2.1.4.2 Checking Modeling Assumptions

Multicolinearity: All our VIFs are under 10

Homoscedasticity: p = .55>.05 so we fail to reject Ho: our errors have constant variance.

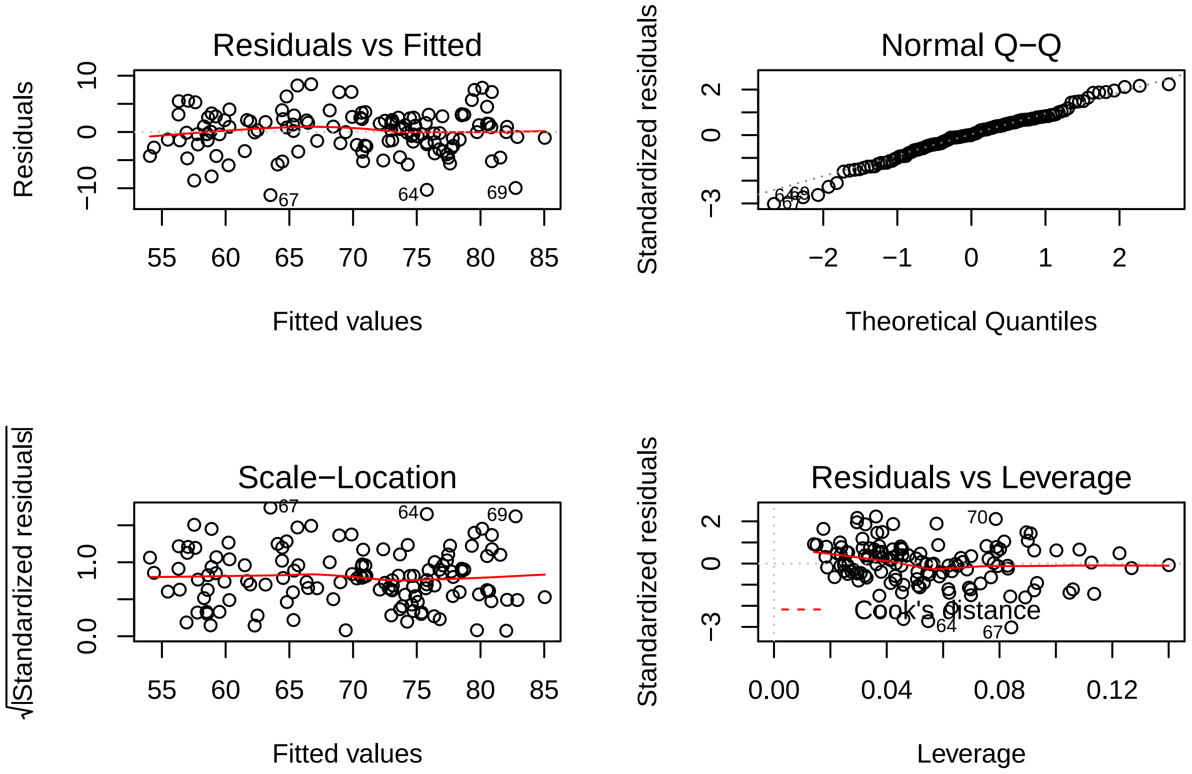
Errors~N(0,sigma^2): p = .16>.05 so we fail to reject Ho: our errors are normally distributed.

Independence: we specifically chose not to use data for multiple years to avoid problems with independence.

Linearity: we did not observe any patterns in our residuals versus fitted diagnostic graph.

With the diagnostic plots of the final fit, we demonstrate with visual indicators that the residuals of the fit are well-behaved. No outliers jump out, and there is no indication that we fit a line when we should fit a curve since the residuals vs. fitted and scale-location graphs looks to be fairly horizontal. The normal qqplot show that the residuals appear to be normally distributed since the standardized residuals fall close to a line when plotted against theoretical quanties. It's not perfect, with some points that fall below and above the line at the lower and upper quantiles, respectively, but overall it is quite good. Finally, examining the residuals vs. leverage graph, we see that no points are more than Cook's distance away. The residuals are all quite low, and the points with highest leverage have close to 0 residual.

```
par(mfrow = c(2,2))
plot(fit.final3)
```



2.1.4.3 WHO and WHR Data

When we combined the WHO and WHR datasets, we went down from 130 countries to 103 countries due to missing data. Therefore, we first want to make sure we at least get a similar regression result with just the WHO predictors with the reduced number of countries in the combined data. Again, we eliminated variables to reduce VIFs below 10 and ended up removing similar variables: under.5.deaths, thin.5.9 (instead of thin.10.19), and income.composition (See Appendix G). We then performed foward and backward stepwise regression using stepAIC then manually eliminating variables one by one that did not have a p-value below 0.01.

```
fit.whr.who.only.1 = lm(life.expectancy~.-under.five.deaths.shift1.log
    -income.composition-thin.5.9.power0.15, data = transformed.who)
#summary(fit.whr.who.only.1)
#vif(fit.whr.who.only.1)

fit.whr.who.only.2 = stepAIC(fit.whr.who.only.1, direction = "both", trace = 0)
# summary(fit.whr.who.only.2)

fit.whr.who.only.3 = update(fit.whr.who.only.2,.~.-gdp.per.cap.log )
# summary(fit.whr.who.only.3)

fit.whr.who.only.4 = update(fit.whr.who.only.3,.~.-polio.power4)
# summary(fit.whr.who.only.4)

fit.whr.who.only.5 = update(fit.whr.who.only.4,.~.-hep.b.power4)
```

```r
# summary(fit.whr.who.only.5)

fit.whr.who.only.6 = update(fit.whr.who.only.5,.~.-total.expenditure)
# summary(fit.whr.who.only.6)

fit.whr.who.only.7 = update(fit.whr.who.only.6,.~.-infant.deaths.shift1.log)
summary(fit.whr.who.only.7)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + hivaids.1oversqrt +
##     thin.10.19.power0.15 + schooling, data = transformed.who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1388 -2.1876 -0.0593  1.9680 10.1533
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          56.53875    3.11419  18.155  < 2e-16 ***
## adult.mortality.sqrt -0.34323    0.09851  -3.484 0.000739 ***
## hivaids.1oversqrt     3.70428    0.49332   7.509 2.79e-11 ***
## thin.10.19.power0.15 -2.05804    0.53362  -3.857 0.000206 ***
## schooling             0.92519    0.19518   4.740 7.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.866 on 98 degrees of freedom
## Multiple R-squared:  0.8438, Adjusted R-squared:  0.8374
## F-statistic: 132.3 on 4 and 98 DF,  p-value: < 2.2e-16
```

```r
ncvTest(fit.whr.who.only.7)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2331086, Df = 1, p = 0.62923
```

```r
shapiro.test(residuals(fit.whr.who.only.7))
```

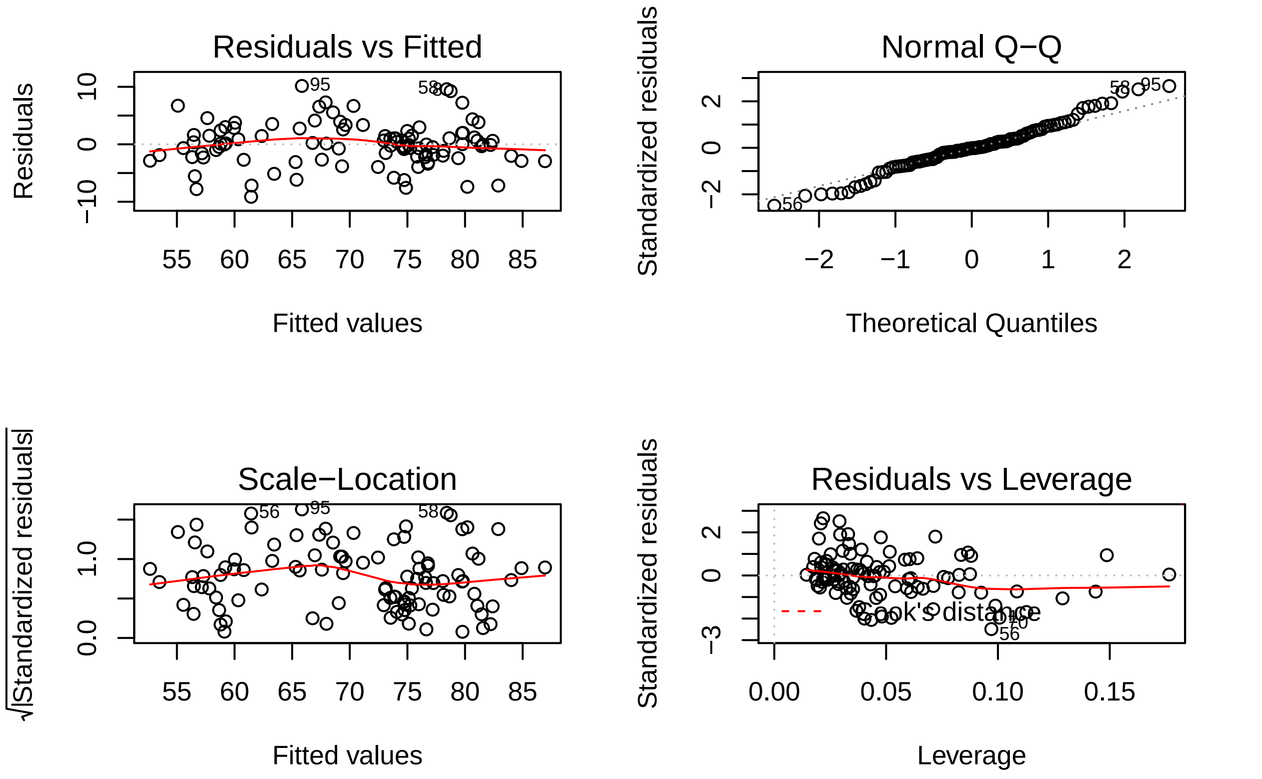
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit.whr.who.only.7)
## W = 0.98042, p-value = 0.1308
```

```r
vif(fit.whr.who.only.7)
```

```
## adult.mortality.sqrt    hivaids.1oversqrt thin.10.19.power0.15
##             1.594921             2.105124             1.765452
##            schooling
##                 2.231589
```

Overall, the results were similar to before, where four predictors remained at the end, three of which were the same (adult.mortality, hivaids, and schooling), but alcohol was replaced by thin.10.19. This is not too concerning since they likely measure similar attributes of general health in the country. Looking at the ncvTest, Shapiro-Wilk test, and diagnostic plots (as discussed above), we were satisfied our assumptions of linear, homoskedastic, independent, normally distributed residuals were met.

```r
par(mfrow = c(2,2))
plot(fit.whr.who.only.7)
```



We then asked if there is a role happiness indicators play on life expectancy. We performed a partial F-test to see if any of the additional happiness indicators are needed to predict life expectancy. From the test, we obtained a p-value of 0.0018, so at least one of the happiness indicators seem to have an impact on life expectancy.

Repeating the same procedure with all the predictors from both datasets (See Appendix H), we arrive at a final model consisting of 6 predictors using our full dataset: adult.mortality, infant.deaths, hivaids, positive.affect, negative.affect, and delivery.quality. adult.mortality and hivaids were two of the four variables we had previously, but infant.deaths was new and we see the appearance of 3 predictors from the WHR dataset.

Looking at the ncvTest, Shapiro-Wilk test, and diagnostic plots, we were satisfied our assumptions of linear, homoskedastic, independent, normally distributed residuals were met.

We also looked at BIC model probabilities of the Final WHO/WHR model vs. WHO Only model vs. WHO plus Full WHR model (containing all WHR predictors). The Final WHO/WHR Model has a BIC probability of ~100%, so we are satisfied that there is added value in adding emotional and social indicators of well-being into our model.

```
# Partial F-test
# Can we improve the final WHO-only model by adding any of the WHR hapiness variables?
# Model from before
fit.who.and.whr.reducedmodel = fit.whr.who.only.7

# Fit the final WHO model plus the happiness indicators
fit.who.and.whr.fullmodel = lm(life.expectancy ~ adult.mortality.sqrt + hivaids.1oversqrt +
    thin.10.19.power0.15 + schooling + social.support + freedom.to.make.life.choices +
    generosity + perceptions.of.corruption + positive.affect + negative.affect +
    democratic.quality + delivery.quality, data=transformed.who.and.whr)

# Get the p-value of the partial F-test
anova(fit.who.and.whr.fullmodel, fit.who.and.whr.reducedmodel)
```

```
## Analysis of Variance Table
##
## Model 1: life.expectancy ~ adult.mortality.sqrt + hivaids.1oversqrt +
##     thin.10.19.power0.15 + schooling + social.support + freedom.to.make.life.choices +
##     generosity + perceptions.of.corruption + positive.affect +
##     negative.affect + democratic.quality + delivery.quality
## Model 2: life.expectancy ~ adult.mortality.sqrt + hivaids.1oversqrt +
##     thin.10.19.power0.15 + schooling
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     90 1123.2
## 2     98 1464.6 -8   -341.47 3.4202 0.001773 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Final fit starting with all predictors (See Appendix H)
fit.who.and.whr.final = lm( life.expectancy ~ adult.mortality.sqrt +
                              infant.deaths.shift1.log + hivaids.1oversqrt +
                              positive.affect + negative.affect + delivery.quality,
                            data = transformed.who.and.whr)
summary(fit.who.and.whr.final)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log +
##     hivaids.1oversqrt + positive.affect + negative.affect + delivery.quality,
##     data = transformed.who.and.whr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.591  -1.986  -0.257   2.330   6.954
##
## Coefficients:
```

```
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                52.5070     3.5138  14.943  < 2e-16 ***
## adult.mortality.sqrt       -0.3008     0.0873  -3.445 0.000847 ***
## infant.deaths.shift1.log   -0.9627     0.2488  -3.868 0.000200 ***
## hivaids.1oversqrt           4.4092     0.4028  10.946  < 2e-16 ***
## positive.affect            13.7794     3.3446   4.120 8.04e-05 ***
## negative.affect            13.9296     4.9942   2.789 0.006373 ** 
## delivery.quality            2.6422     0.5497   4.806 5.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.416 on 96 degrees of freedom
## Multiple R-squared:  0.8805, Adjusted R-squared:  0.8731 
## F-statistic: 117.9 on 6 and 96 DF,  p-value: < 2.2e-16
```

```r
vif(fit.who.and.whr.final)
```

```
##     adult.mortality.sqrt infant.deaths.shift1.log        hivaids.1oversqrt 
##                 1.604683                 1.764358                 1.798164 
##          positive.affect          negative.affect         delivery.quality 
##                 1.394716                 1.238993                 2.014922
```

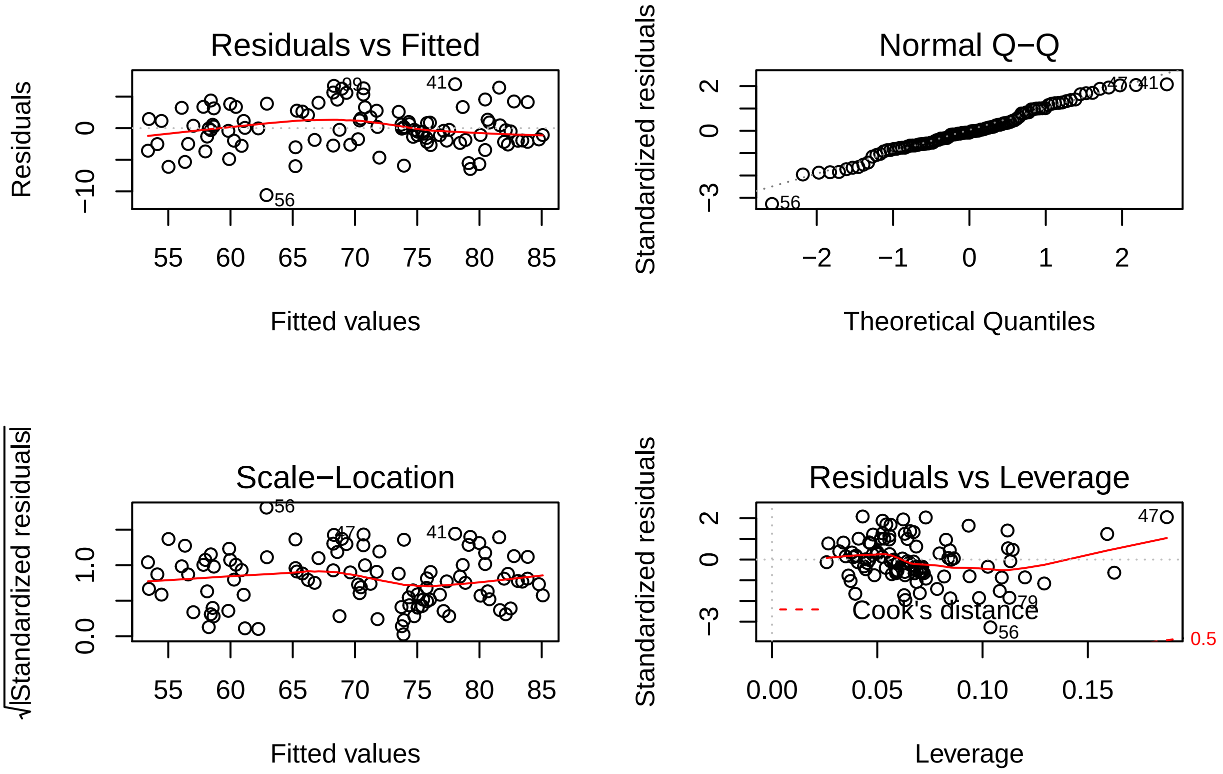
```r
ncvTest(fit.who.and.whr.final)
```

```
## Non-constant Variance Score Test 
## Variance formula: ~ fitted.values 
## Chisquare = 0.6295583, Df = 1, p = 0.42752
```

```r
shapiro.test(residuals(fit.who.and.whr.final))
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  residuals(fit.who.and.whr.final)
## W = 0.98343, p-value = 0.2261
```

```r
par(mfrow = c(2,2))
plot(fit.who.and.whr.final)
```



```
# BIC probabilities of models
thebics=c(BIC(fit.who.and.whr.final),BIC(fit.who.and.whr.reducedmodel),BIC(fit.who.and.whr.full
eBIC <- exp(-0.5*(thebics-min(thebics)))
modelprobs <- eBIC/sum(eBIC)
names(modelprobs)=c("Final Model","WHO Only Model","WHO and Full WHR Model")
modelprobs
```

```
##            Final Model         WHO Only Model WHO and Full WHR Model
##           9.998965e-01           1.026667e-04           7.892760e-07
```

3 Appendices

3.1 Appendix A: WHO/UN Data Cleanup

```
# Load WHO data
lifeData = read.csv("Life Expectancy Data.csv")

# Remove rows with NA
lifeData = na.omit(lifeData)

# Reorder data so outcome (life expectancy) is first and grab Country, Year, and Status
lifeData.new = cbind(LifeExpectancy=lifeData$Life.expectancy, lifeData[ , 1:3])

# Group other data according to common themes
lifeData.new = cbind(lifeData.new,
                     AdultMortality=lifeData$Adult.Mortality, # Deaths
                     InfantDeaths=lifeData$infant.deaths,
                     UnderFiveDeaths=lifeData$under.five.deaths,
                     HIVAIDS=lifeData$HIV.AIDS,
                     Measles=lifeData$Measles, # Diseases
                     HepB=lifeData$Hepatitis.B, # Vaccinations
                     Polio=lifeData$Polio, Diphtheria=lifeData$Diphtheria,
                     PercentExpenditure=lifeData$percentage.expenditure, # Economic Resources
                     TotalExpenditure=lifeData$Total.expenditure,
                     IncomeComposition=lifeData$Income.composition.of.resources,
                     GDP=lifeData$GDP,
                     Alcohol=lifeData$Alcohol, BMI=lifeData$BMI, # Health/Nutrition
                     Thin59=lifeData$thinness.5.9.years,
                     Thin1019=lifeData$thinness..1.19.years,
                     Population=lifeData$Population, # Other (population/schooling)
                     Schooling=lifeData$Schooling)
```

3.2 Appendix B: World Happiness Report Data Cleanup

```
# Load WHR Data
whr.df = read.csv("Table 2.1.csv")

# Look at structure of WHR Data
str(whr.df)
```

```
## 'data.frame':    1111 obs. of  22 variables:
##  $ WP5.Country                                     : Factor w/ 163 levels "001 United States'
##  $ Country                                         : Factor w/ 163 levels "Afghanistan",..: 
##  $ Year                                            : int  2008 2009 2010 2011 2012 2013 2014
##  $ Life.Ladder                                     : num  3.72 4.4 4.76 3.83 3.78 ...
##  $ Log.GDP.per.capita                              : num  7.18 7.34 7.4 7.44 7.55 ...
##  $ Social.support                                 : num  0.451 0.552 0.539 0.521 0.521 ...
##  $ Healthy.life.expectancy.at.birth               : num  47.9 48.3 48.7 49.1 49.4 ...
##  $ Freedom.to.make.life.choices                   : num  0.718 0.679 0.6 0.496 0.531 ...
##  $ Generosity                                     : num  0.18 0.201 0.135 0.172 0.244 ...
##  $ Perceptions.of.corruption                      : num  0.882 0.85 0.707 0.731 0.776 ...
##  $ Positive.affect                                : num  0.518 0.584 0.618 0.611 0.71 ...
##  $ Negative.affect                                : num  0.258 0.237 0.275 0.267 0.268 ...
##  $ Confidence.in.national.government              : num  0.612 0.612 0.299 0.307 0.435 ...
##  $ Democratic.Quality                             : num  -1.96 -2.08 -2.02 -1.95 -1.87 ...
##  $ Delivery.Quality                               : num  -1.67 -1.65 -1.63 -1.62 -1.43 ...
##  $ Most.people.can.be.trusted..Gallup             : num  NA 0.286 0.276 NA NA ...
##  $ Most.people.can.be.trusted..WVS.round.1981.1984: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Most.people.can.be.trusted..WVS.round.1989.1993: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Most.people.can.be.trusted..WVS.round.1994.1998: num  NA NA NA NA NA ...
##  $ Most.people.can.be.trusted..WVS.round.1999.2004: num  NA NA NA NA NA ...
##  $ Most.people.can.be.trusted..WVS.round.2005.2009: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Most.people.can.be.trusted..WVS.round.2010.2014: num  NA NA NA NA NA NA NA NA NA NA ...
```

```
# Look at what years have the most data available
table(whr.df$Year)
```

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
##   27   90  102  110  114  124  146  142  137  119
```

```
# Consider only 2011 data
whr.2011 = whr.df[whr.df$Year==2011,]

# Examine which columns have a lot of missing data
na_count = data.frame(sapply(whr.2011, function(y) sum(length(which(is.na(y))))))
na_count
```

```
##                                                  sapply.whr.2011..function.y..sum.length.whi
## WP5.Country
## Country
## Year
## Life.Ladder
## Log.GDP.per.capita
## Social.support
## Healthy.life.expectancy.at.birth
## Freedom.to.make.life.choices
## Generosity
## Perceptions.of.corruption
## Positive.affect
## Negative.affect
## Confidence.in.national.government
## Democratic.Quality
## Delivery.Quality
## Most.people.can.be.trusted..Gallup
## Most.people.can.be.trusted..WVS.round.1981.1984
## Most.people.can.be.trusted..WVS.round.1989.1993
## Most.people.can.be.trusted..WVS.round.1994.1998
## Most.people.can.be.trusted..WVS.round.1999.2004
## Most.people.can.be.trusted..WVS.round.2005.2009
## Most.people.can.be.trusted..WVS.round.2010.2014
```

```
# Take subset of data (remove columns)
whr.2011 = subset(whr.2011, select = -c(Confidence.in.national.government, # Missing data
                                        Most.people.can.be.trusted..Gallup,
                                        Most.people.can.be.trusted..WVS.round.1981.1984,
                                        Most.people.can.be.trusted..WVS.round.1989.1993,
                                        Most.people.can.be.trusted..WVS.round.1994.1998,
                                        Most.people.can.be.trusted..WVS.round.1999.2004,
                                        Most.people.can.be.trusted..WVS.round.2005.2009,
                                        Most.people.can.be.trusted..WVS.round.2010.2014,

                                        Year, # Only looking at 2011
                                        WP5.Country, # Not a predictor.
                                        Life.Ladder, # Linear combination of other predictors
                                        Log.GDP.per.capita)) # GDP already in WHO data
na_count = data.frame(sapply(whr.2011, function(y) sum(length(which(is.na(y))))))
na_count
```

```
##                                  sapply.whr.2011..function.y..sum.length.which.is.na.y.....
## Country                                                                                   0
## Social.support                                                                            1
## Healthy.life.expectancy.at.birth                                                          1
## Freedom.to.make.life.choices                                                              2
## Generosity                                                                                3
## Perceptions.of.corruption                                                                 8
## Positive.affect                                                                           1
## Negative.affect                                                                           0
## Democratic.Quality                                                                        1
## Delivery.Quality                                                                          1
```

```
# Take subset of data (remove rows)
whr.2011.clean = whr.2011[!is.na(whr.2011$Perceptions.of.corruption),]
na_count = data.frame(sapply(whr.2011.clean, function(y) sum(length(which(is.na(y))))))
na_count
```

```
##                                  sapply.whr.2011.clean..function.y..sum.length.which.is.na.
## Country
## Social.support
## Healthy.life.expectancy.at.birth
## Freedom.to.make.life.choices
## Generosity
## Perceptions.of.corruption
## Positive.affect
## Negative.affect
## Democratic.Quality
## Delivery.Quality
```

3.3 Appendix C: WHO/UN Data for 2011 Cleanup

```
# Load 2011 data
lifeData.2011 = read.csv("Life Expectancy Data 2011.csv")

# Remove rows with NA
lifeData.2011 = na.omit(lifeData.2011)

# Reorder data so outcome (life expectancy) is first and grab Country, Year, and Status
lifeData.new = cbind(LifeExpectancy=lifeData.2011$Life.expectancy, lifeData.2011[ , 1:3])

# Group other data according to common themes
lifeData.new = cbind(lifeData.new,
                     AdultMortality=lifeData.2011$Adult.Mortality, # Deaths
                     InfantDeaths=lifeData.2011$infant.deaths,
                     UnderFiveDeaths=lifeData.2011$under.five.deaths,
                     HIVAIDS=lifeData.2011$HIV.AIDS,
                     Measles=lifeData.2011$Measles, # Diseases
                     HepB=lifeData.2011$Hepatitis.B, # Vaccinations
                     Polio=lifeData.2011$Polio, Diphtheria=lifeData.2011$Diphtheria,
                     PercentExpenditure=lifeData.2011$percentage.expenditure, # Economic Resou
                     TotalExpenditure=lifeData.2011$Total.expenditure,
                     IncomeComposition=lifeData.2011$Income.composition.of.resources,
                     GDP=lifeData.2011$GDP,
                     Alcohol=lifeData.2011$Alcohol, BMI=lifeData.2011$BMI, # Health/Nutrition
                     Thin59=lifeData.2011$thinness.5.9.years,
                     Thin1019=lifeData.2011$thinness..1.19.years,
                     Population=lifeData.2011$Population, # Other (population/schooling)
                     Schooling=lifeData.2011$Schooling)
lifeData.2011 = lifeData.new
```

3.4 Appendix D: Transformations for WHO Data

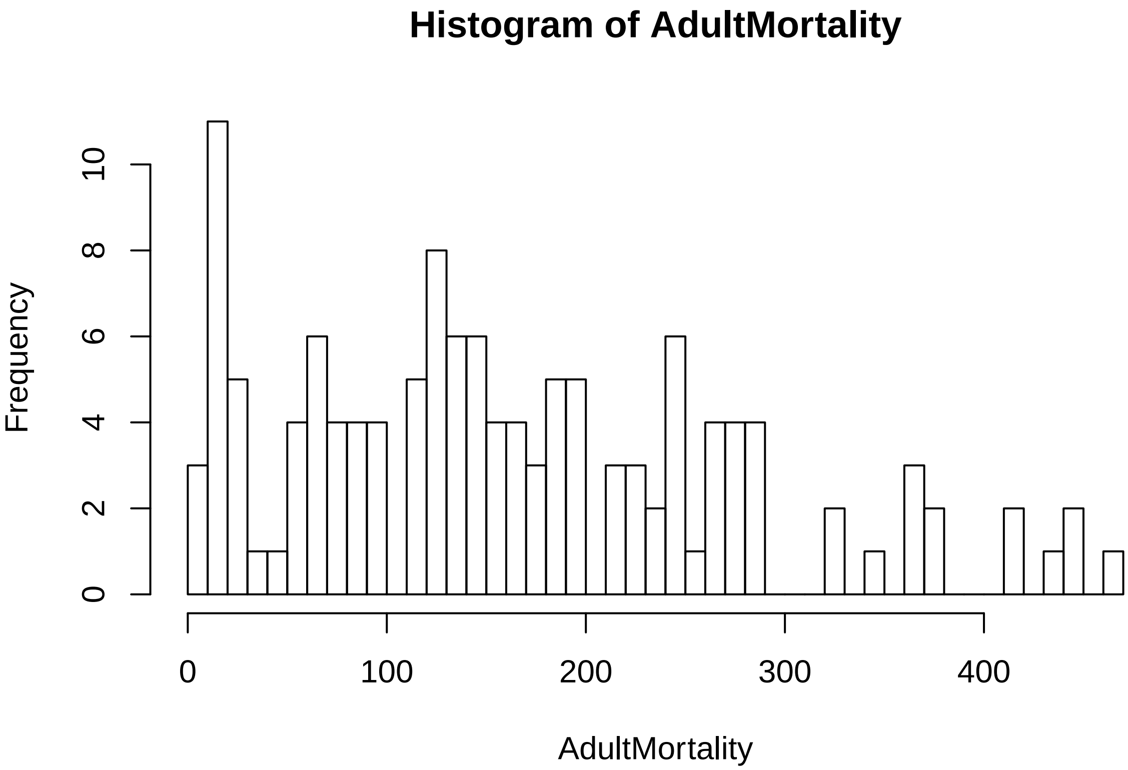
```
###########################################
# Using powerTransform

# Dataframe for using powerTransform to determine good transformations for predictors
lifeData2011.tf = lifeData.2011

# Make 0's 1e-12 (data must be strictly positive)
lifeData2011.tf[which(lifeData2011.tf==0, arr.ind = TRUE)] = 1e-12
```

```r
# Use powerTransform to give idea of potential transformations
summary(powerTransform(with(lifeData2011.tf, cbind(AdultMortality, InfantDeaths,
    Alcohol, PercentExpenditure, HepB, Measles, BMI, UnderFiveDeaths, Polio,
    TotalExpenditure, Diphtheria, HIVAIDS, GDP, Population, Thin1019, Thin59,
    IncomeComposition, Schooling))))
```

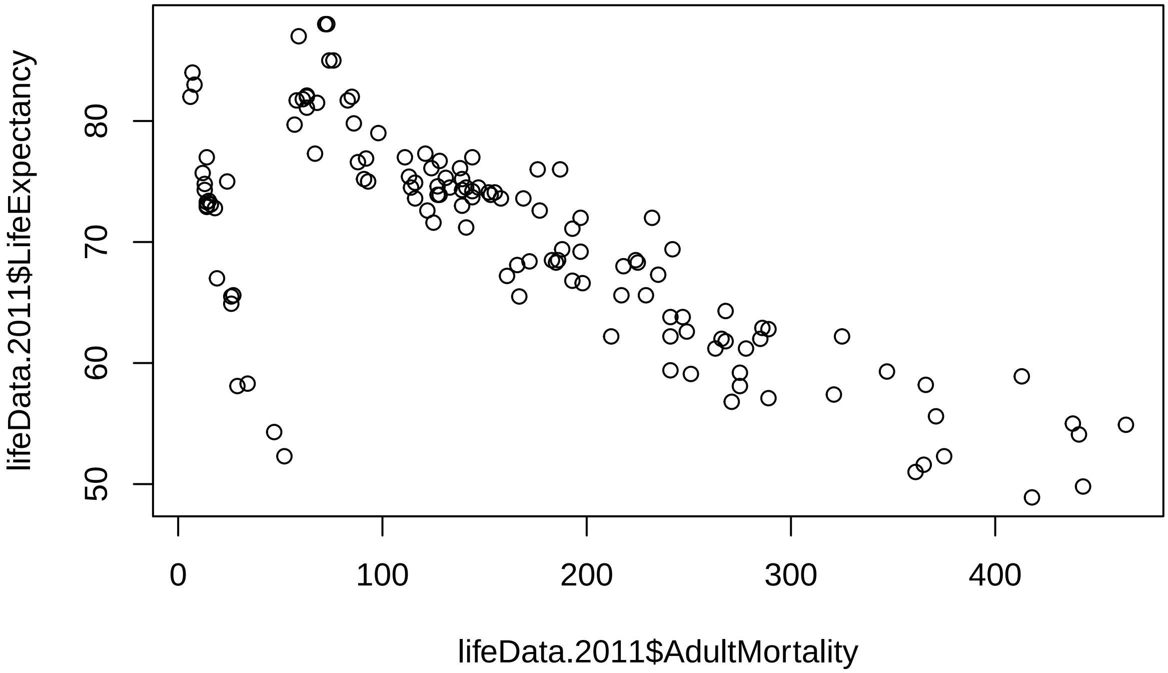
```
## Warning in estimateTransform.default(X, Y, weights, family, ...):
## Convergence failure: return code = 1

## bcPower Transformations to Multinormality
##                     Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## AdultMortality        0.7189        0.72       0.5389       0.8988
## InfantDeaths          0.1400        0.14       0.1184       0.1616
## Alcohol               0.4779        0.50       0.3679       0.5880
## PercentExpenditure    0.4822        0.50       0.1402       0.8243
## HepB                  3.9443        3.94       3.3126       4.5760
## Measles               0.0794        0.08       0.0617       0.0971
## BMI                   0.6462        1.00      -0.7953       2.0878
## UnderFiveDeaths       0.1492        0.15       0.1306       0.1679
## Polio                 4.0736        4.07       3.5151       4.6322
## TotalExpenditure      0.8408        1.00       0.5348       1.1468
## Diphtheria            4.6398        4.64       3.9554       5.3243
## HIVAIDS              -0.5266       -0.50      -0.6755      -0.3777
## GDP                   0.1249        0.12       0.0482       0.2016
## Population            0.0728        0.07       0.0160       0.1296
## Thin1019              0.1290        0.13       0.0200       0.2380
## Thin59                0.1379        0.14       0.0324       0.2434
## IncomeComposition     1.4136        1.00       0.9392       1.8879
## Schooling             1.4830        1.48       1.1181       1.8480
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                                               LRT df
## LR test, lambda = (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0) 1868.436 18
##                                                                pval
## LR test, lambda = (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0) < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                                               LRT df
## LR test, lambda = (1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1) 8219.701 18
##                                                                pval
## LR test, lambda = (1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1) < 2.22e-16
```

```r
###########################################
# Status
# Our "clean, transformed" data frame
lifeData.df = data.frame(developed = (lifeData.2011$Status == "Developed")) # Dummy variable
```

```
############################################
# AdultMortality
with(lifeData.2011,hist(AdultMortality,50))
```



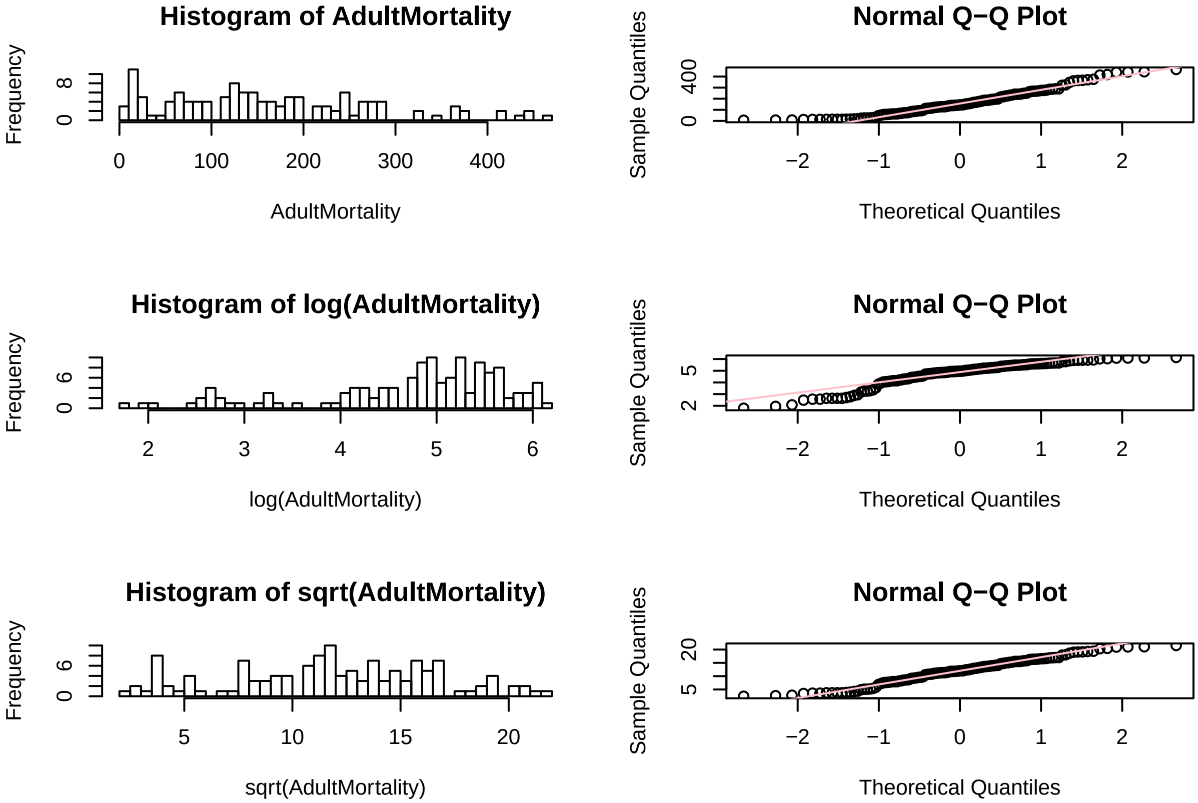
```
plot(lifeData.2011$AdultMortality,lifeData.2011$LifeExpectancy)
```



```r
summary(lm(LifeExpectancy~AdultMortality,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ AdultMortality, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.090  -2.980   1.053   3.639  12.840
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    79.435068   0.950677   83.56   <2e-16 ***
## AdultMortality -0.058564   0.004798  -12.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.128 on 128 degrees of freedom
## Multiple R-squared:  0.5379, Adjusted R-squared:  0.5343
## F-statistic:   149 on 1 and 128 DF,  p-value: < 2.2e-16
```

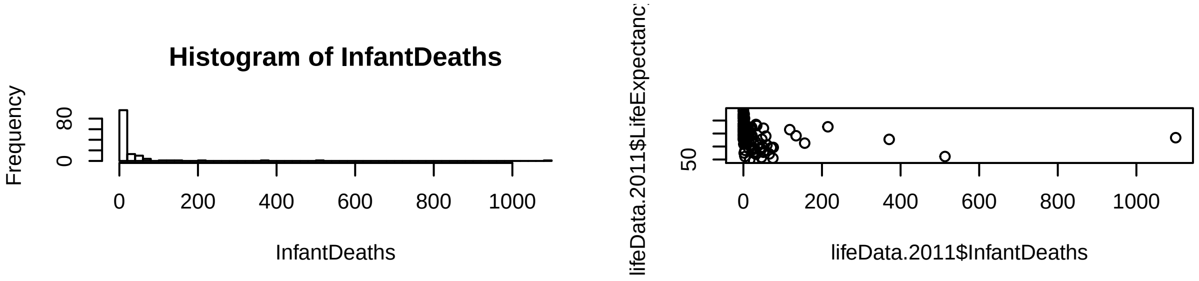
```r
# significant

par(mfrow = c(3,2))
with(lifeData.2011,hist(AdultMortality,50))
with(lifeData.2011,qqnorm(AdultMortality))
with(lifeData.2011,qqline(AdultMortality,col="pink"))
with(lifeData.2011,hist(log(AdultMortality),50))
with(lifeData.2011,qqnorm(log(AdultMortality)))
with(lifeData.2011,qqline(log(AdultMortality),col="pink"))
with(lifeData.2011,hist(sqrt(AdultMortality),50))
with(lifeData.2011,qqnorm(sqrt(AdultMortality)))
with(lifeData.2011,qqline(sqrt(AdultMortality),col="pink"))
```



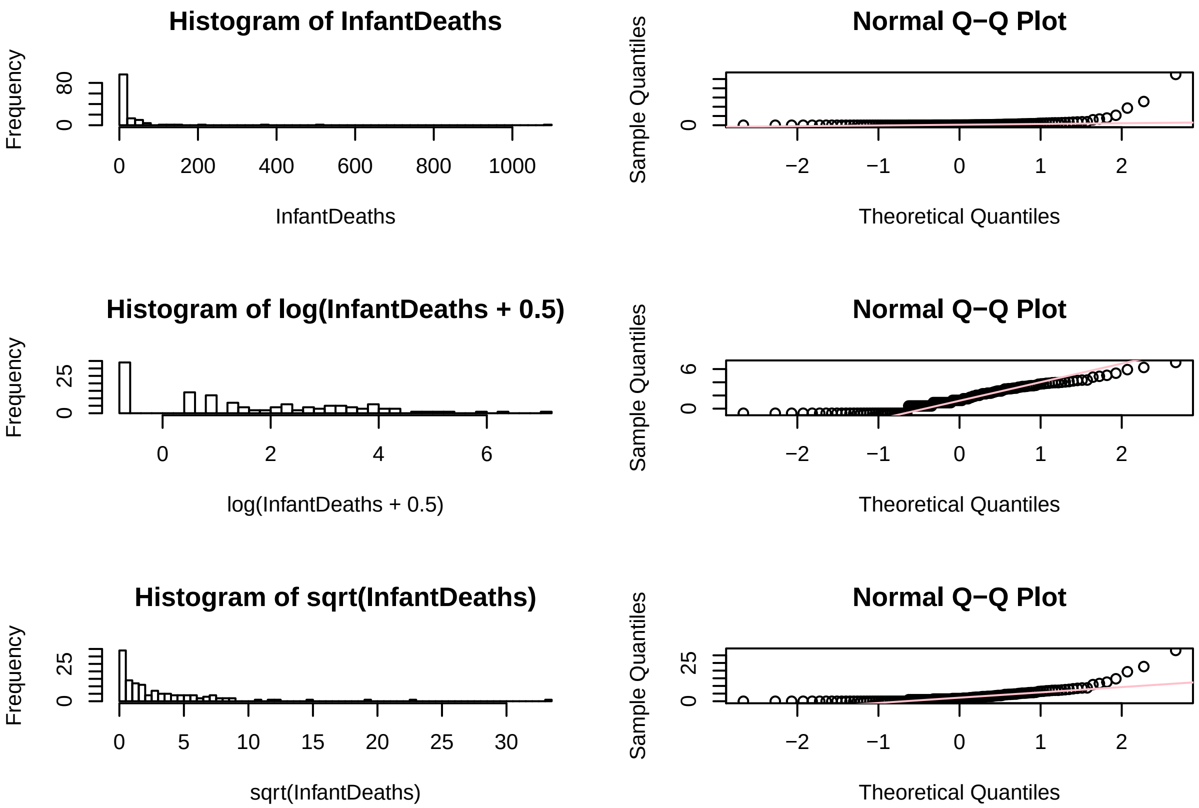
```
# use sqrt

lifeData.df <- cbind(lifeData.df,data.frame(adult.mortality.sqrt=sqrt(lifeData.2011[,'AdultMor

##############################################
# InfantDeaths

with(lifeData.2011,hist(InfantDeaths,50))
plot(lifeData.2011$InfantDeaths,lifeData.2011$LifeExpectancy)
# very skewed

summary(lm(LifeExpectancy~InfantDeaths,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ InfantDeaths, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.051  -7.104   2.348   5.607  17.648
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.352389   0.805899  87.297   <2e-16 ***
## InfantDeaths -0.015436   0.006898  -2.238    0.027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.844 on 128 degrees of freedom
## Multiple R-squared:  0.03764,    Adjusted R-squared:  0.03012
## F-statistic: 5.007 on 1 and 128 DF,  p-value: 0.02698
```

```
# significant

par(mfrow = c(3,2))
```



```
with(lifeData.2011,hist(InfantDeaths,50))
with(lifeData.2011,qqnorm(InfantDeaths))
with(lifeData.2011,qqline(InfantDeaths,col="pink"))
with(lifeData.2011,hist(log(InfantDeaths+0.5),50))
with(lifeData.2011,qqnorm(log(InfantDeaths+0.5)))
with(lifeData.2011,qqline(log(InfantDeaths+0.5),col="pink"))
with(lifeData.2011,hist(sqrt(InfantDeaths),50))
with(lifeData.2011,qqnorm(sqrt(InfantDeaths)))
with(lifeData.2011,qqline(sqrt(InfantDeaths),col="pink"))
```



```
# not sure but log seems to help, kind of confirms powertransform. Need to eliminate zeroes --

#boxTidwell(lifeData.2011$LifeExpectancy~lifeData.2011$InfantDeaths+0.5,other.x=~lifeData.2011

#fit=lm(LifeExpectancy~InfantDeaths,data=lifeData.2011)
#boxcox(fit)

#fit=lm(LifeExpectancy~log(InfantDeaths+0.5),data=lifeData.2011)
#boxcox(fit)

# maybe come back later if there is time
lifeData.df <- cbind(lifeData.df,data.frame(infant.deaths.log=log(lifeData.2011[,'InfantDeaths


############################################
# UnderFiveDeaths

with(lifeData.2011,hist(UnderFiveDeaths,50))
plot(lifeData.2011$LifeExpectancy,lifeData.2011$UnderFiveDeaths)
# very scewed

summary(lm(LifeExpectancy~UnderFiveDeaths,data=lifeData.2011))
```

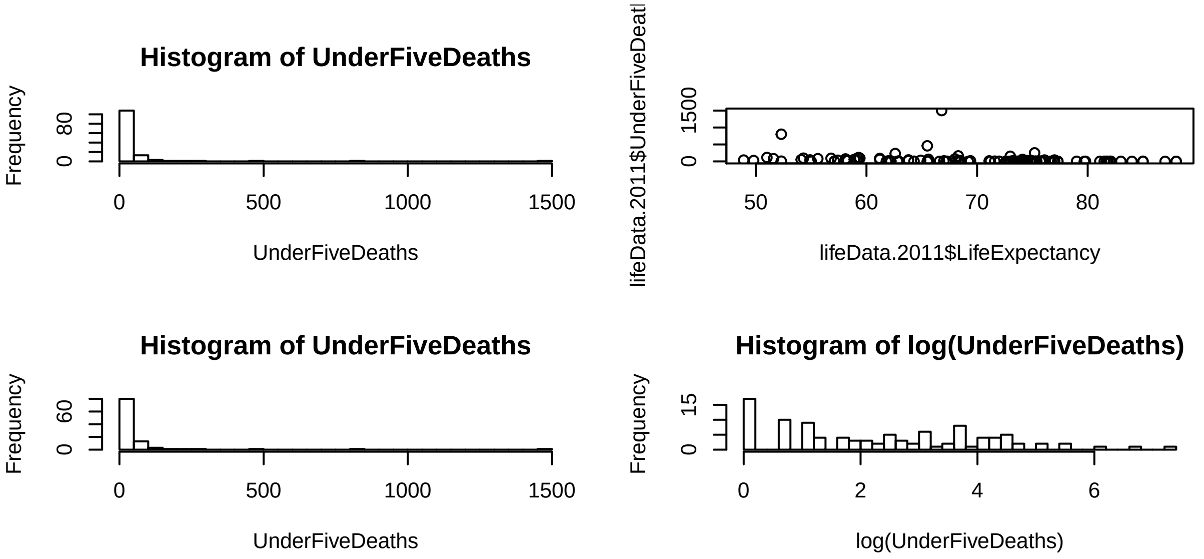
```
##
## Call:
```

```
## lm(formula = LifeExpectancy ~ UnderFiveDeaths, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.034  -7.143   2.296   5.552  17.596
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     70.403635   0.802187  87.765   <2e-16 ***
## UnderFiveDeaths -0.012348   0.004961  -2.489   0.0141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.805 on 128 degrees of freedom
## Multiple R-squared:  0.04616,    Adjusted R-squared:  0.03871
## F-statistic: 6.194 on 1 and 128 DF,  p-value: 0.0141
```

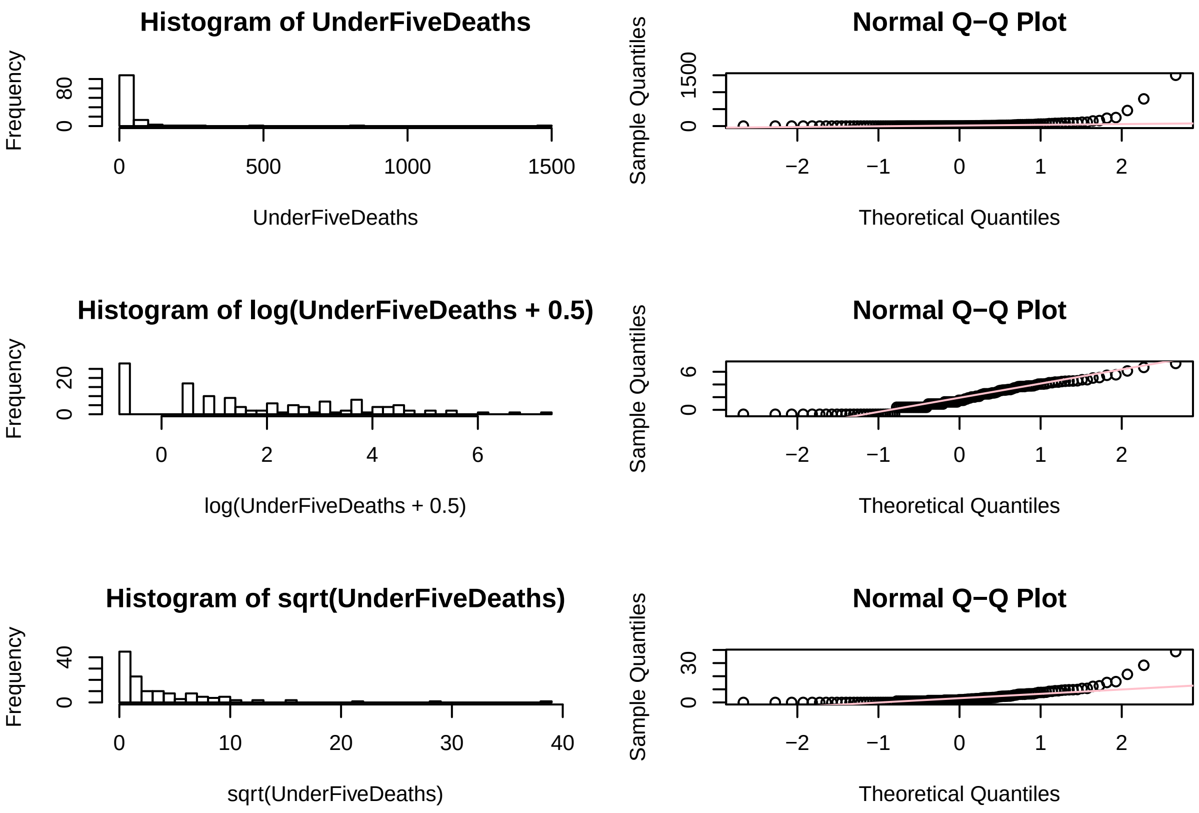
```r
# significant

# bunch of zeroes, let's see how many
with(lifeData.2011,sum(UnderFiveDeaths==0)/length(UnderFiveDeaths)*100) # 22%
```

```
## [1] 21.53846
```

```r
with(subset(lifeData.2011,UnderFiveDeaths!=0),hist(UnderFiveDeaths,50)) #and the pesky tail ag
with(subset(lifeData.2011,UnderFiveDeaths!=0),hist(log(UnderFiveDeaths),50))
# OK, let's just log it

par(mfrow = c(3,2))
```



```r
with(lifeData.2011,hist(UnderFiveDeaths,50))
with(lifeData.2011,qqnorm(UnderFiveDeaths))
with(lifeData.2011,qqline(UnderFiveDeaths,col="pink"))
with(lifeData.2011,hist(log(UnderFiveDeaths+0.5),50))
with(lifeData.2011,qqnorm(log(UnderFiveDeaths+0.5)))
with(lifeData.2011,qqline(log(UnderFiveDeaths+0.5),col="pink"))
with(lifeData.2011,hist(sqrt(UnderFiveDeaths),50))
with(lifeData.2011,qqnorm(sqrt(UnderFiveDeaths)))
with(lifeData.2011,qqline(sqrt(UnderFiveDeaths),col="pink"))
```

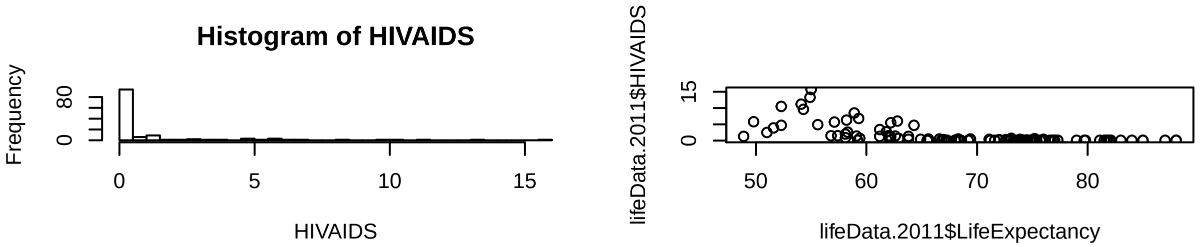


```
# not sure but log seems to help, kind of confirms powertransform. Need to eliminate zeroes --

# maybe come back later if there is time
lifeData.df <- cbind(lifeData.df,data.frame(under.five.deaths.log=log(lifeData.2011[,'UnderFiv


###########################################
# HIVAIDS

with(lifeData.2011,hist(HIVAIDS,50))
plot(lifeData.2011$LifeExpectancy,lifeData.2011$HIVAIDS)
# very scewed

summary(lm(LifeExpectancy~HIVAIDS,data=lifeData.2011))
```

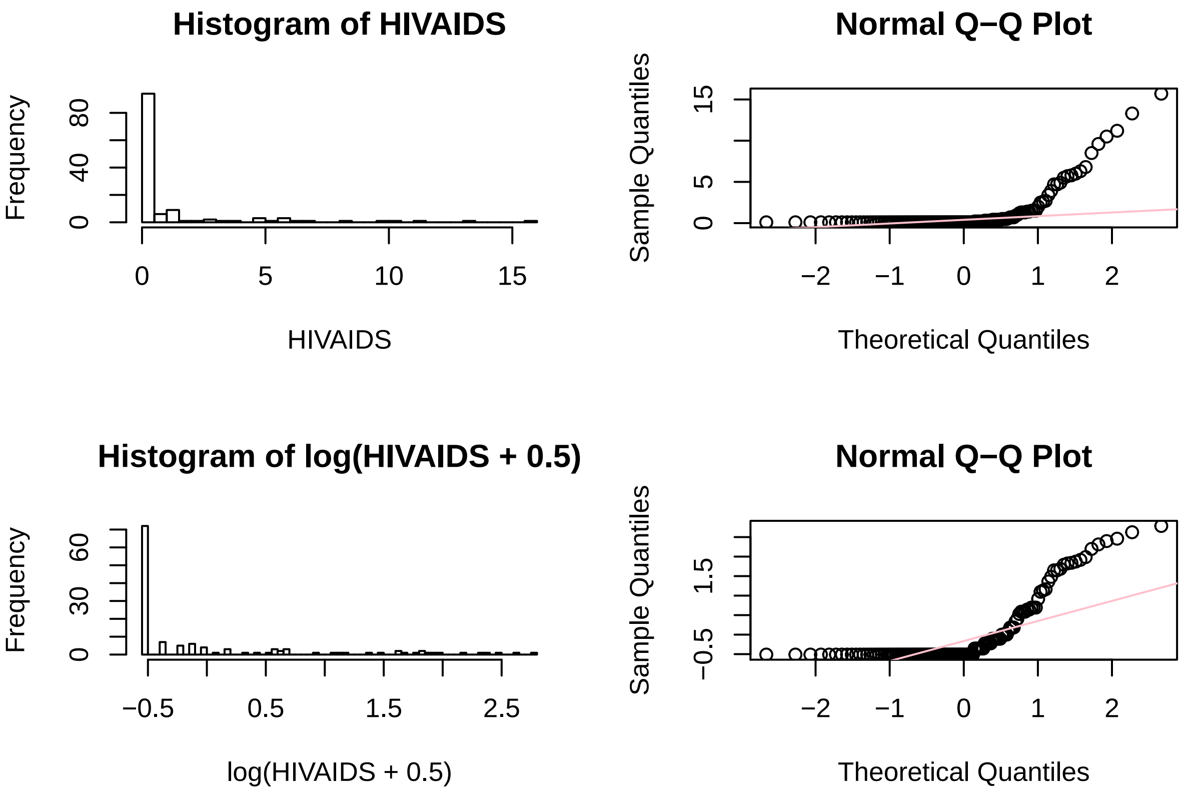
```
##
## Call:
## lm(formula = LifeExpectancy ~ HIVAIDS, data = lifeData.2011)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.944  -4.967   1.273   3.805  15.706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  72.4987     0.6841 105.978  < 2e-16 ***
## HIVAIDS      -2.0419     0.2266  -9.013 2.45e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.051 on 128 degrees of freedom
## Multiple R-squared:  0.3882, Adjusted R-squared:  0.3835
## F-statistic: 81.23 on 1 and 128 DF,  p-value: 2.449e-15
```

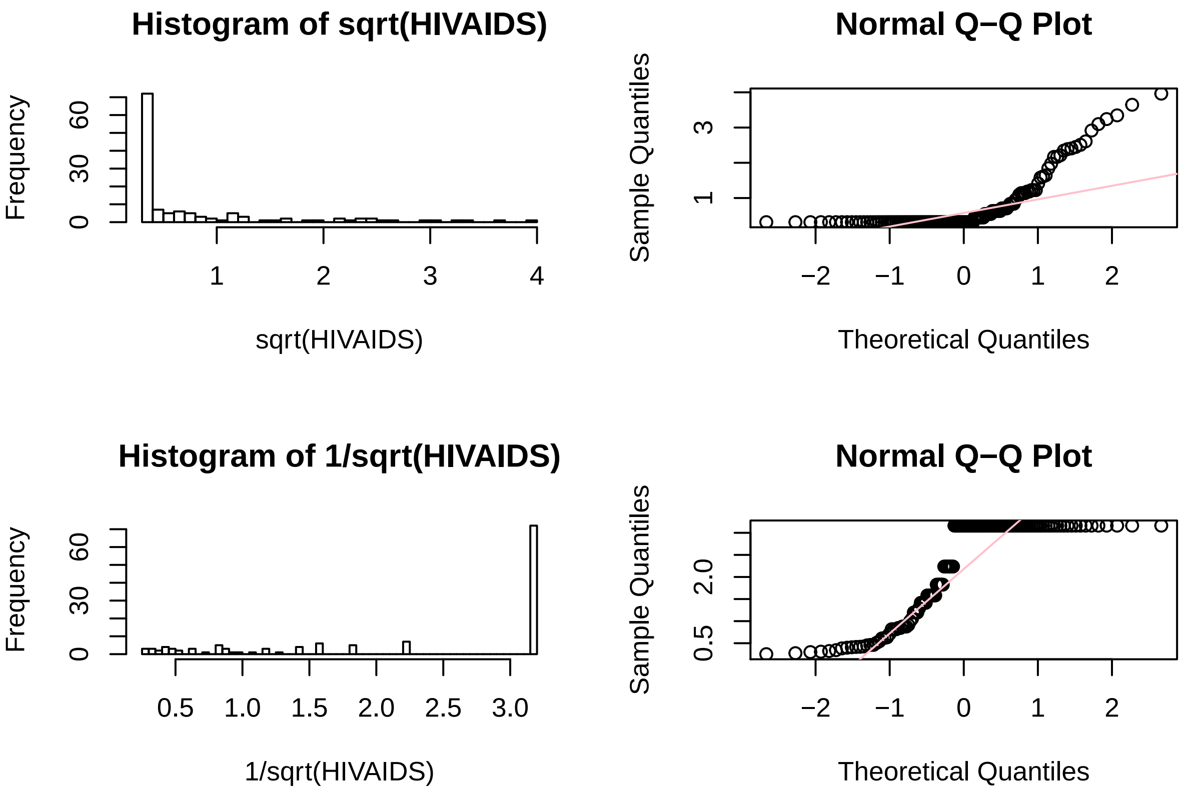
```
# significant

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



```
with(lifeData.2011,hist(HIVAIDS,50))
with(lifeData.2011,qqnorm(HIVAIDS))
with(lifeData.2011,qqline(HIVAIDS,col="pink"))
with(lifeData.2011,hist(log(HIVAIDS+0.5),50))
with(lifeData.2011,qqnorm(log(HIVAIDS+0.5)))
with(lifeData.2011,qqline(log(HIVAIDS+0.5),col="pink"))
```



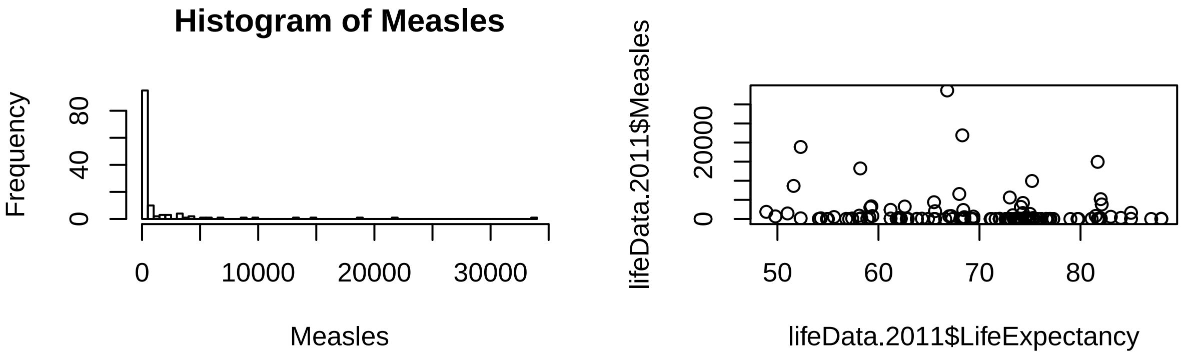
```
with(lifeData.2011,hist(sqrt(HIVAIDS),50))
with(lifeData.2011,qqnorm(sqrt(HIVAIDS)))
with(lifeData.2011,qqline(sqrt(HIVAIDS),col="pink"))
with(lifeData.2011,hist(1/sqrt(HIVAIDS),50))
with(lifeData.2011,qqnorm(1/sqrt(HIVAIDS)))
with(lifeData.2011,qqline(1/sqrt(HIVAIDS),col="pink"))
```



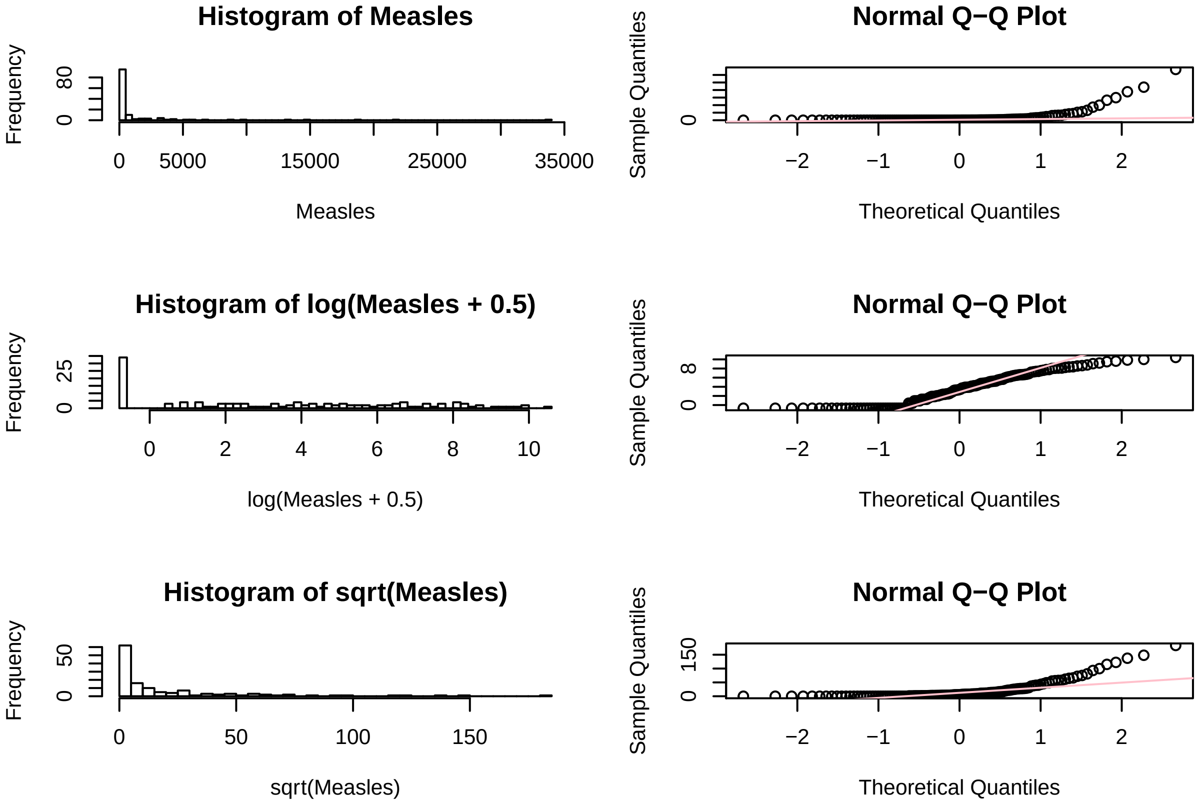
```
# 1/sqrt per powertransform... not looking perfect

# use 1/sqrt
lifeData.df <- cbind(lifeData.df,data.frame(hivaids.1oversqrt=1/sqrt(lifeData.2011[,'HIVAIDS']


##############################################
# Measles

with(lifeData.2011,hist(Measles,50))
plot(lifeData.2011$LifeExpectancy,lifeData.2011$Measles)
# very scewed

summary(lm(LifeExpectancy~Measles,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ Measles, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.866  -7.188   2.632   5.407  17.797
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.2190762  0.8279064  84.815   <2e-16 ***
## Measles     -0.0002431  0.0001795  -1.354    0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.951 on 128 degrees of freedom
## Multiple R-squared:  0.01413,    Adjusted R-squared:  0.006429
## F-statistic: 1.835 on 1 and 128 DF,  p-value: 0.178
```

```
# not significant, skip it

par(mfrow = c(3,2))
```



```
with(lifeData.2011,hist(Measles,50))
with(lifeData.2011,qqnorm(Measles))
with(lifeData.2011,qqline(Measles,col="pink"))
with(lifeData.2011,hist(log(Measles+0.5),50))
with(lifeData.2011,qqnorm(log(Measles+0.5)))
with(lifeData.2011,qqline(log(Measles+0.5),col="pink"))
with(lifeData.2011,hist(sqrt(Measles),50))
with(lifeData.2011,qqnorm(sqrt(Measles)))
with(lifeData.2011,qqline(sqrt(Measles),col="pink"))
```

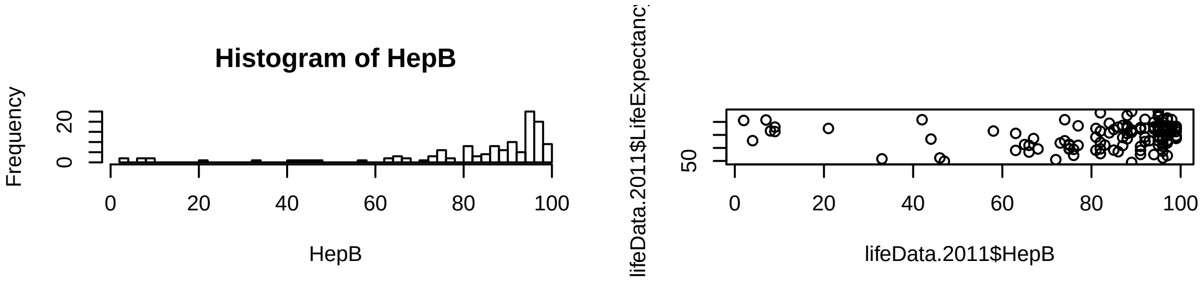


```
# with(lifeData.2011,hist(1/sqrt(Measles),50))
# with(lifeData.2011,qqnorm(1/sqrt(Measles)))
# with(lifeData.2011,qqline(1/sqrt(Measles),col="pink"))
# log it if need to add it back in

# use log
# lifeData.df <- cbind(lifeData.df,data.frame(measles=log(lifeData.2011[,'Measles']+0.5)))


############################################
# HepB

with(lifeData.2011,hist(HepB,50))
plot(lifeData.2011$HepB,lifeData.2011$LifeExpectancy)
# skewed

summary(lm(LifeExpectancy~HepB,data=lifeData.2011))
```

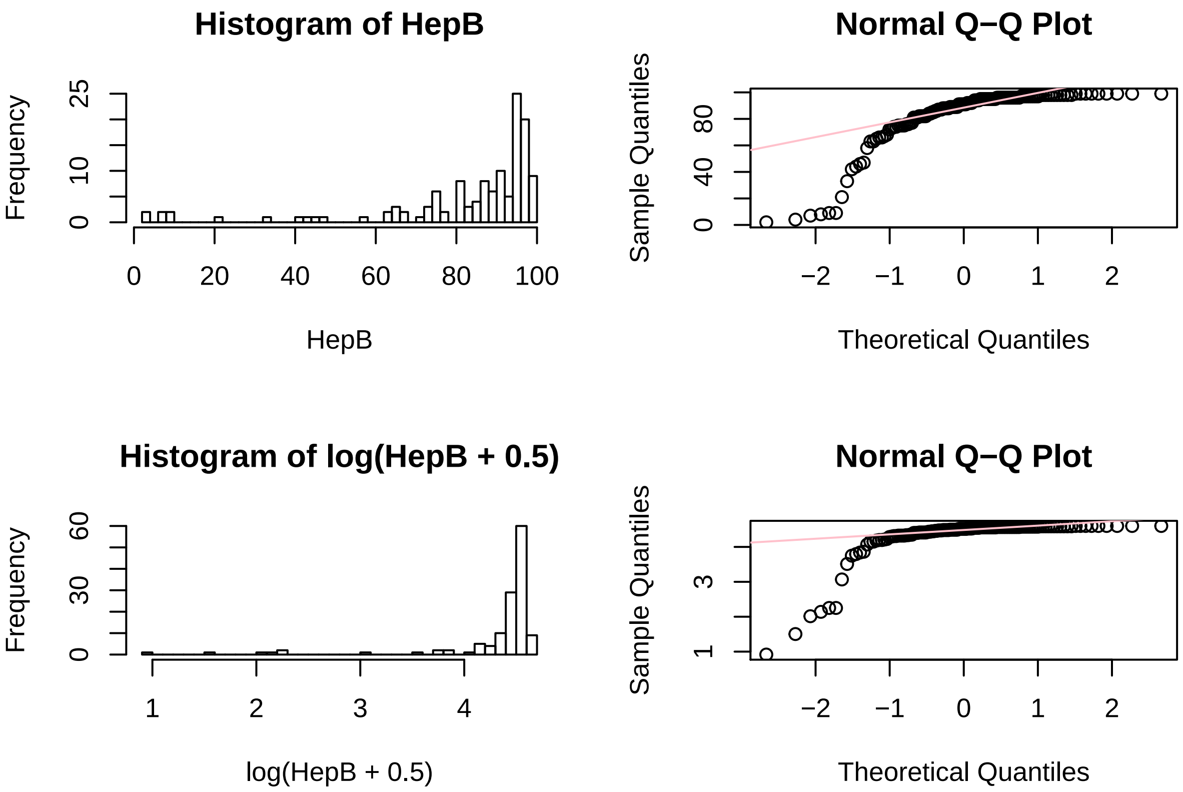
```
##
## Call:
## lm(formula = LifeExpectancy ~ HepB, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.238  -6.703   2.315   5.691  17.863
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.10810    3.05827   21.62   <2e-16 ***
## HepB         0.04527    0.03564    1.27    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.959 on 128 degrees of freedom
## Multiple R-squared:  0.01245,    Adjusted R-squared:  0.004738
## F-statistic: 1.614 on 1 and 128 DF,  p-value: 0.2062
```

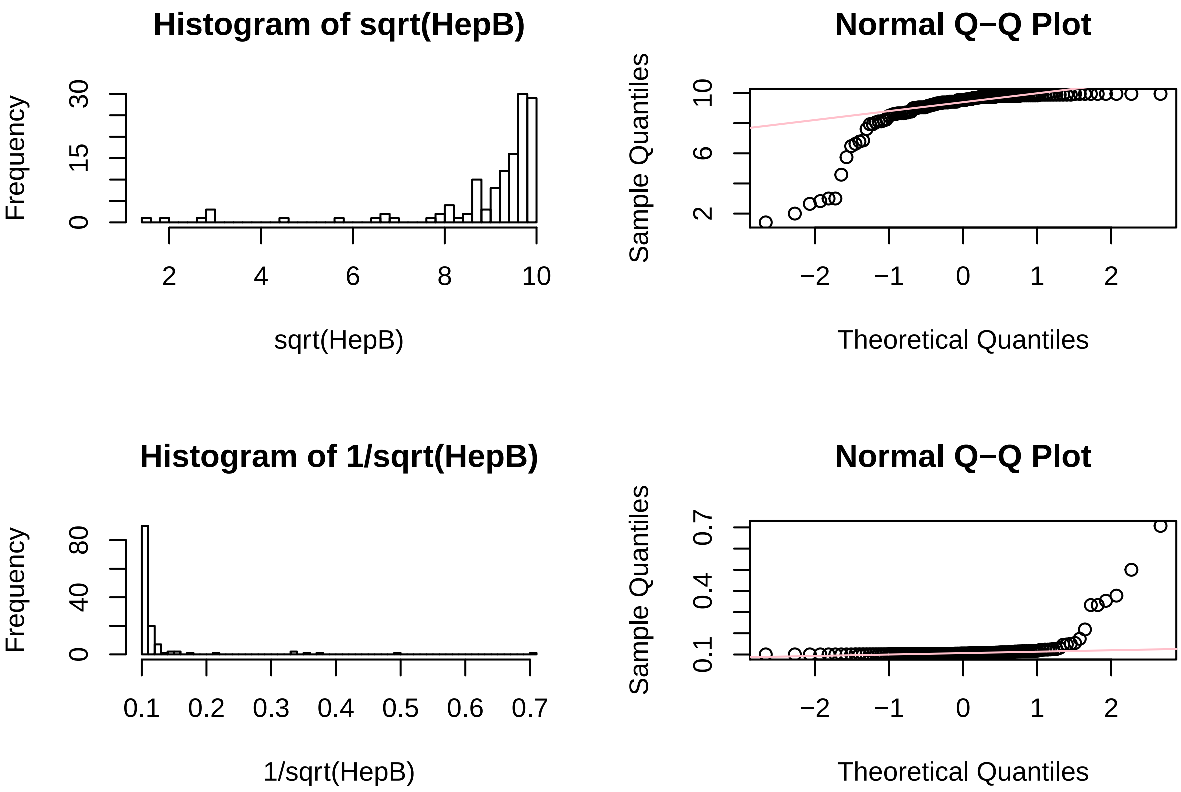
```
# not significant

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



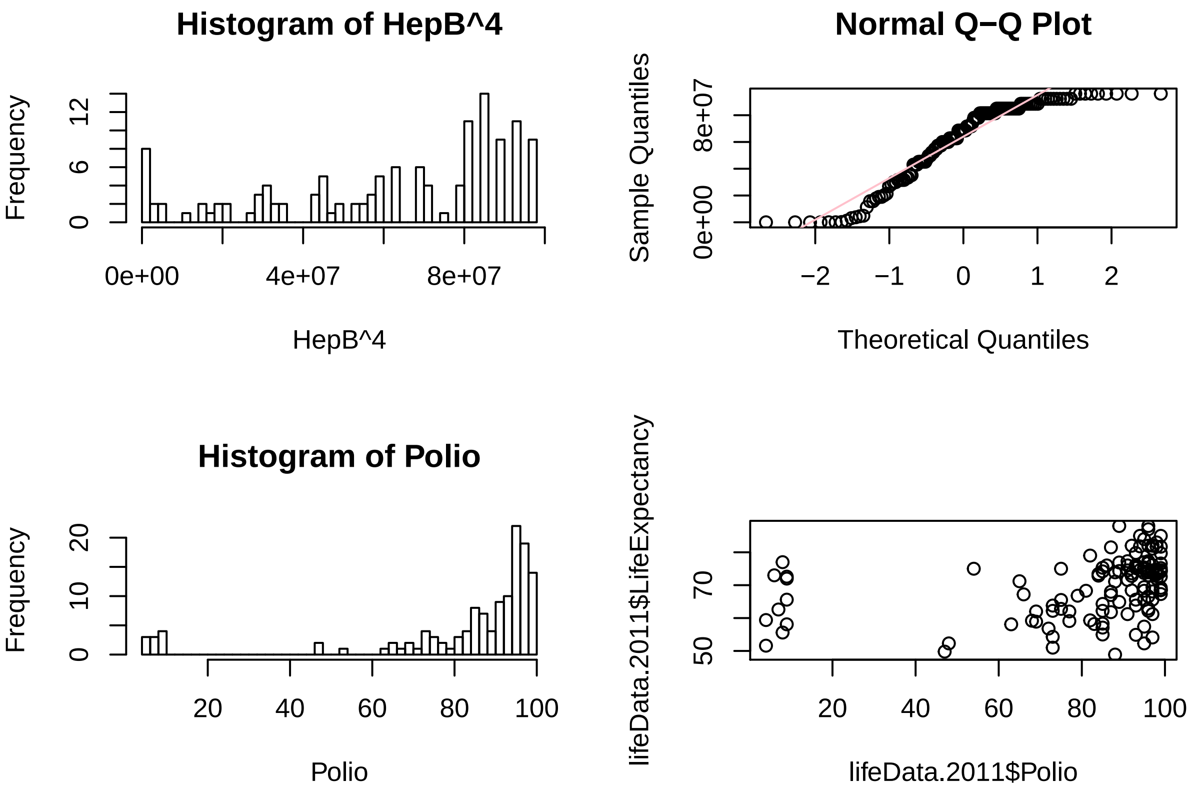
```
with(lifeData.2011,hist(HepB,50))
with(lifeData.2011,qqnorm(HepB))
with(lifeData.2011,qqline(HepB,col="pink"))
with(lifeData.2011,hist(log(HepB+0.5),50))
with(lifeData.2011,qqnorm(log(HepB+0.5)))
with(lifeData.2011,qqline(log(HepB+0.5),col="pink"))
```



```
with(lifeData.2011,hist(sqrt(HepB),50))
with(lifeData.2011,qqnorm(sqrt(HepB)))
with(lifeData.2011,qqline(sqrt(HepB),col="pink"))
with(lifeData.2011,hist(1/sqrt(HepB),50))
with(lifeData.2011,qqnorm(1/sqrt(HepB)))
with(lifeData.2011,qqline(1/sqrt(HepB),col="pink"))
```



```
with(lifeData.2011,hist(HepB^4,50))
with(lifeData.2011,qqnorm(HepB^4))
with(lifeData.2011,qqline(HepB^4,col="pink"))
# ^4 per powertransform... not looking perfect

# use ^4
lifeData.df <- cbind(lifeData.df,data.frame(hepb.power4=(lifeData.2011[,'HepB'])^4))


###########################################
# Polio

with(lifeData.2011,hist(Polio,50))
plot(lifeData.2011$Polio,lifeData.2011$LifeExpectancy)
```

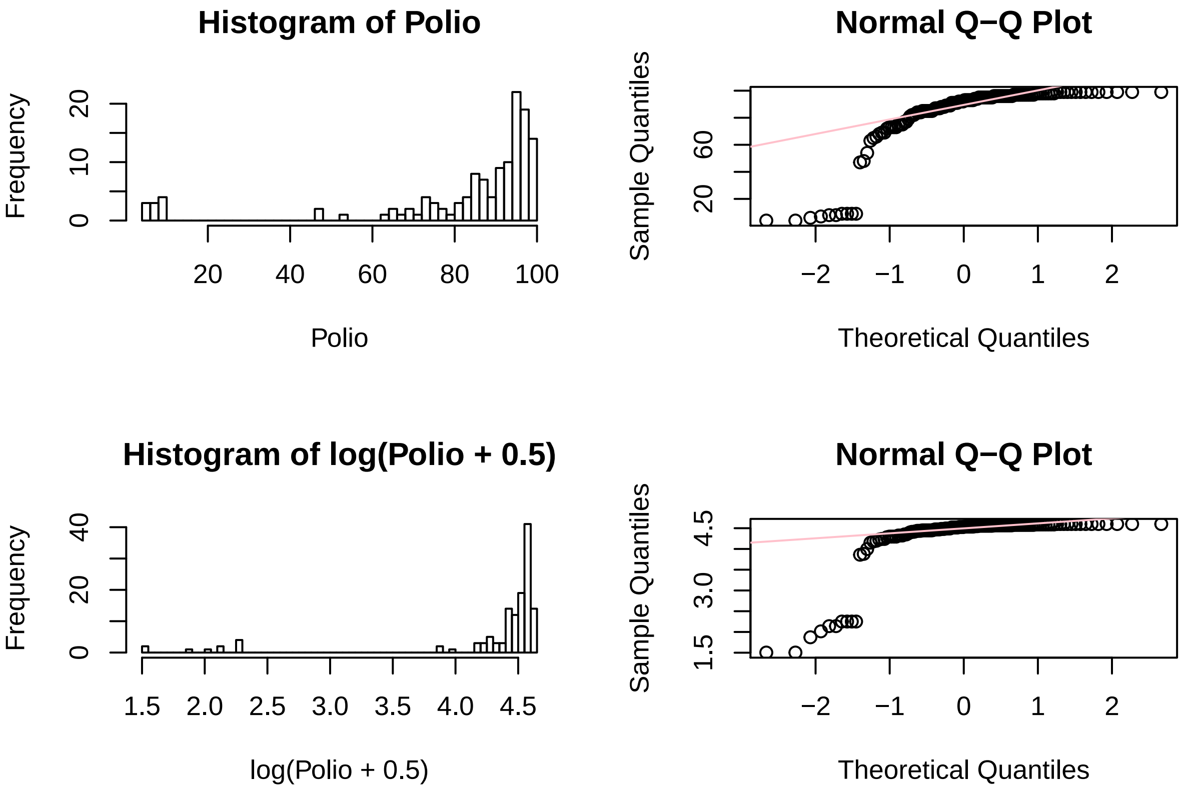


```
# skewed

summary(lm(LifeExpectancy~Polio,data=lifeData.2011))
```

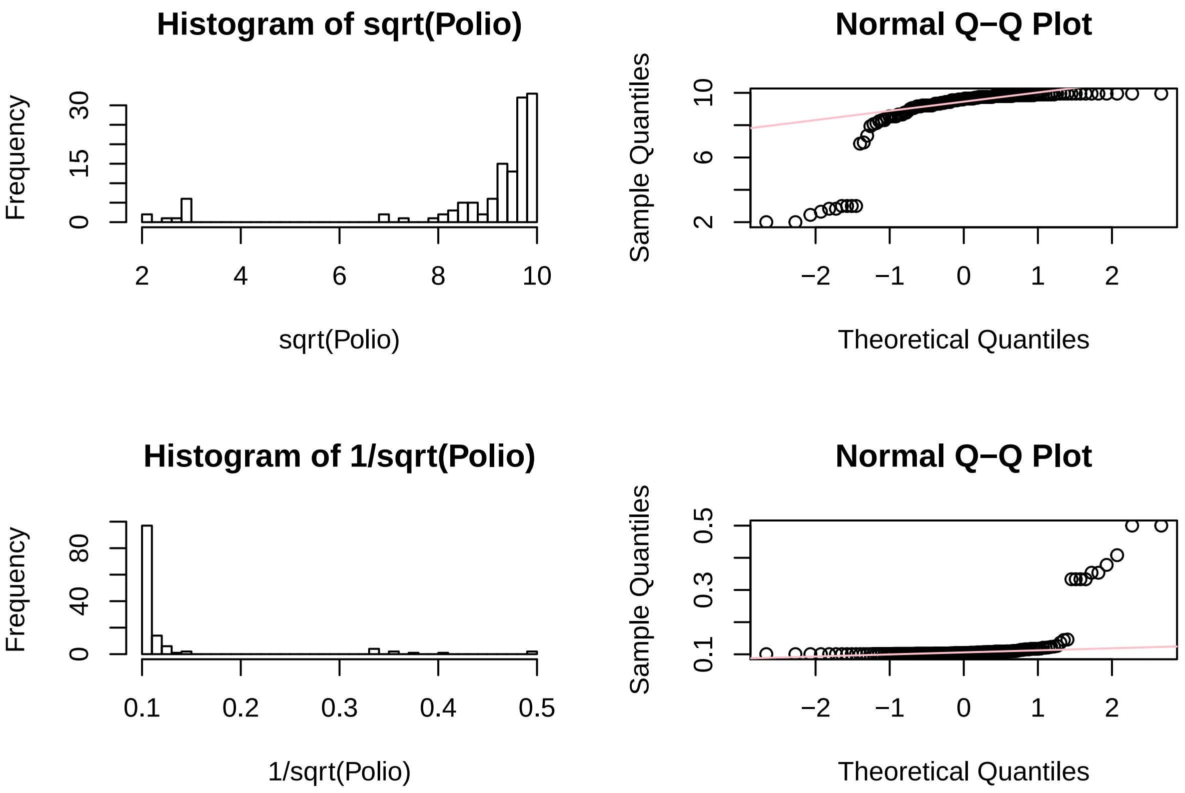
```
##
## Call:
## lm(formula = LifeExpectancy ~ Polio, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.638  -5.926   1.756   5.144  17.331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.04568    2.63689  22.392  < 2e-16 ***
## Polio        0.13060    0.03056   4.274 3.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.434 on 128 degrees of freedom
## Multiple R-squared:  0.1249, Adjusted R-squared:  0.118
## F-statistic: 18.27 on 1 and 128 DF,  p-value: 3.72e-05
```

```
# significant

# powertransform says ^4
```

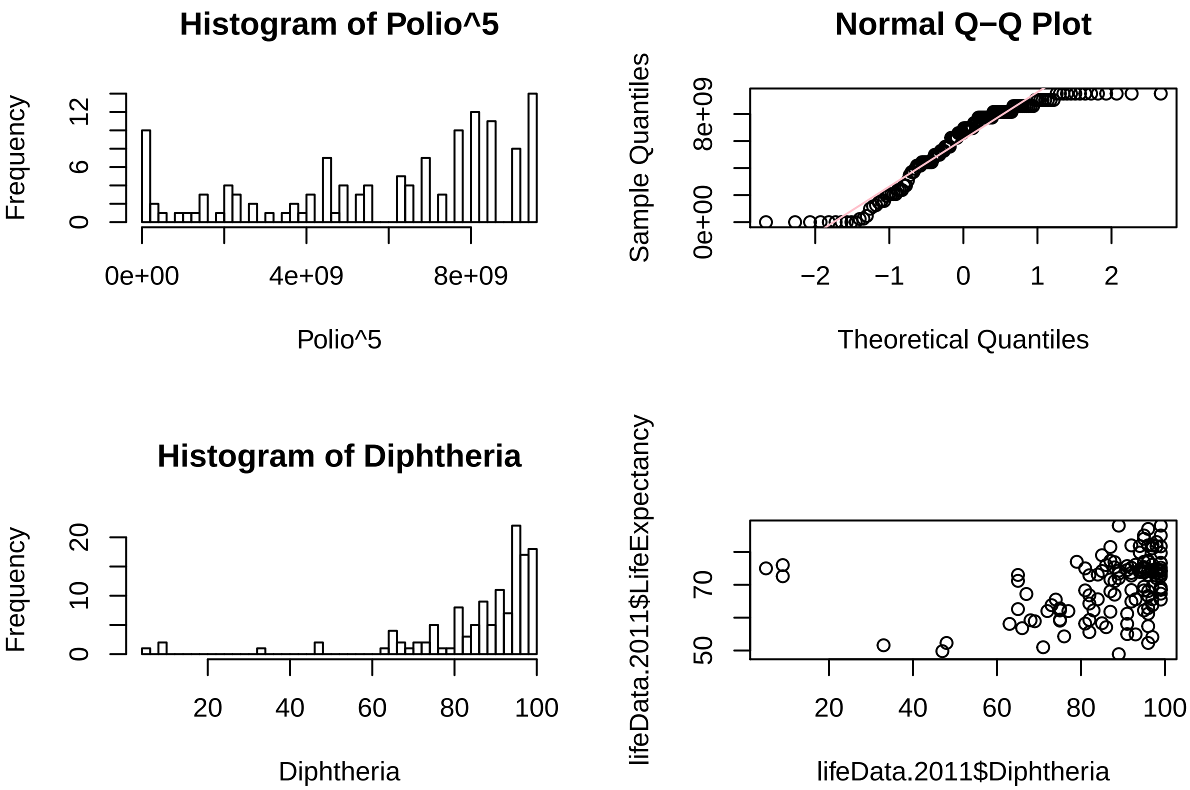
```
par(mfrow = c(2,2))
with(lifeData.2011,hist(Polio,50))
with(lifeData.2011,qqnorm(Polio))
with(lifeData.2011,qqline(Polio,col="pink"))
with(lifeData.2011,hist(log(Polio+0.5),50))
with(lifeData.2011,qqnorm(log(Polio+0.5)))
with(lifeData.2011,qqline(log(Polio+0.5),col="pink"))
```



```
with(lifeData.2011,hist(sqrt(Polio),50))
with(lifeData.2011,qqnorm(sqrt(Polio)))
with(lifeData.2011,qqline(sqrt(Polio),col="pink"))
with(lifeData.2011,hist(1/sqrt(Polio),50))
with(lifeData.2011,qqnorm(1/sqrt(Polio)))
with(lifeData.2011,qqline(1/sqrt(Polio),col="pink"))
```



```
with(lifeData.2011,hist(Polio^5,50))
with(lifeData.2011,qqnorm(Polio^5))
with(lifeData.2011,qqline(Polio^5,col="pink"))
# ^4 per powertransform... not looking perfect

# use ^4
lifeData.df <- cbind(lifeData.df,data.frame(polio.power4=(lifeData.2011[,'Polio'])^4))


##############################################
# Diphtheria

with(lifeData.2011,hist(Diphtheria,50))
plot(lifeData.2011$Diphtheria,lifeData.2011$LifeExpectancy)
```

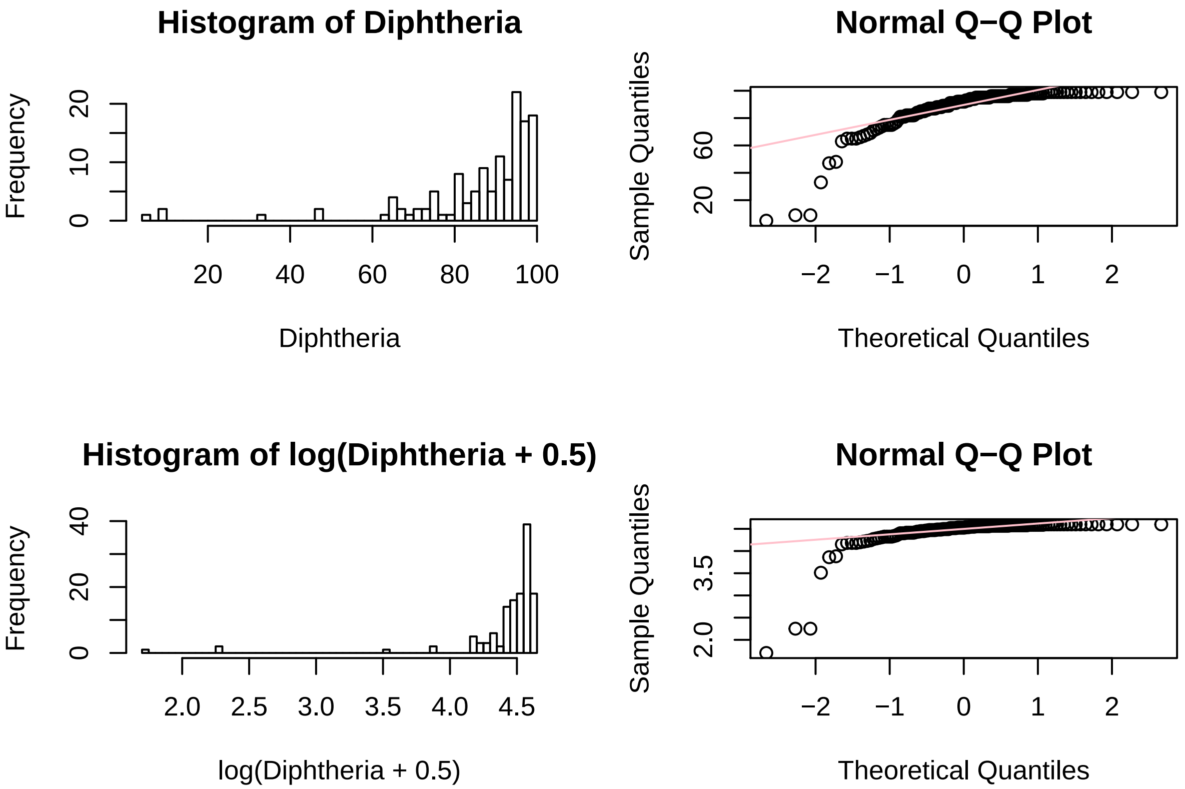


```
# skewed

summary(lm(LifeExpectancy~Diphtheria,data=lifeData.2011))
```

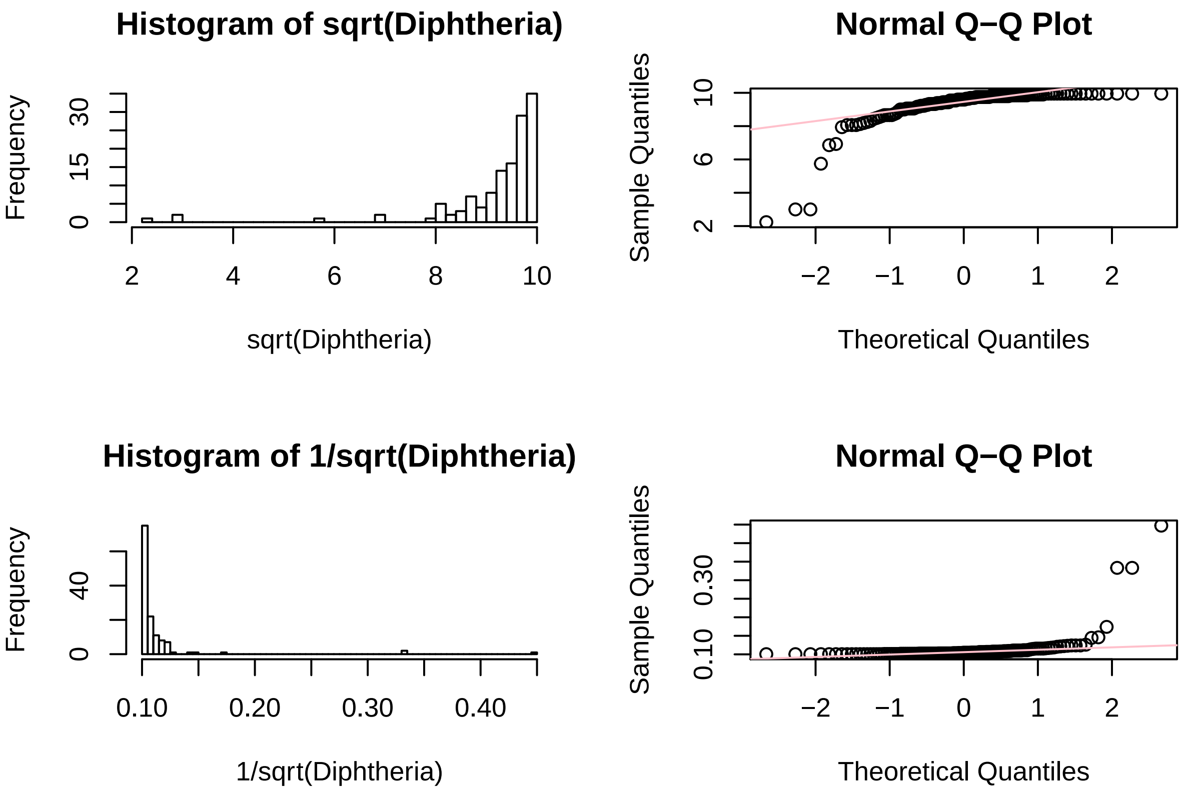
```
##
## Call:
## lm(formula = LifeExpectancy ~ Diphtheria, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.346  -5.928   1.718   5.180  18.869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 55.65550    3.93145  14.156  < 2e-16 ***
## Diphtheria   0.16394    0.04453   3.682  0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.573 on 128 degrees of freedom
## Multiple R-squared:  0.09576,    Adjusted R-squared:  0.0887
## F-statistic: 13.56 on 1 and 128 DF,  p-value: 0.0003402
```

```
# significant

# powertransform says ^5
```

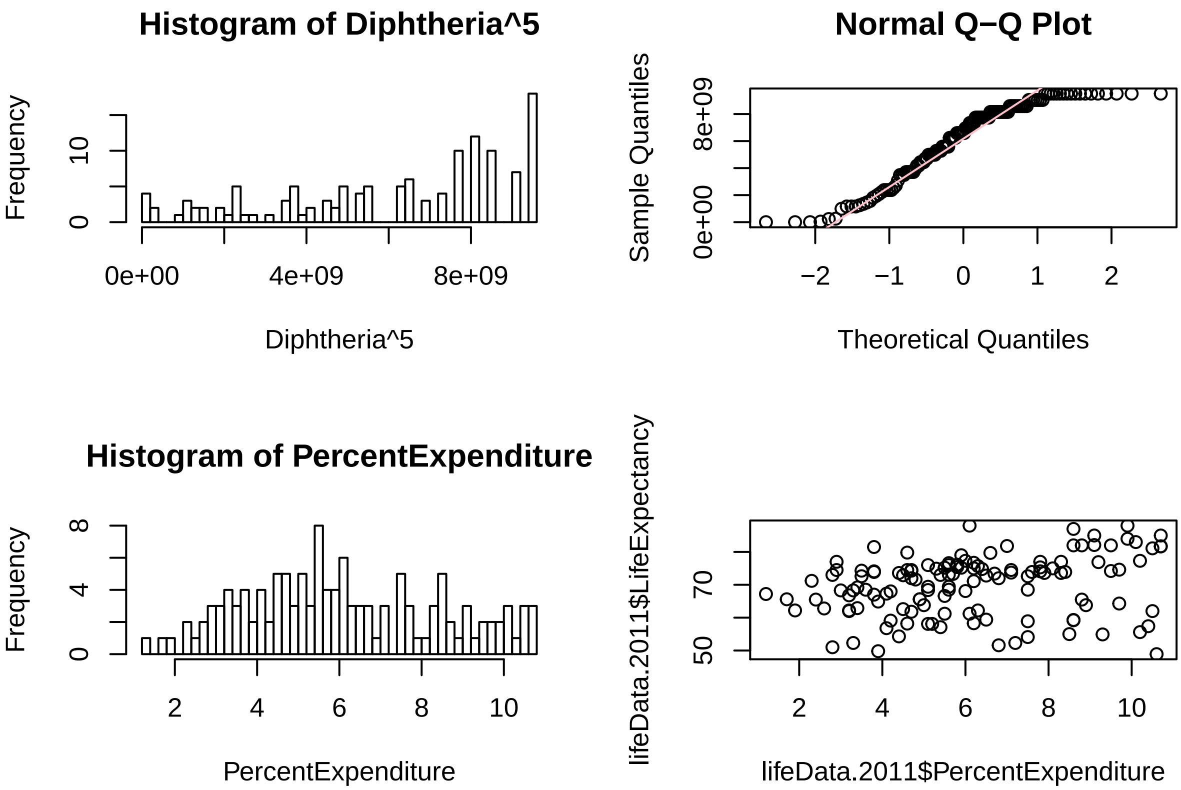
```
par(mfrow = c(2,2))
with(lifeData.2011,hist(Diphtheria,50))
with(lifeData.2011,qqnorm(Diphtheria))
with(lifeData.2011,qqline(Diphtheria,col="pink"))
with(lifeData.2011,hist(log(Diphtheria+0.5),50))
with(lifeData.2011,qqnorm(log(Diphtheria+0.5)))
with(lifeData.2011,qqline(log(Diphtheria+0.5),col="pink"))
```



```
with(lifeData.2011,hist(sqrt(Diphtheria),50))
with(lifeData.2011,qqnorm(sqrt(Diphtheria)))
with(lifeData.2011,qqline(sqrt(Diphtheria),col="pink"))
with(lifeData.2011,hist(1/sqrt(Diphtheria),50))
with(lifeData.2011,qqnorm(1/sqrt(Diphtheria)))
with(lifeData.2011,qqline(1/sqrt(Diphtheria),col="pink"))
```



```
with(lifeData.2011,hist(Diphtheria^5,50))
with(lifeData.2011,qqnorm(Diphtheria^5))
with(lifeData.2011,qqline(Diphtheria^5,col="pink"))
# ^5 per powertransform... not looking perfect

# use ^5
lifeData.df <- cbind(lifeData.df,data.frame(diphtheria.power5=(lifeData.2011[,'Diphtheria'])^5)


############################################
# PercentExpenditure

with(lifeData.2011,hist(PercentExpenditure,50))
plot(lifeData.2011$PercentExpenditure,lifeData.2011$LifeExpectancy)
```

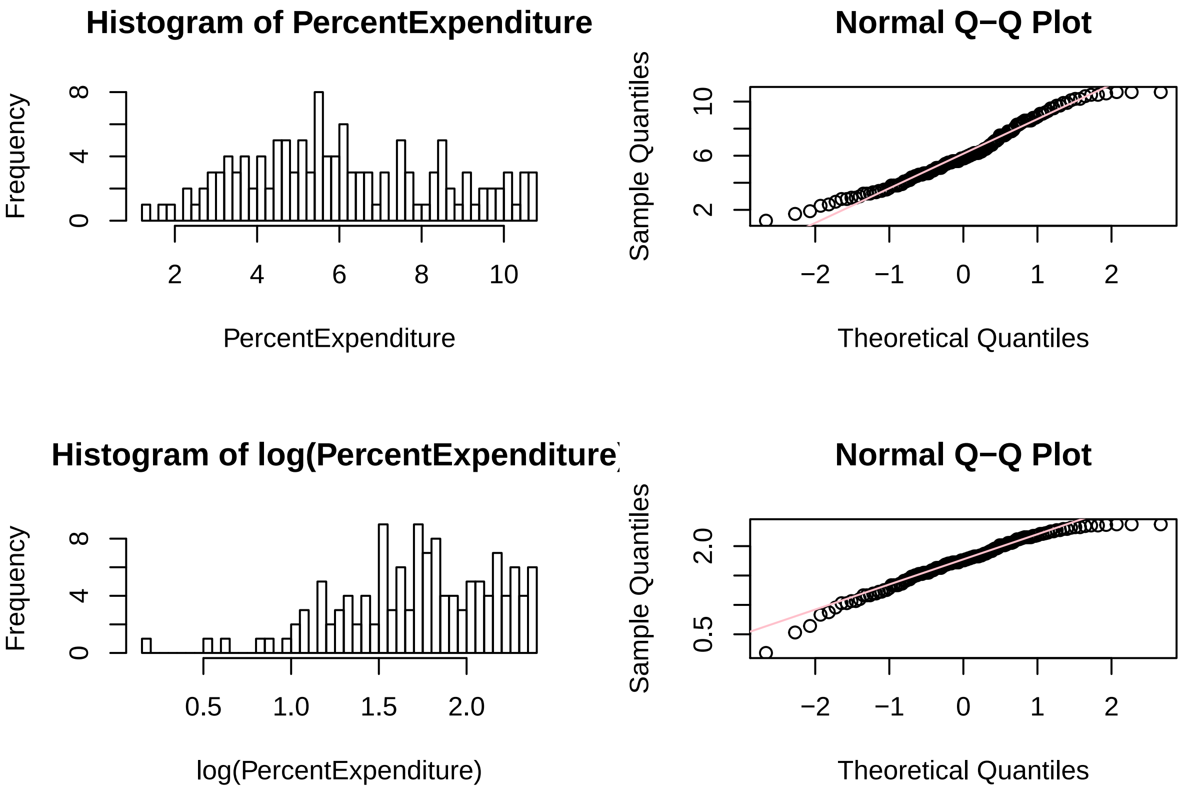


```
# skewed

summary(lm(LifeExpectancy~PercentExpenditure,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ PercentExpenditure, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.145  -5.080   1.942   6.056  18.173
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         64.1092     2.1384  29.980  < 2e-16 ***
## PercentExpenditure   0.9374     0.3252   2.882  0.00463 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.736 on 128 degrees of freedom
## Multiple R-squared:  0.06094,    Adjusted R-squared:  0.05361
## F-statistic: 8.307 on 1 and 128 DF,  p-value: 0.004633
```

```
# significant

# powertransform says sqrt
```

```
par(mfrow = c(2,2))
with(lifeData.2011,hist(PercentExpenditure,50))
with(lifeData.2011,qqnorm(PercentExpenditure))
with(lifeData.2011,qqline(PercentExpenditure,col="pink"))
with(lifeData.2011,hist(log(PercentExpenditure),50))
with(lifeData.2011,qqnorm(log(PercentExpenditure)))
with(lifeData.2011,qqline(log(PercentExpenditure),col="pink"))
```



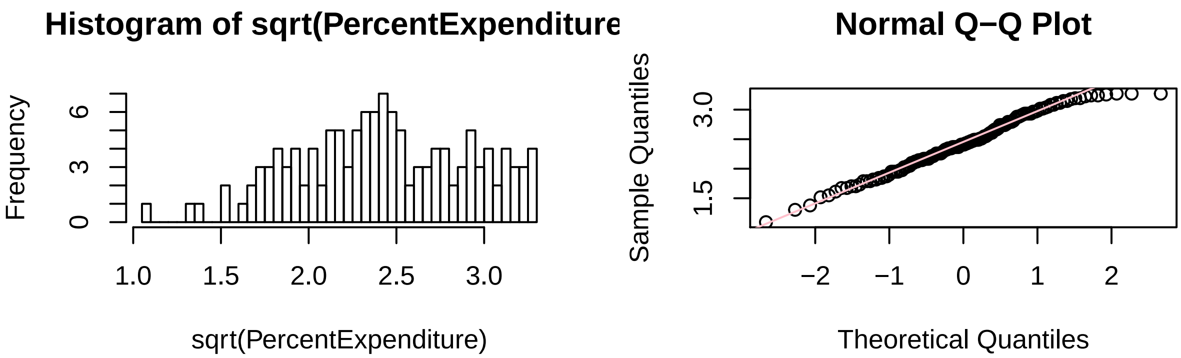
Histogram of PercentExpenditure

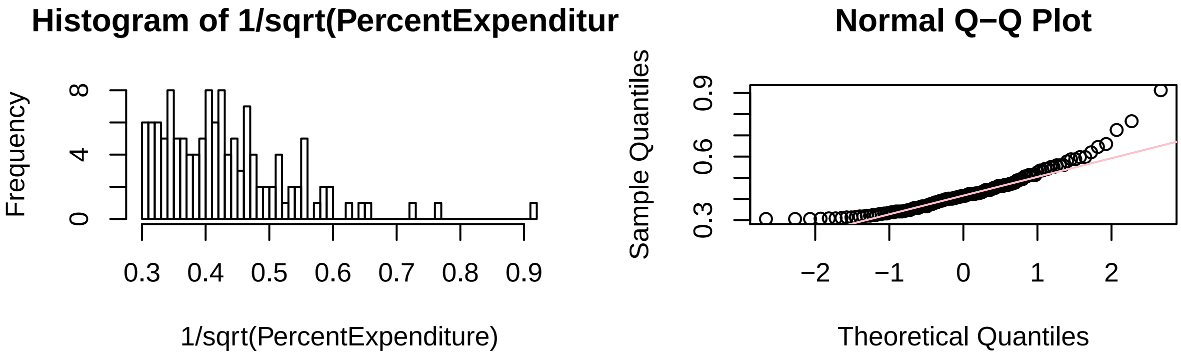
Normal Q−Q Plot

Histogram of log(PercentExpenditure)

Normal Q−Q Plot

```
with(lifeData.2011,hist(sqrt(PercentExpenditure),50))
with(lifeData.2011,qqnorm(sqrt(PercentExpenditure)))
with(lifeData.2011,qqline(sqrt(PercentExpenditure),col="pink"))
with(lifeData.2011,hist(1/sqrt(PercentExpenditure),50))
with(lifeData.2011,qqnorm(1/sqrt(PercentExpenditure)))
with(lifeData.2011,qqline(1/sqrt(PercentExpenditure),col="pink"))
```

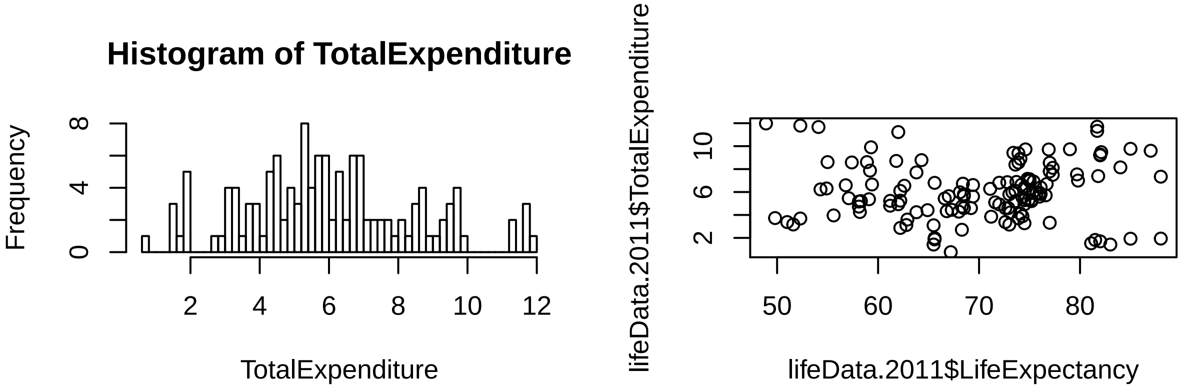




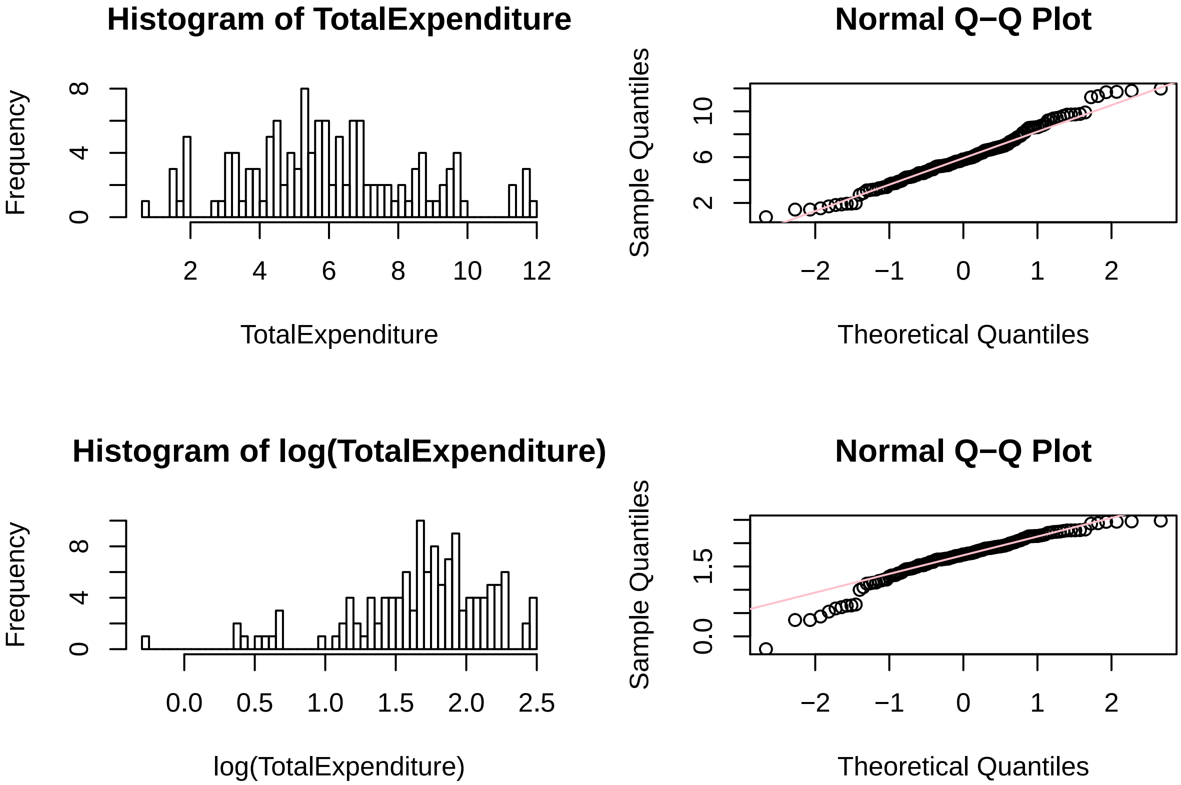
```
# use sqrt
lifeData.df <- cbind(lifeData.df,data.frame(percent.expenditure.sqrt=sqrt(lifeData.2011[,'Perce


############################################
# TotalExpenditure

# no zeroes

with(lifeData.2011,hist(TotalExpenditure,50))
plot(lifeData.2011$LifeExpectancy,lifeData.2011$TotalExpenditure)
# looks good

summary(lm(LifeExpectancy~TotalExpenditure,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ TotalExpenditure, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.963  -6.576   2.879   5.466  18.816
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        68.8594     2.0830  33.059   <2e-16 ***
## TotalExpenditure   0.1672     0.3211   0.521    0.603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.006 on 128 degrees of freedom
## Multiple R-squared:  0.002114,   Adjusted R-squared:  -0.005682
## F-statistic: 0.2712 on 1 and 128 DF,  p-value: 0.6034
```

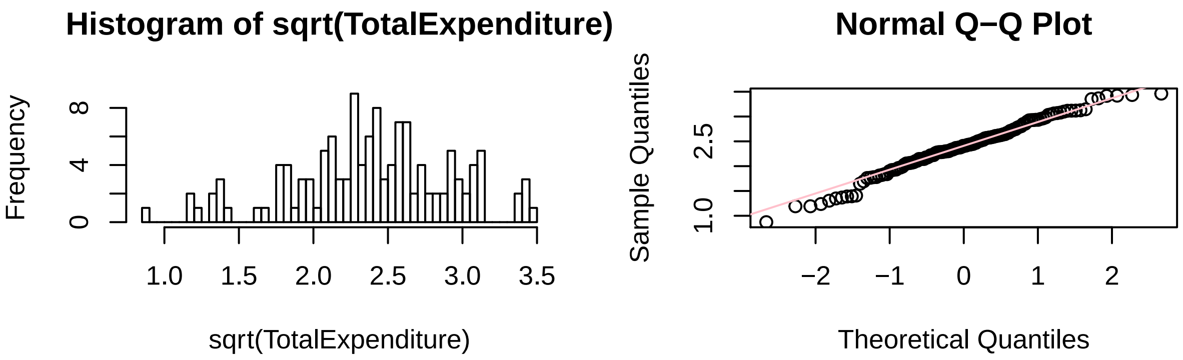
```r
# significant

# powertransform as is
par(mfrow = c(2,2))
```

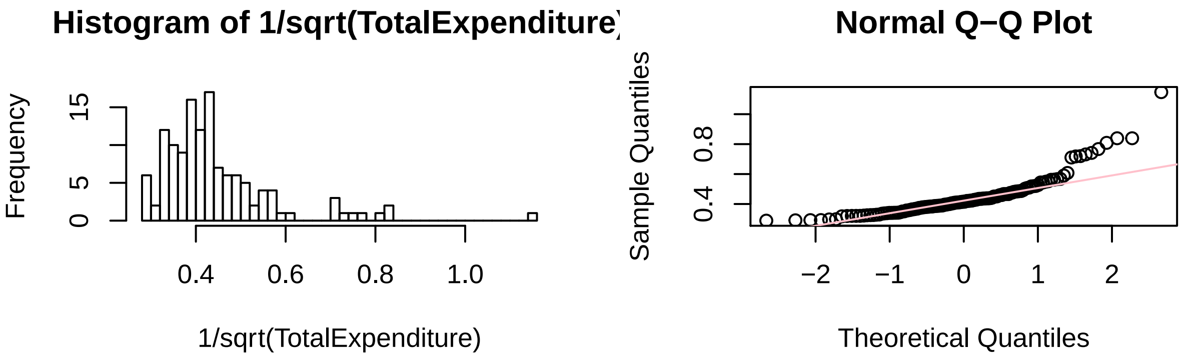


```r
with(lifeData.2011,hist(TotalExpenditure,50))
with(lifeData.2011,qqnorm(TotalExpenditure))
with(lifeData.2011,qqline(TotalExpenditure,col="pink"))
with(lifeData.2011,hist(log(TotalExpenditure),50))
with(lifeData.2011,qqnorm(log(TotalExpenditure)))
with(lifeData.2011,qqline(log(TotalExpenditure),col="pink"))
```



```
with(lifeData.2011,hist(sqrt(TotalExpenditure),50))
with(lifeData.2011,qqnorm(sqrt(TotalExpenditure)))
with(lifeData.2011,qqline(sqrt(TotalExpenditure),col="pink"))
with(lifeData.2011,hist(1/sqrt(TotalExpenditure),50))
with(lifeData.2011,qqnorm(1/sqrt(TotalExpenditure)))
with(lifeData.2011,qqline(1/sqrt(TotalExpenditure),col="pink"))
```

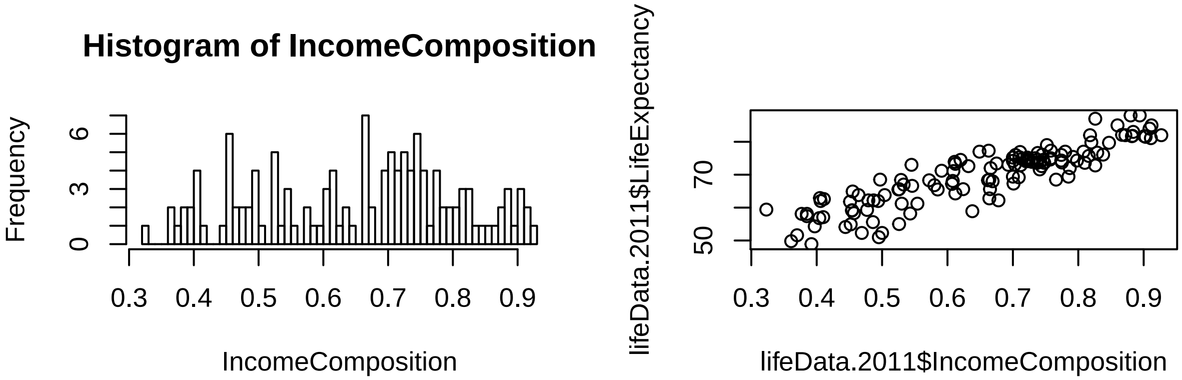




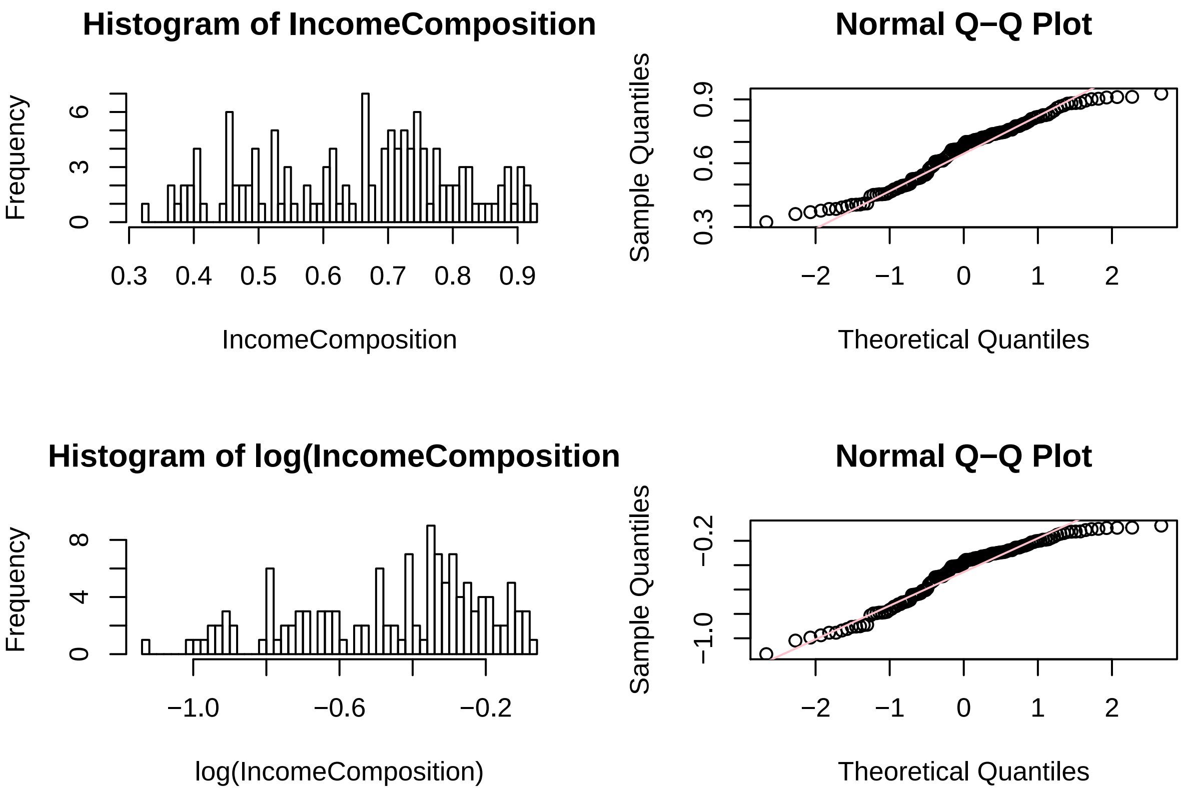
```r
# use as is
lifeData.df <- cbind(lifeData.df,data.frame(total.expenditure=lifeData.2011[,'TotalExpenditure


############################################
# IncomeComposition

# no zeroes

with(lifeData.2011,hist(IncomeComposition,50))
plot(lifeData.2011$IncomeComposition,lifeData.2011$LifeExpectancy)
# wow

summary(lm(LifeExpectancy~IncomeComposition,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ IncomeComposition, data = lifeData.2011)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6456  -2.2784   0.2695   2.3971   8.7735
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          36.095      1.537   23.48   <2e-16 ***
## IncomeComposition   51.617      2.287   22.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.039 on 128 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:  0.7977
## F-statistic: 509.6 on 1 and 128 DF,  p-value: < 2.2e-16
```

```
# significant

# powertransform says as is
par(mfrow = c(2,2))
```



```
with(lifeData.2011,hist(IncomeComposition,50))
with(lifeData.2011,qqnorm(IncomeComposition))
with(lifeData.2011,qqline(IncomeComposition,col="pink"))
with(lifeData.2011,hist(log(IncomeComposition),50))
with(lifeData.2011,qqnorm(log(IncomeComposition)))
with(lifeData.2011,qqline(log(IncomeComposition),col="pink"))
```

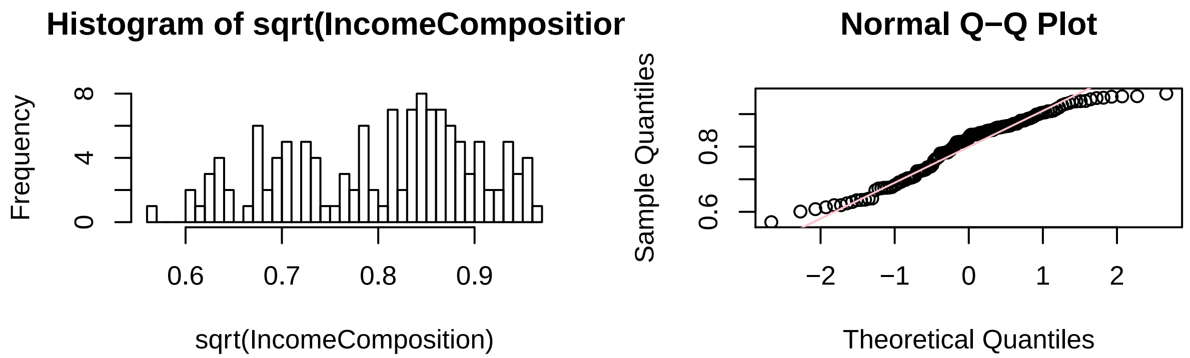

Histogram of IncomeComposition

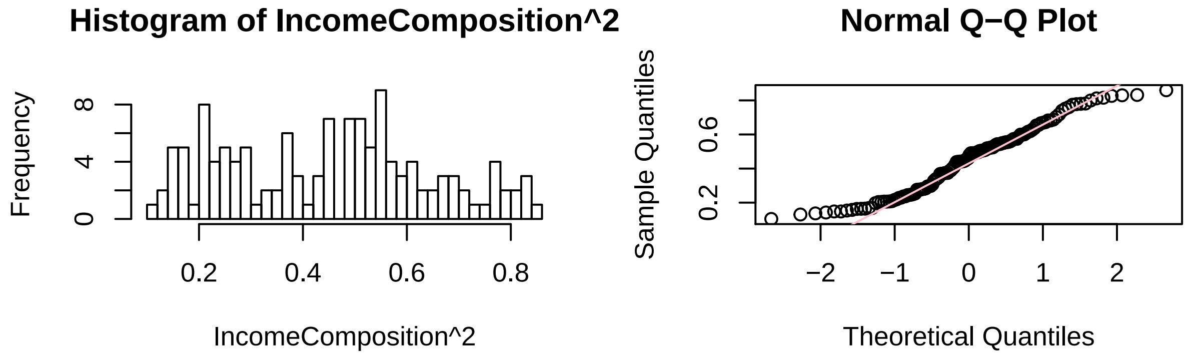
Normal Q−Q Plot

Histogram of log(IncomeComposition

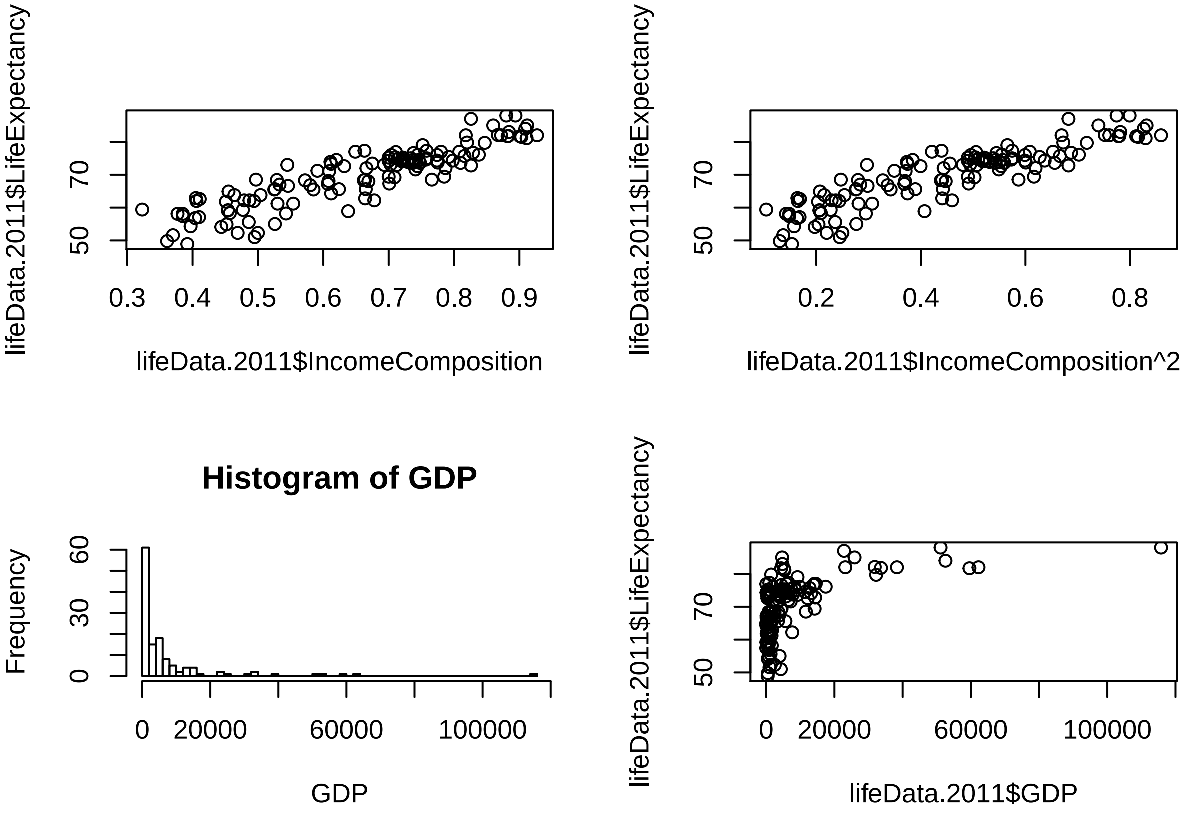
Normal Q−Q Plot

```
with(lifeData.2011,hist(sqrt(IncomeComposition),50))
with(lifeData.2011,qqnorm(sqrt(IncomeComposition)))
with(lifeData.2011,qqline(sqrt(IncomeComposition),col="pink"))
with(lifeData.2011,hist(IncomeComposition^2,50))
with(lifeData.2011,qqnorm(IncomeComposition^2))
with(lifeData.2011,qqline(IncomeComposition^2,col="pink"))
```





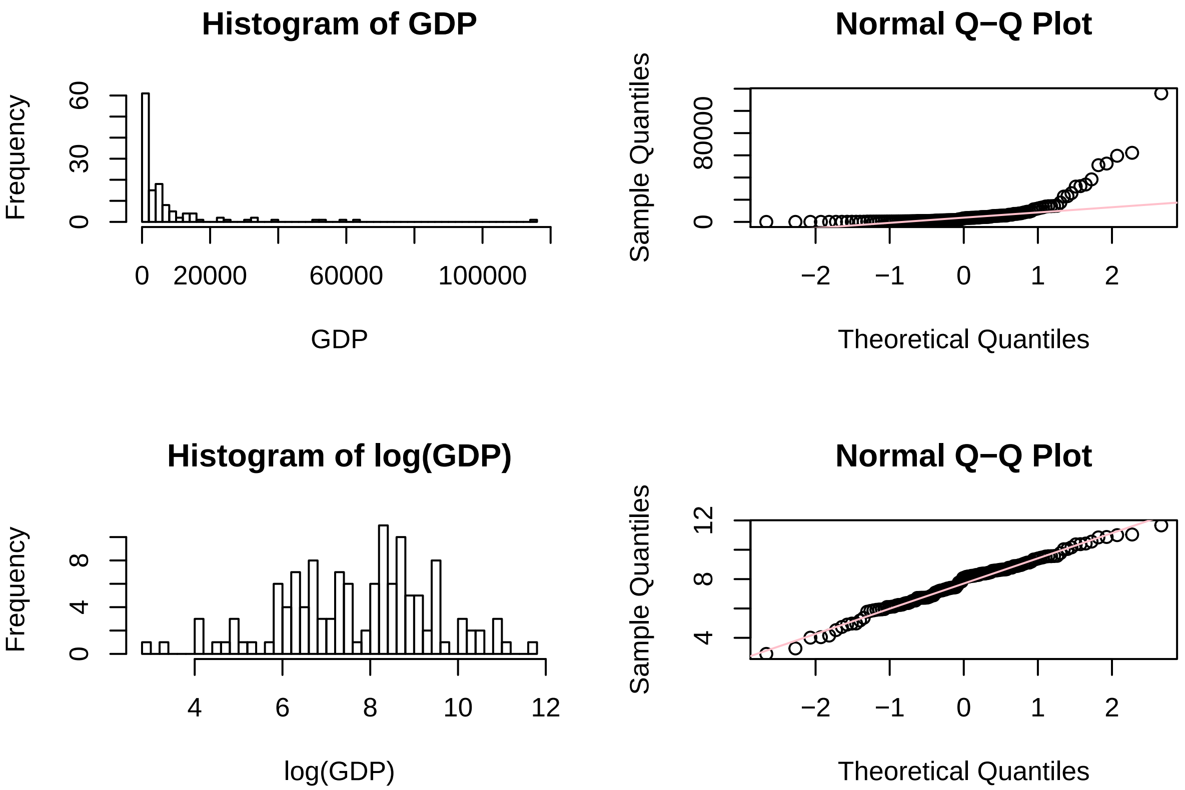
```
# hmm

plot(lifeData.2011$IncomeComposition,lifeData.2011$LifeExpectancy)
plot(lifeData.2011$IncomeComposition^2,lifeData.2011$LifeExpectancy)


# use as is
lifeData.df <- cbind(lifeData.df,data.frame(income.composition.power2=lifeData.2011[,'IncomeCo


###########################################
# GDP

# no zeroes

with(lifeData.2011,hist(GDP,50))
plot(lifeData.2011$GDP,lifeData.2011$LifeExpectancy)
```



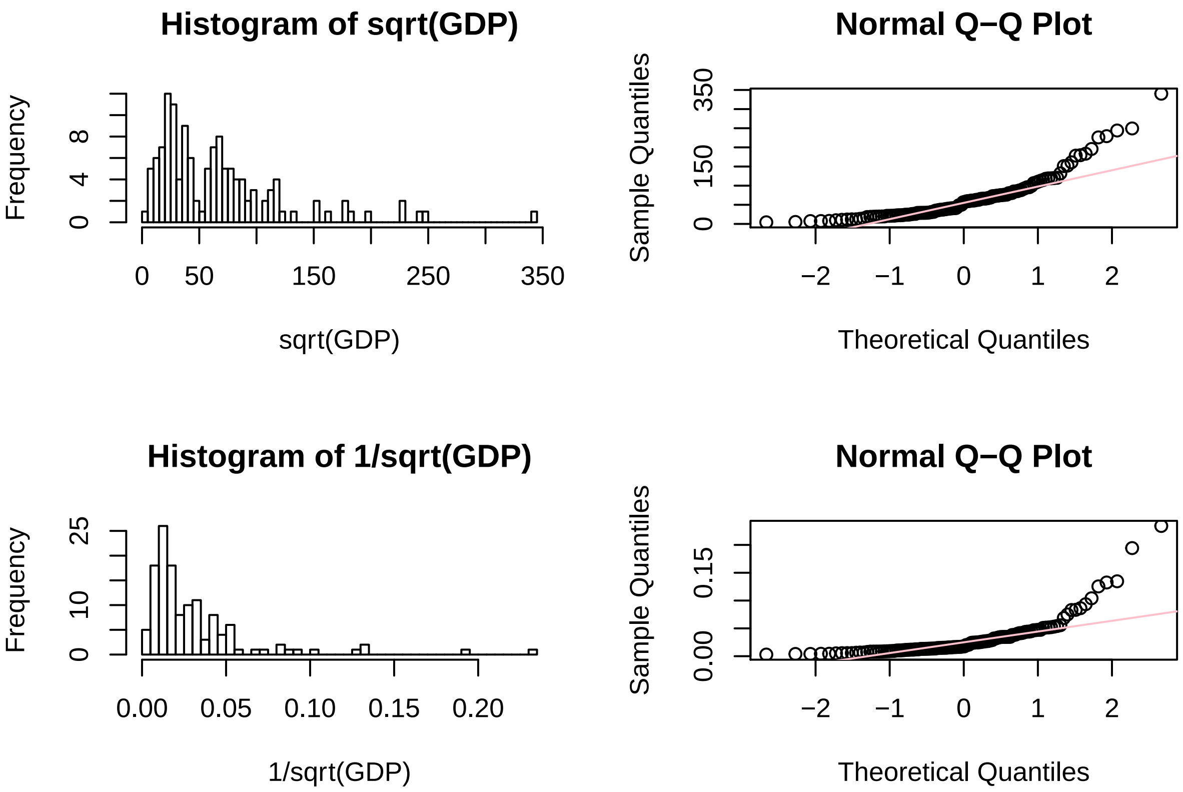
```
# skewed

summary(lm(LifeExpectancy~GDP,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ GDP, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.689  -5.204   1.576   5.481  16.066
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.745e+01  7.552e-01  89.308  < 2e-16 ***
## GDP         3.175e-04  4.528e-05   7.011 1.21e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.663 on 128 degrees of freedom
## Multiple R-squared:  0.2775, Adjusted R-squared:  0.2718
## F-statistic: 49.15 on 1 and 128 DF,  p-value: 1.215e-10
```

```
# significant
# powertransform says log
par(mfrow = c(2,2))
with(lifeData.2011,hist(GDP,50))
with(lifeData.2011,qqnorm(GDP))
with(lifeData.2011,qqline(GDP,col="pink"))
with(lifeData.2011,hist(log(GDP),50))
with(lifeData.2011,qqnorm(log(GDP)))
with(lifeData.2011,qqline(log(GDP),col="pink"))
```



```
with(lifeData.2011,hist(sqrt(GDP),50))
with(lifeData.2011,qqnorm(sqrt(GDP)))
with(lifeData.2011,qqline(sqrt(GDP),col="pink"))
with(lifeData.2011,hist(1/sqrt(GDP),50))
with(lifeData.2011,qqnorm(1/sqrt(GDP)))
with(lifeData.2011,qqline(1/sqrt(GDP),col="pink"))
```

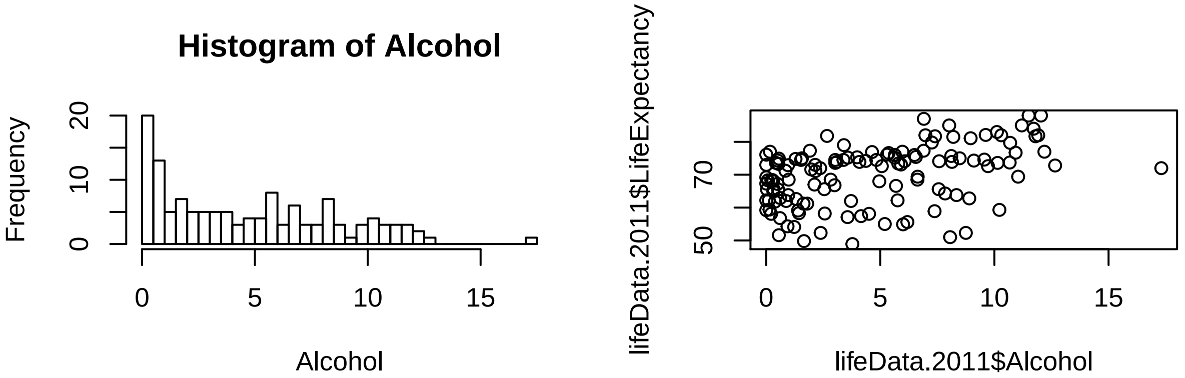


```r
# clearly log

# use log
lifeData.df <- cbind(lifeData.df,data.frame(gdp.log=log(lifeData.2011[,'GDP'])))


############################################
# Alcohol

#  no zeroes

with(lifeData.2011,hist(Alcohol,50))
plot(lifeData.2011$Alcohol,lifeData.2011$LifeExpectancy)
# skewed

summary(lm(LifeExpectancy~Alcohol,data=lifeData.2011))
```

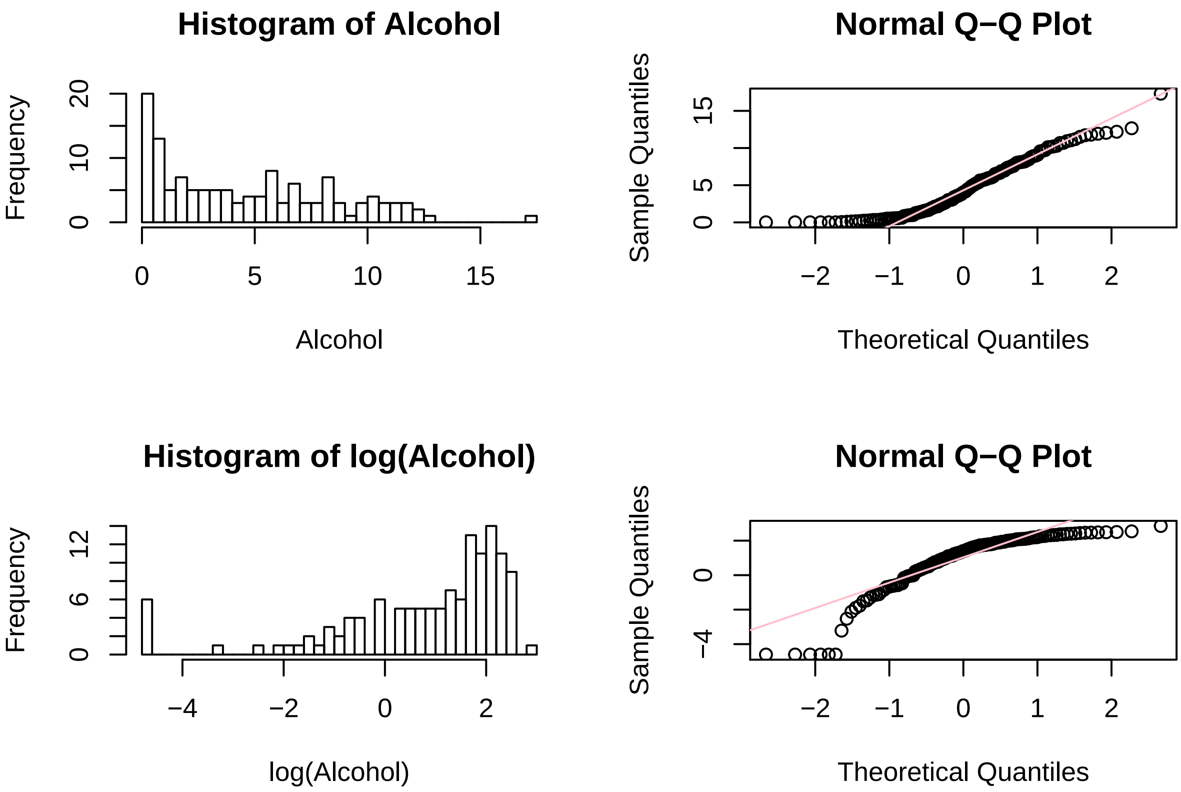
```
##
## Call:
## lm(formula = LifeExpectancy ~ Alcohol, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.110  -5.452   2.041   6.057  15.006
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.2847     1.1268  57.939  < 2e-16 ***
## Alcohol       0.9709     0.1844   5.266 5.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 128 degrees of freedom
## Multiple R-squared:  0.1781, Adjusted R-squared:  0.1717
## F-statistic: 27.73 on 1 and 128 DF,  p-value: 5.711e-07
```

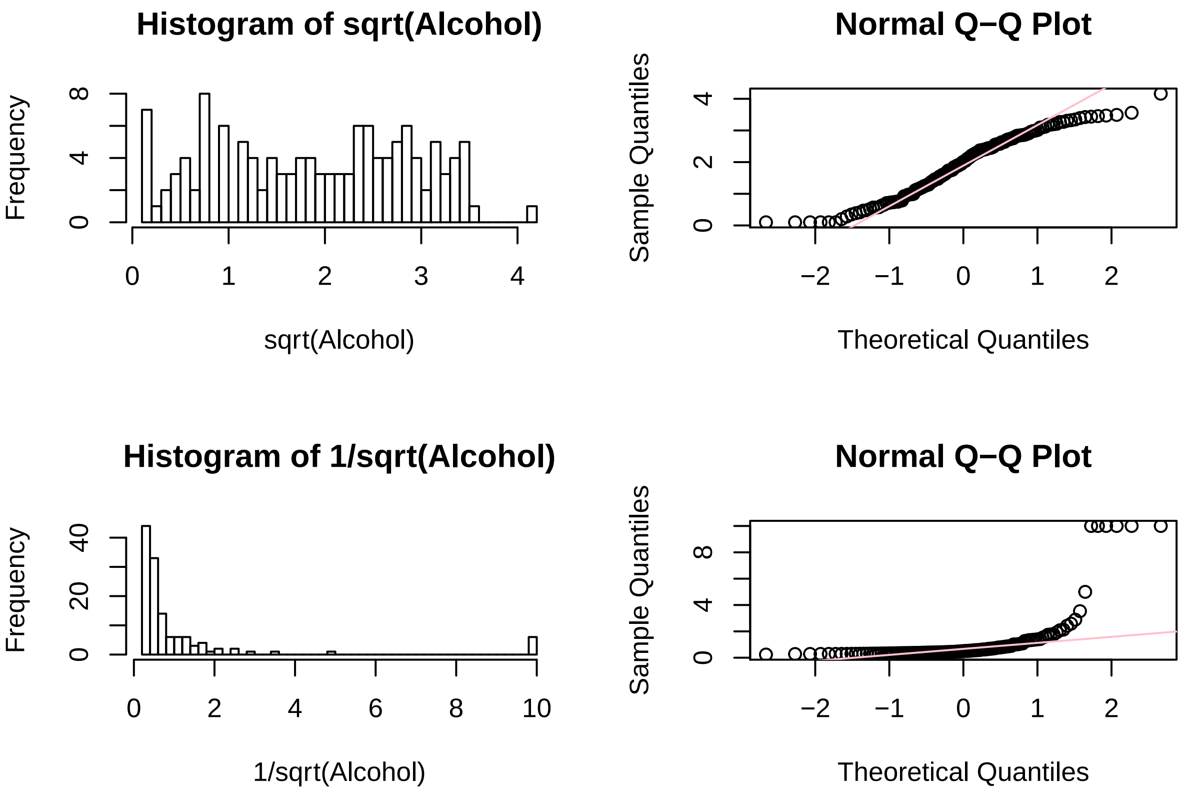
```
# significant

# powertransform says sqrt
par(mfrow = c(2,2))
```



```
with(lifeData.2011,hist(Alcohol,50))
with(lifeData.2011,qqnorm(Alcohol))
with(lifeData.2011,qqline(Alcohol,col="pink"))
with(lifeData.2011,hist(log(Alcohol),50))
with(lifeData.2011,qqnorm(log(Alcohol)))
with(lifeData.2011,qqline(log(Alcohol),col="pink"))
```



```r
with(lifeData.2011,hist(sqrt(Alcohol),50))
with(lifeData.2011,qqnorm(sqrt(Alcohol)))
with(lifeData.2011,qqline(sqrt(Alcohol),col="pink"))
with(lifeData.2011,hist(1/sqrt(Alcohol),50))
with(lifeData.2011,qqnorm(1/sqrt(Alcohol)))
with(lifeData.2011,qqline(1/sqrt(Alcohol),col="pink"))
```

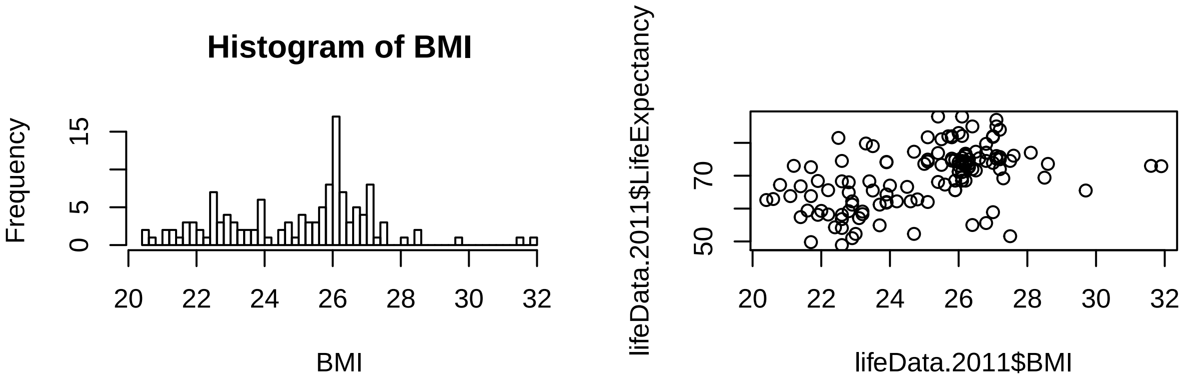


```r
# sqrt

# use sqrt
lifeData.df <- cbind(lifeData.df,data.frame(alcohol.sqrt=sqrt(lifeData.2011[,'Alcohol'])))


############################################
# BMI

#  no zeroes

with(lifeData.2011,hist(BMI,50))
plot(lifeData.2011$BMI,lifeData.2011$LifeExpectancy)
# ok

summary(lm(LifeExpectancy~BMI,data=lifeData.2011))
```

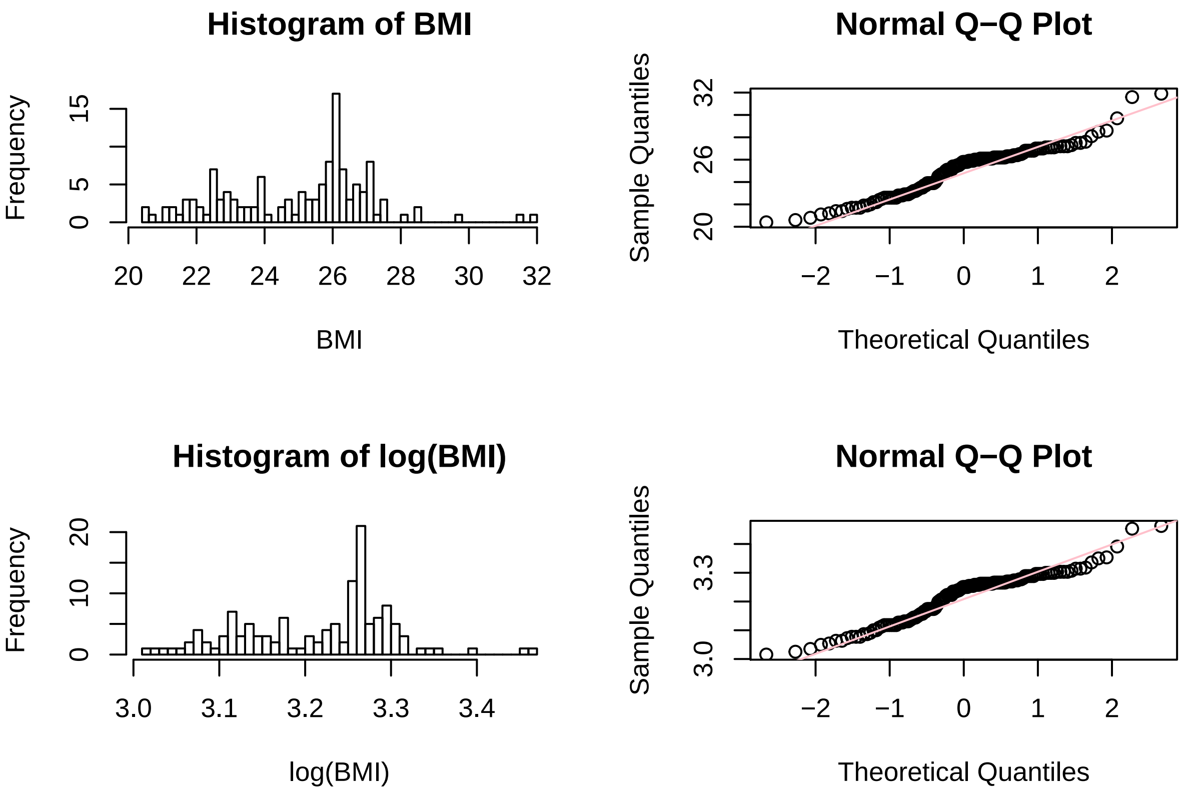
```
##
## Call:
## lm(formula = LifeExpectancy ~ BMI, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.243  -5.324   1.126   4.357  17.428
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.9107     8.0036   2.363   0.0196 *
## BMI           2.0339     0.3183   6.390 2.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.85 on 128 degrees of freedom
## Multiple R-squared:  0.2418, Adjusted R-squared:  0.2359
## F-statistic: 40.83 on 1 and 128 DF,  p-value: 2.821e-09
```

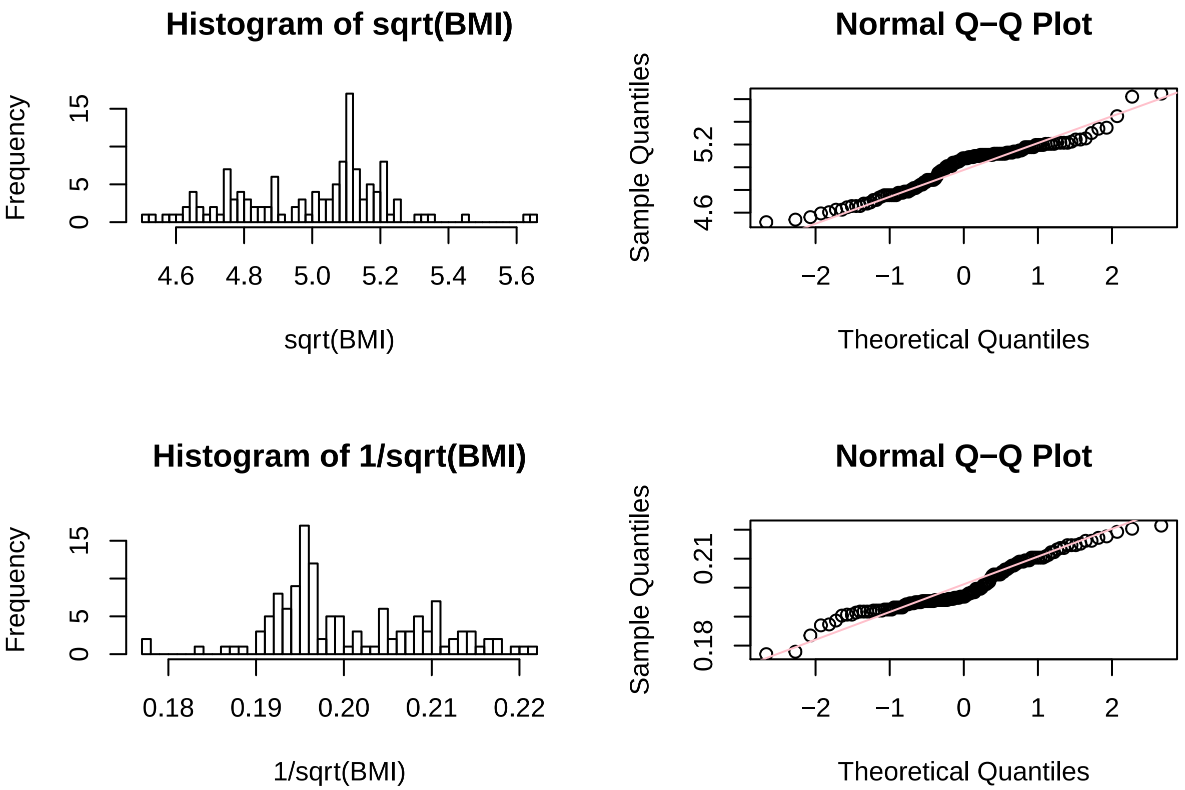
```
# significant

# powertransform says no transform
par(mfrow = c(2,2))
```



```
with(lifeData.2011,hist(BMI,50))
with(lifeData.2011,qqnorm(BMI))
with(lifeData.2011,qqline(BMI,col="pink"))
with(lifeData.2011,hist(log(BMI),50))
with(lifeData.2011,qqnorm(log(BMI)))
with(lifeData.2011,qqline(log(BMI),col="pink"))
```



```
with(lifeData.2011,hist(sqrt(BMI),50))
with(lifeData.2011,qqnorm(sqrt(BMI)))
with(lifeData.2011,qqline(sqrt(BMI),col="pink"))
with(lifeData.2011,hist(1/sqrt(BMI),50))
with(lifeData.2011,qqnorm(1/sqrt(BMI)))
with(lifeData.2011,qqline(1/sqrt(BMI),col="pink"))
```

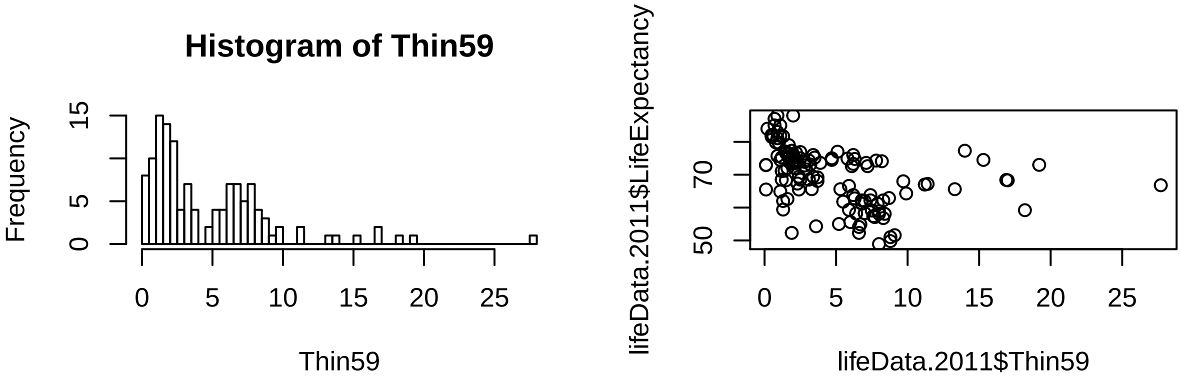


```
# not looking too good but no transform

# no transform
lifeData.df <- cbind(lifeData.df,data.frame(bmi=lifeData.2011[,'BMI']))


############################################
# Thin59

#  no zeroes

with(lifeData.2011,hist(Thin59,50))
plot(lifeData.2011$Thin59,lifeData.2011$LifeExpectancy)
# skewed

summary(lm(LifeExpectancy~Thin59,data=lifeData.2011))
```

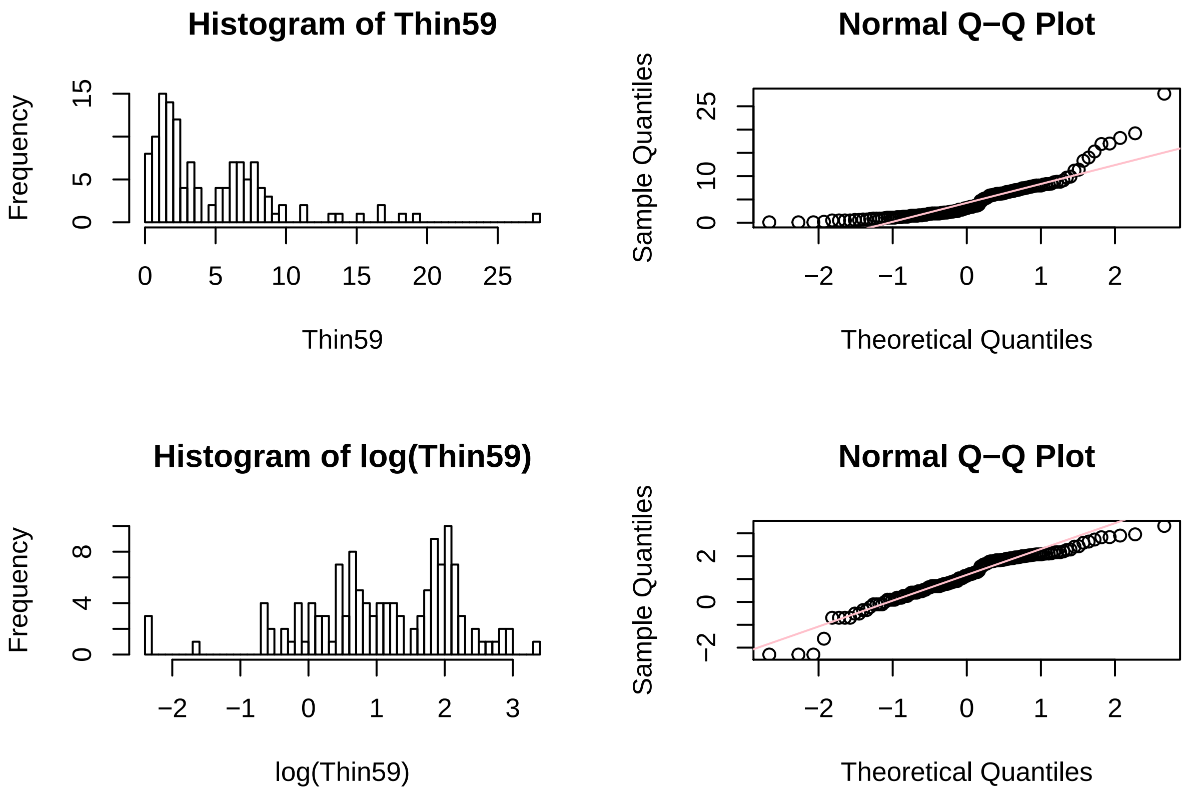
```
##
## Call:
## lm(formula = LifeExpectancy ~ Thin59, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.042  -5.704   1.773   5.072  16.641
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.9756     1.0438  70.869  < 2e-16 ***
## Thin59       -0.8598     0.1593  -5.398 3.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.136 on 128 degrees of freedom
## Multiple R-squared:  0.1855, Adjusted R-squared:  0.1791
## F-statistic: 29.14 on 1 and 128 DF,  p-value: 3.15e-07
```

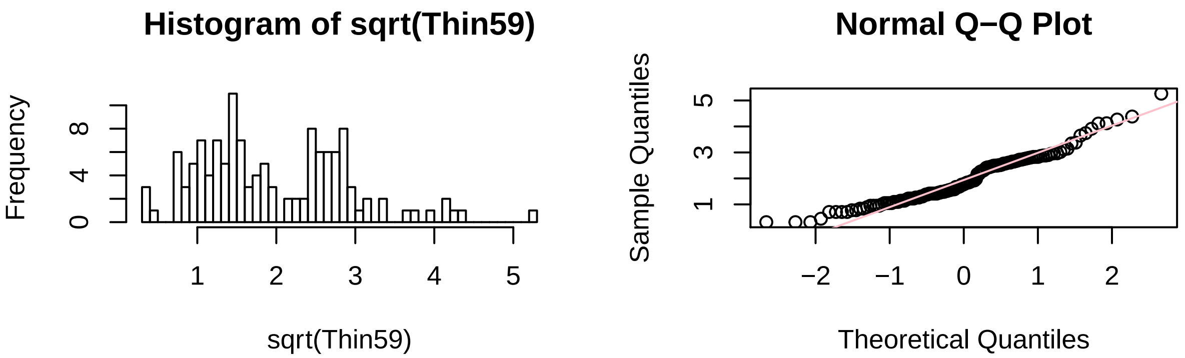
```
# significant

# powertransform says 0.14
par(mfrow = c(2,2))
```

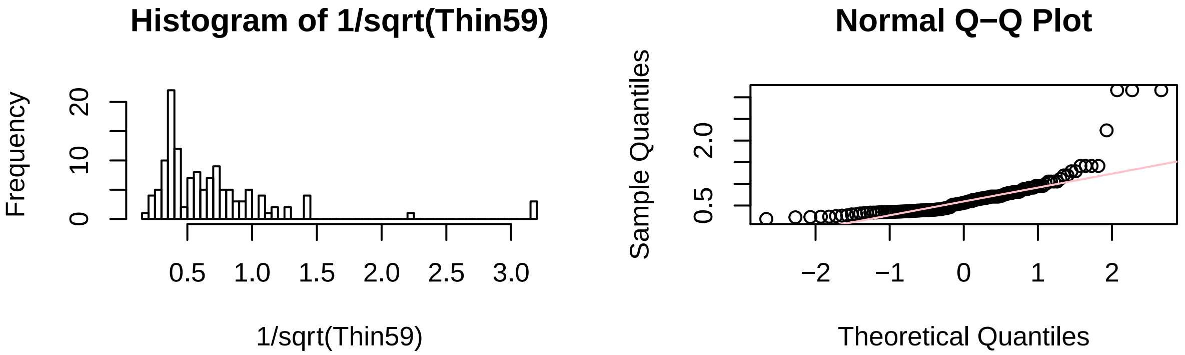


```
with(lifeData.2011,hist(Thin59,50))
with(lifeData.2011,qqnorm(Thin59))
with(lifeData.2011,qqline(Thin59,col="pink"))
with(lifeData.2011,hist(log(Thin59),50))
with(lifeData.2011,qqnorm(log(Thin59)))
with(lifeData.2011,qqline(log(Thin59),col="pink"))
```

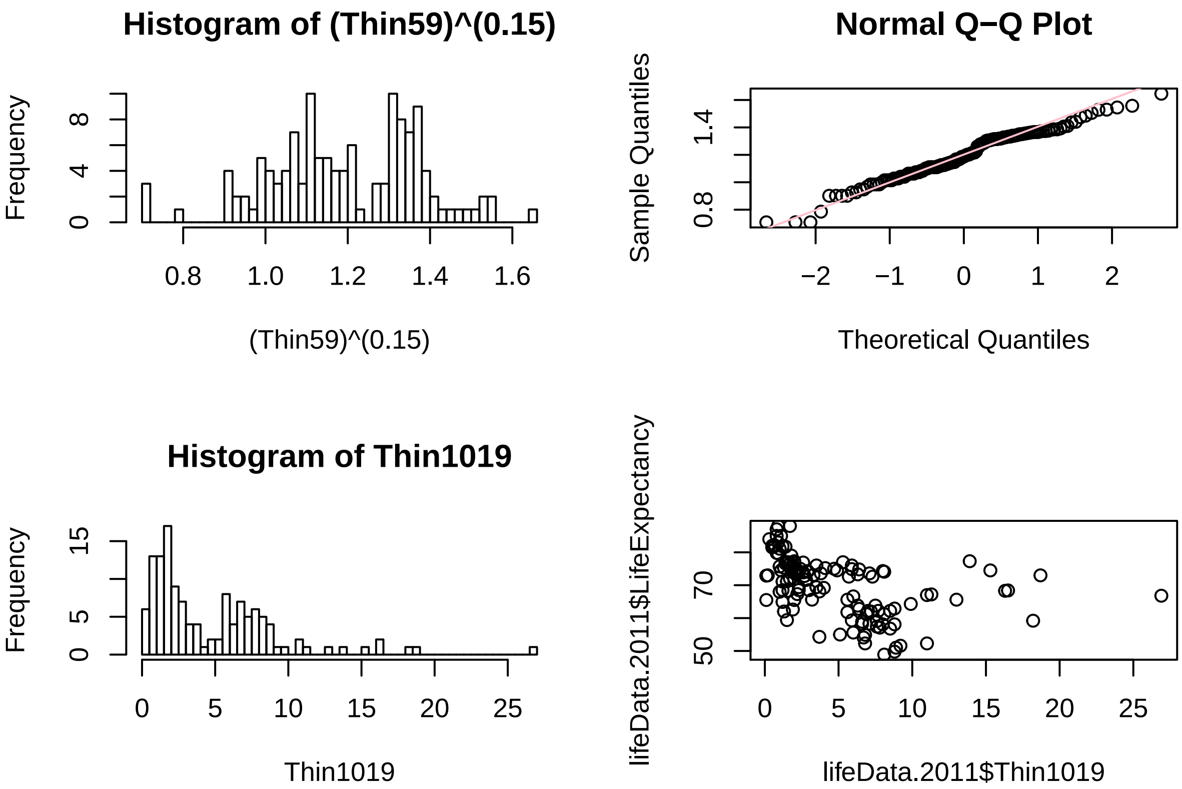


```
with(lifeData.2011,hist(sqrt(Thin59),50))
with(lifeData.2011,qqnorm(sqrt(Thin59)))
with(lifeData.2011,qqline(sqrt(Thin59),col="pink"))
with(lifeData.2011,hist(1/sqrt(Thin59),50))
with(lifeData.2011,qqnorm(1/sqrt(Thin59)))
with(lifeData.2011,qqline(1/sqrt(Thin59),col="pink"))
```





```
# not looking ok

with(lifeData.2011,hist((Thin59)^(0.15),50))
with(lifeData.2011,qqnorm((Thin59)^(0.15)))
with(lifeData.2011,qqline((Thin59)^(0.15),col="pink"))
# ok, use ^(0.15)

lifeData.df <- cbind(lifeData.df,data.frame(thin59.power0.15=(lifeData.2011[,'Thin59'])^(0.15)

############################################
# Thin1019

#  no zeroes

with(lifeData.2011,hist(Thin1019,50))
plot(lifeData.2011$Thin1019,lifeData.2011$LifeExpectancy)
```

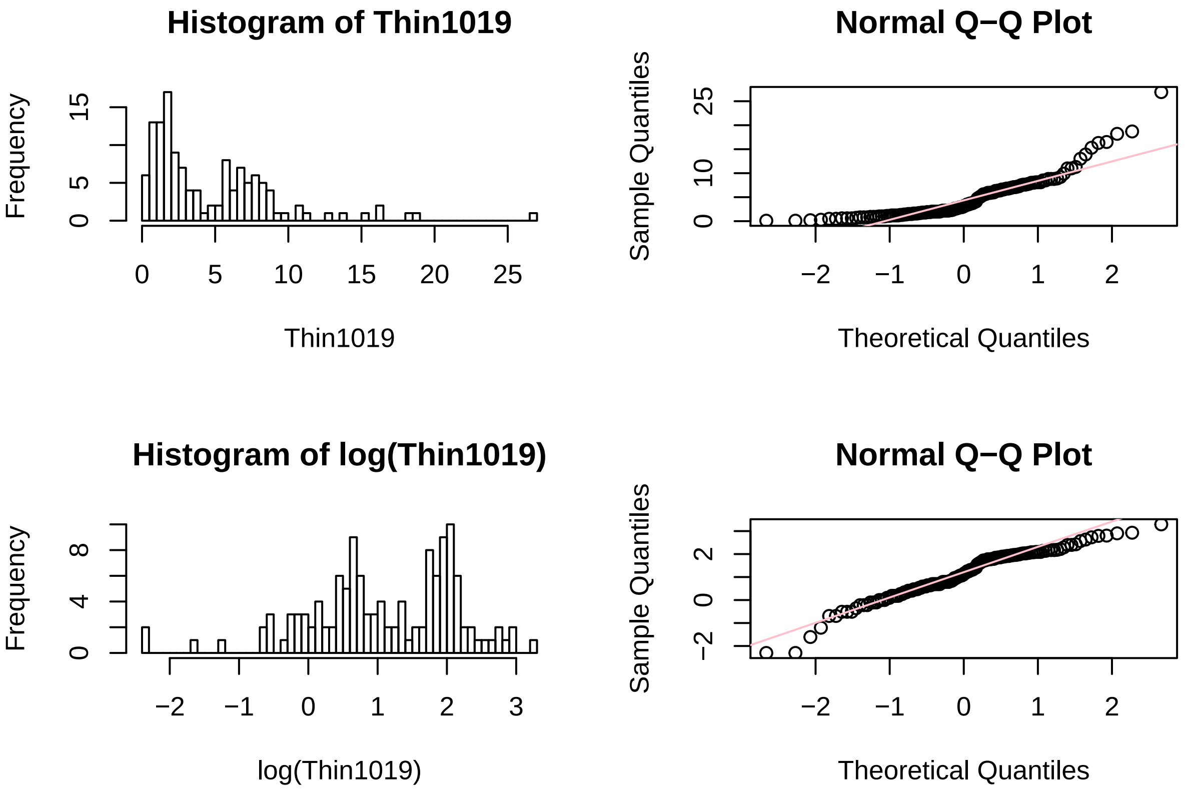


```
# skewed

summary(lm(LifeExpectancy~Thin1019,data=lifeData.2011))
```

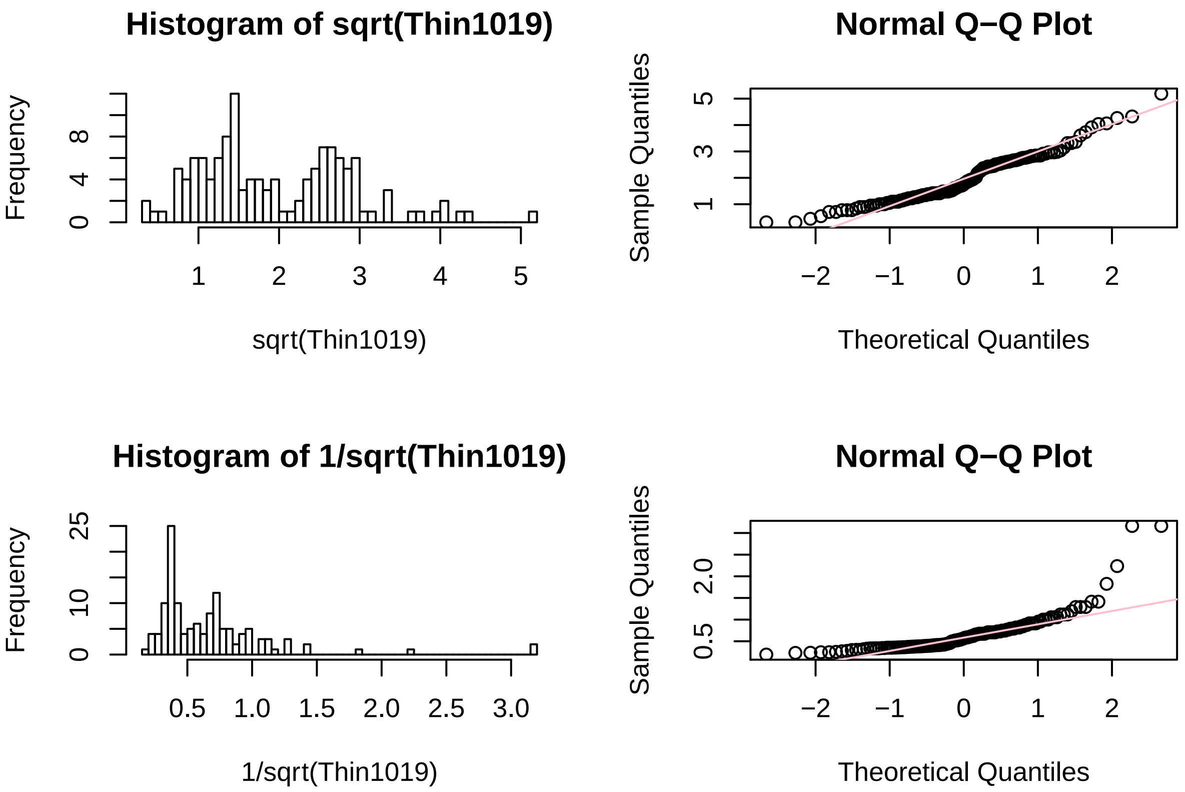
```
##
## Call:
## lm(formula = LifeExpectancy ~ Thin1019, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.848  -5.548   1.501   5.087  17.760
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  74.3769     1.0309  72.148  < 2e-16 ***
## Thin1019     -0.9419     0.1580  -5.961 2.28e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.976 on 128 degrees of freedom
## Multiple R-squared:  0.2173, Adjusted R-squared:  0.2112
## F-statistic: 35.53 on 1 and 128 DF,  p-value: 2.281e-08
```

```
# significant

# powertransform says ^0.13
```

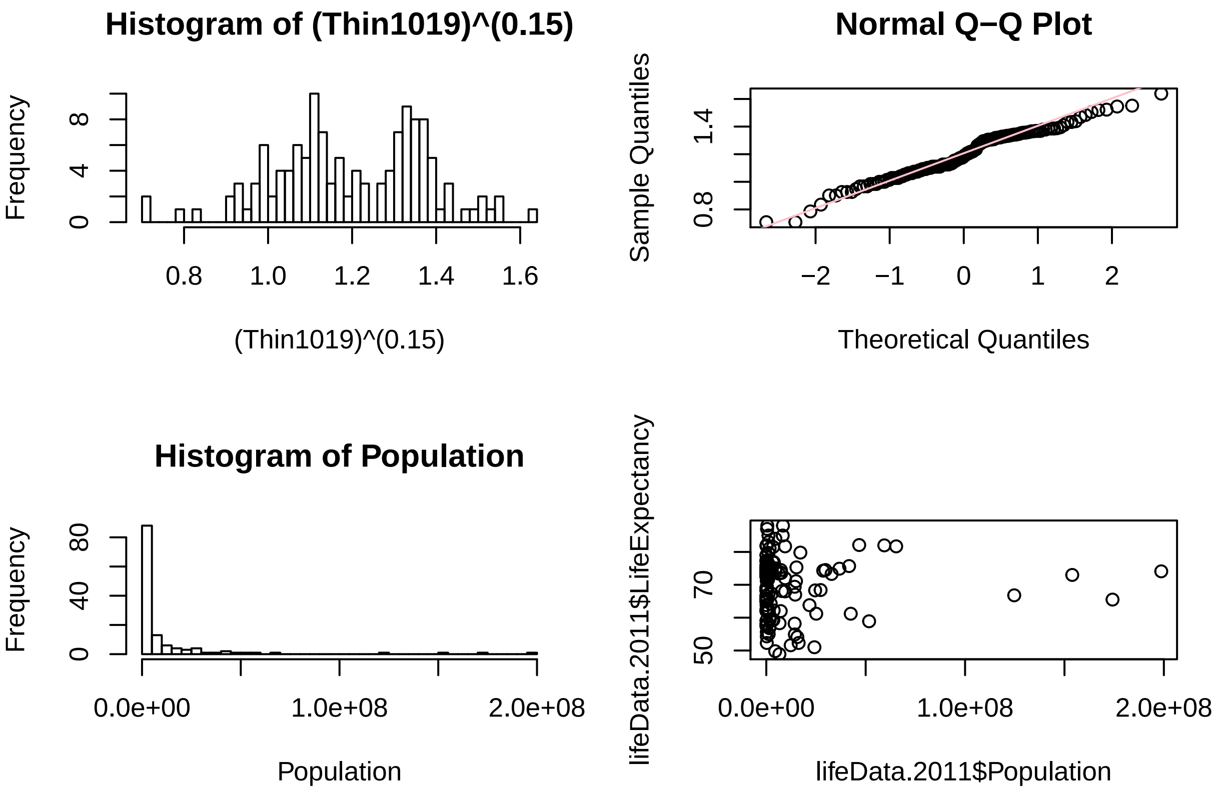
```r
par(mfrow = c(2,2))
with(lifeData.2011,hist(Thin1019,50))
with(lifeData.2011,qqnorm(Thin1019))
with(lifeData.2011,qqline(Thin1019,col="pink"))
with(lifeData.2011,hist(log(Thin1019),50))
with(lifeData.2011,qqnorm(log(Thin1019)))
with(lifeData.2011,qqline(log(Thin1019),col="pink"))
```



```r
with(lifeData.2011,hist(sqrt(Thin1019),50))
with(lifeData.2011,qqnorm(sqrt(Thin1019)))
with(lifeData.2011,qqline(sqrt(Thin1019),col="pink"))
with(lifeData.2011,hist(1/sqrt(Thin1019),50))
with(lifeData.2011,qqnorm(1/sqrt(Thin1019)))
with(lifeData.2011,qqline(1/sqrt(Thin1019),col="pink"))
```



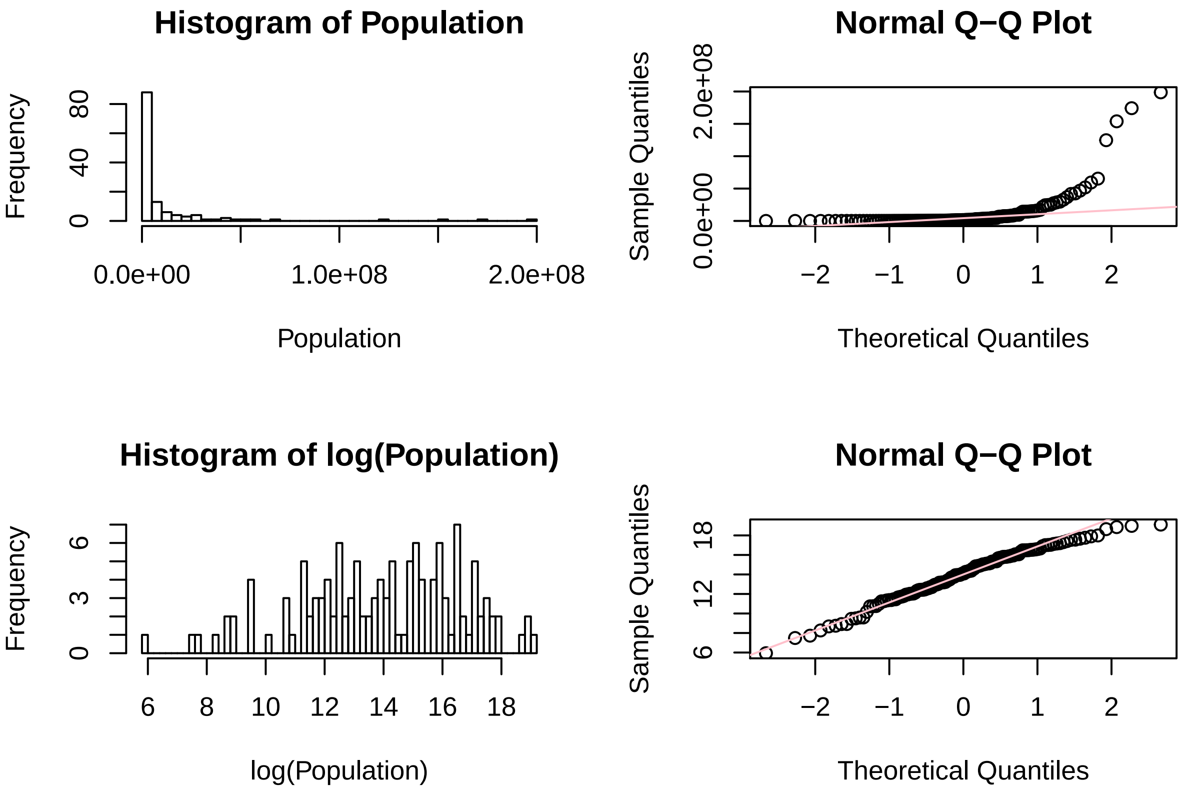
```
with(lifeData.2011,hist((Thin1019)^(0.15),50))
with(lifeData.2011,qqnorm((Thin1019)^(0.15)))
with(lifeData.2011,qqline((Thin1019)^(0.15),col="pink"))

# use ^0.15
lifeData.df <- cbind(lifeData.df,data.frame(thin1019.power0.15=(lifeData.2011[,'Thin1019'])^(0


############################################
# Population

#  no zeroes

with(lifeData.2011,hist(Population,50))
plot(lifeData.2011$Population,lifeData.2011$LifeExpectancy)
```



```
# skewed

summary(lm(LifeExpectancy~Population,data=lifeData.2011))
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ Population, data = lifeData.2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.935  -6.976   2.591   5.555  18.198
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.980e+01  8.507e-01  82.054   <2e-16 ***
## Population  5.315e-09  2.633e-08   0.202     0.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.014 on 128 degrees of freedom
## Multiple R-squared:  0.0003183,  Adjusted R-squared:  -0.007492
## F-statistic: 0.04075 on 1 and 128 DF,  p-value: 0.8403
```

```
# powertransform says log
par(mfrow = c(2,2))
with(lifeData.2011,hist(Population,50))
with(lifeData.2011,qqnorm(Population))
with(lifeData.2011,qqline(Population,col="pink"))
with(lifeData.2011,hist(log(Population),50))
with(lifeData.2011,qqnorm(log(Population)))
with(lifeData.2011,qqline(log(Population),col="pink"))
```

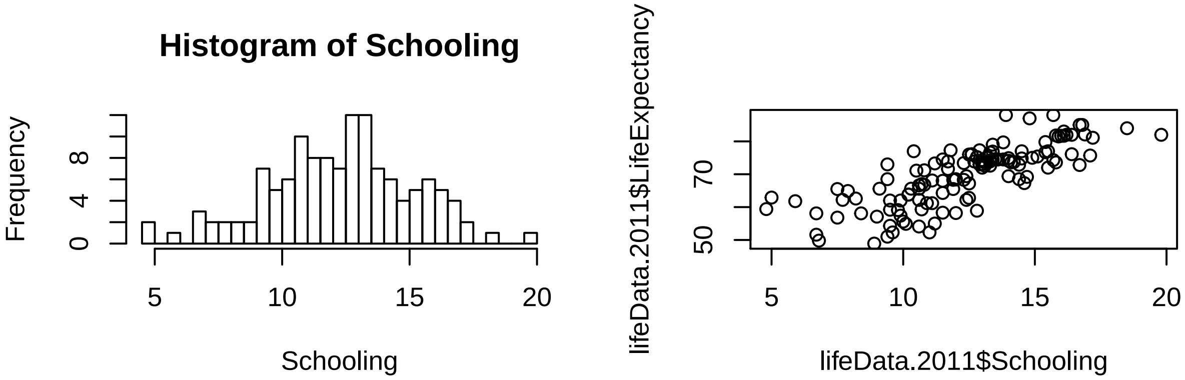


```
# log it
lifeData.df <- cbind(lifeData.df,data.frame(population.log=log(lifeData.2011[,'Population'])))

###########################################
# Schooling

#  no zeroes

with(lifeData.2011,hist(Schooling,50))
plot(lifeData.2011$Schooling,lifeData.2011$LifeExpectancy)
# skewed

summary(lm(LifeExpectancy~Schooling,data=lifeData.2011))
```

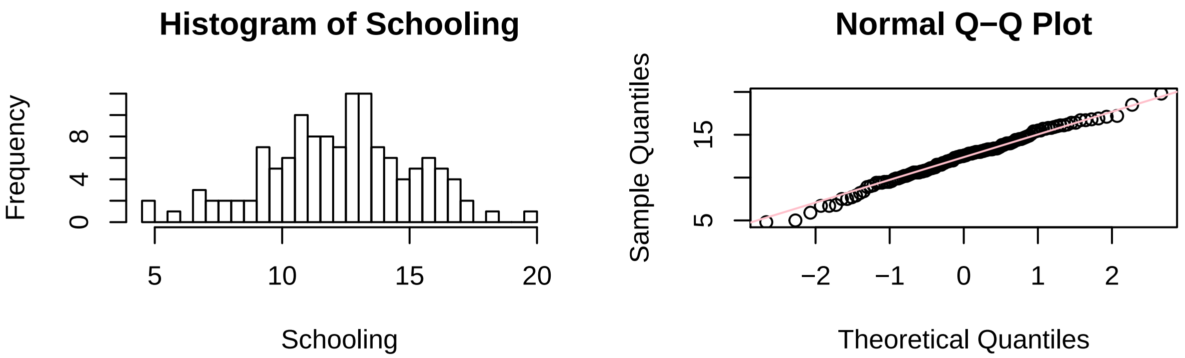
```
##
## Call:
## lm(formula = LifeExpectancy ~ Schooling, data = lifeData.2011)
##
## Residuals:
```

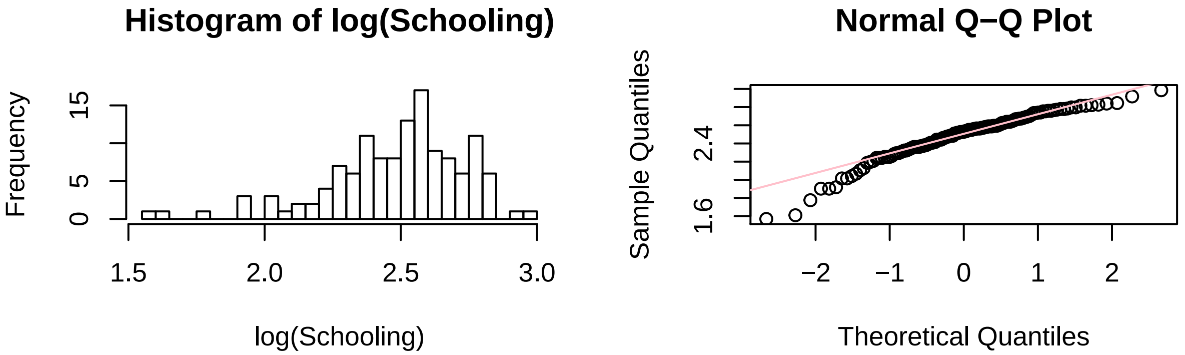
```
##     Min      1Q  Median      3Q     Max 
## -14.237  -3.690   0.768   3.620  14.339 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  39.5161     2.2412   17.63   <2e-16 ***
## Schooling     2.4565     0.1769   13.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.694 on 128 degrees of freedom
## Multiple R-squared:  0.6011, Adjusted R-squared:  0.598 
## F-statistic: 192.9 on 1 and 128 DF,  p-value: < 2.2e-16
```

```
# significant

# powertransform says ^1.5
par(mfrow = c(2,2))
```

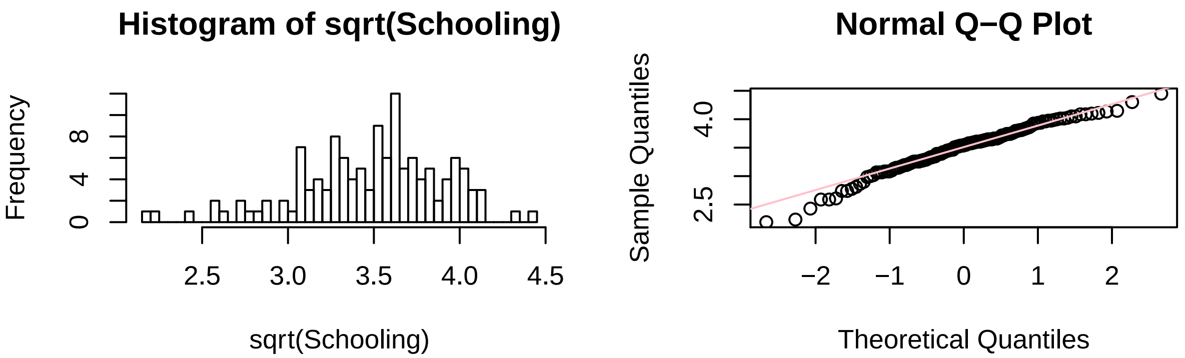


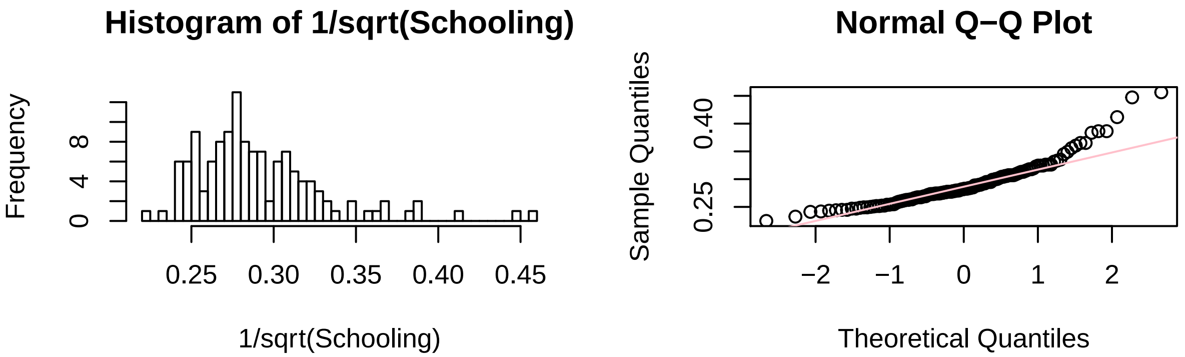
```
with(lifeData.2011,hist(Schooling,50))
with(lifeData.2011,qqnorm(Schooling))
with(lifeData.2011,qqline(Schooling,col="pink"))
with(lifeData.2011,hist(log(Schooling),50))
with(lifeData.2011,qqnorm(log(Schooling)))
with(lifeData.2011,qqline(log(Schooling),col="pink"))
```



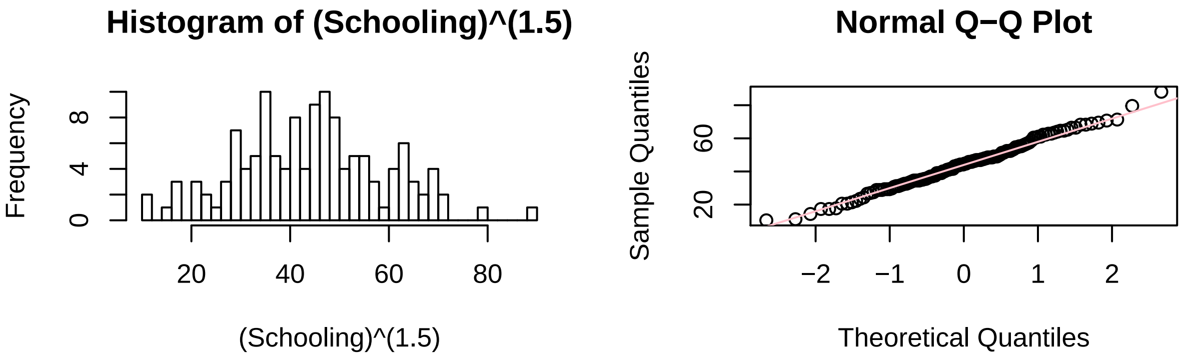


```
with(lifeData.2011,hist(sqrt(Schooling),50))
with(lifeData.2011,qqnorm(sqrt(Schooling)))
with(lifeData.2011,qqline(sqrt(Schooling),col="pink"))
with(lifeData.2011,hist(1/sqrt(Schooling),50))
with(lifeData.2011,qqnorm(1/sqrt(Schooling)))
with(lifeData.2011,qqline(1/sqrt(Schooling),col="pink"))
```





```
with(lifeData.2011,hist((Schooling)^(1.5),50))
with(lifeData.2011,qqnorm((Schooling)^(1.5)))
with(lifeData.2011,qqline((Schooling)^(1.5),col="pink"))
# prefer not to transform -- looks ok

# no transform
lifeData.df <- cbind(lifeData.df,data.frame(schooling=(lifeData.2011[,'Schooling'])))
```



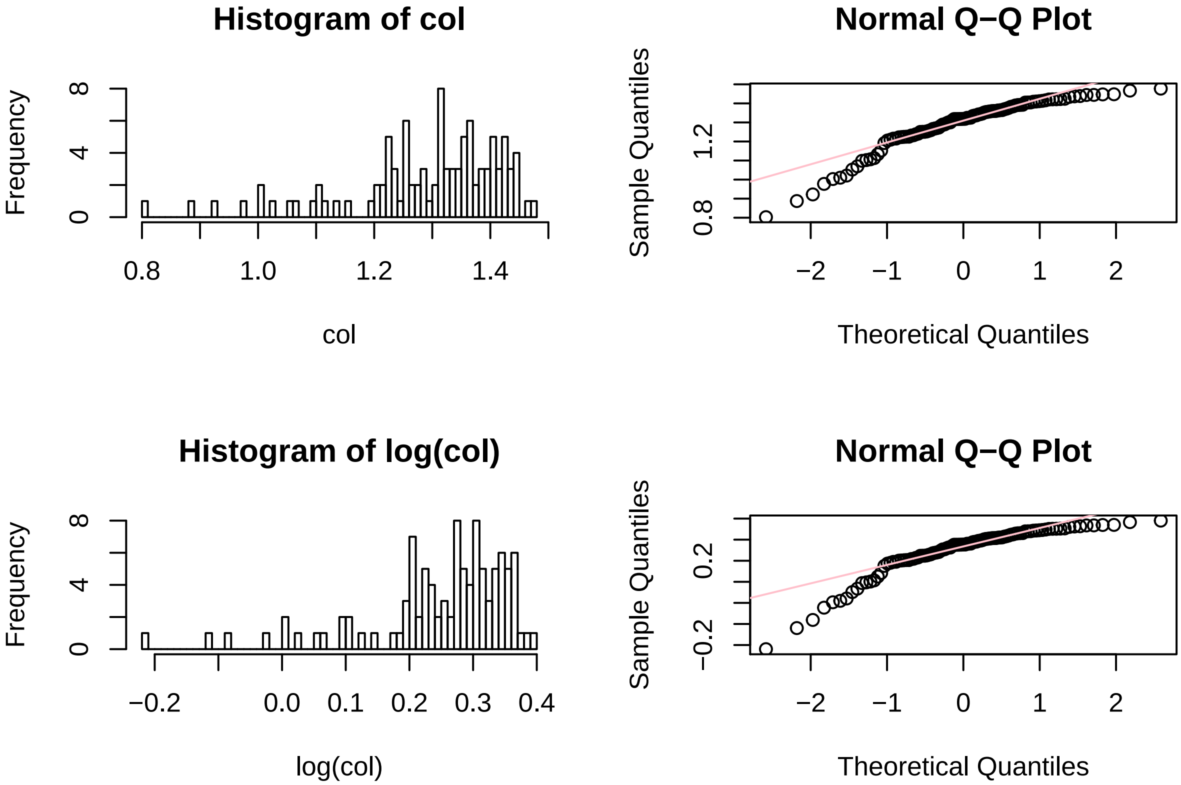
3.5 Appendix E: Transformations for WHR Data

```
## Transforming WHR.
# Social.support
# Healthy.life.expectancy.at.birth
# Freedom.to.make.life.choices
```

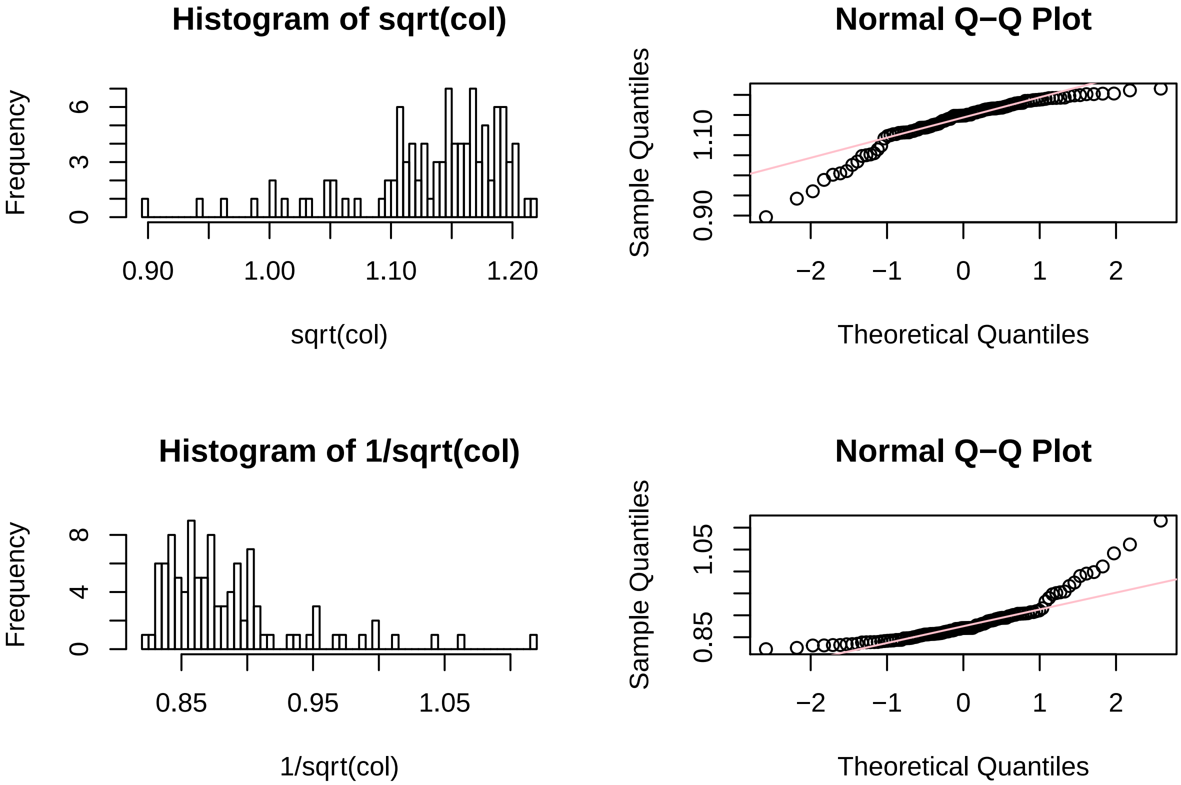
```r
# Generosity
# Perceptions.of.corruption
# Positive.affect
# Negative.affect
# Democratic.Quality
# Delivery.Quality

col = whr.and.who.2011.clean$Social.support+.5
summary(powerTransform(col))
```

```
## bcPower Transformation to Normality
##     Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## col    5.4886        5.49       3.6542       7.3231
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 46.44428  1 9.4261e-12
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 29.34644  1 6.0528e-08
```

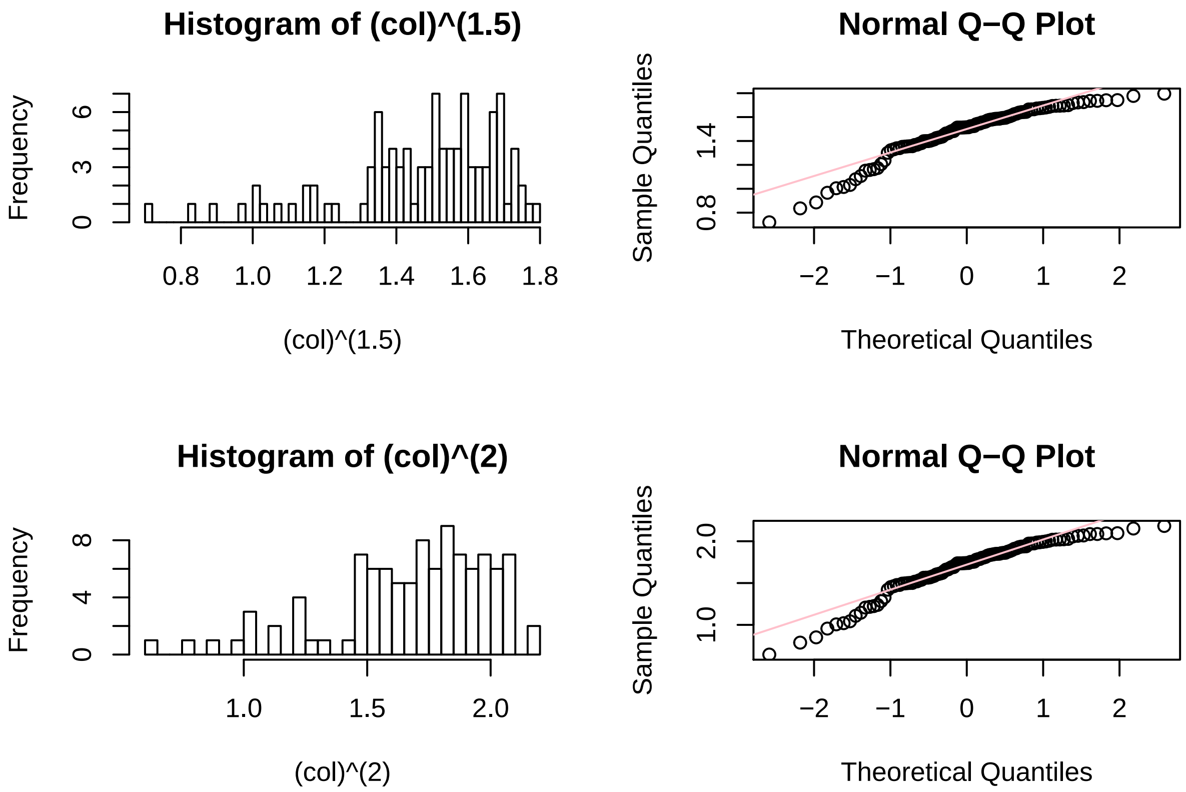
```r
par(mfrow = c(2,2))
with(whr.and.who.2011.clean,hist(col,50))
with(whr.and.who.2011.clean,qqnorm(col))
with(whr.and.who.2011.clean,qqline(col,col="pink"))
with(whr.and.who.2011.clean,hist(log(col),50))
with(whr.and.who.2011.clean,qqnorm(log(col)))
with(whr.and.who.2011.clean,qqline(log(col),col="pink"))
```



```
with(whr.and.who.2011.clean,hist(sqrt(col),50))
with(whr.and.who.2011.clean,qqnorm(sqrt(col)))
with(whr.and.who.2011.clean,qqline(sqrt(col),col="pink"))
with(whr.and.who.2011.clean,hist(1/sqrt(col),50))
with(whr.and.who.2011.clean,qqnorm(1/sqrt(col)))
with(whr.and.who.2011.clean,qqline(1/sqrt(col),col="pink"))
```



```
with(whr.and.who.2011.clean,hist((col)^(1.5),50))
with(whr.and.who.2011.clean,qqnorm((col)^(1.5)))
with(whr.and.who.2011.clean,qqline((col)^(1.5),col="pink"))
with(whr.and.who.2011.clean,hist((col)^(2),50))
with(whr.and.who.2011.clean,qqnorm((col)^(2)))
with(whr.and.who.2011.clean,qqline((col)^(2),col="pink"))
```



```r
#boxplot(whr.and.who.2011.clean$Measles)
```

3.6 Appendix F: Eliminating Variables with VIFs for WHO Data

```r
# Add Life Expectancy to clean and transformed dataframe
lifeData.df = cbind(data.frame(life.expectancy=(lifeData.2011[,'LifeExpectancy'])),lifeData.df

# Linear regression with all transformed variables
fit.2011 = lm(life.expectancy~.,data=lifeData.df)
summary(fit.2011) # Nothing significant
```

```
##
## Call:
## lm(formula = life.expectancy ~ ., data = lifeData.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5736 -1.6548 -0.2424  1.8117  6.6538
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                6.433e+01  6.285e+00  10.236  < 2e-16 ***
## developedTRUE             -2.609e-01  1.057e+00  -0.247 0.805466
## adult.mortality.sqrt      -3.277e-01  7.620e-02  -4.301 3.68e-05 ***
## infant.deaths.log          3.569e-01  1.184e+00   0.301 0.763702
## under.five.deaths.log     -7.284e-01  1.190e+00  -0.612 0.541577
## hivaids.1oversqrt          1.605e+00  4.502e-01   3.565 0.000537 ***
## hepb.power4                9.593e-09  1.888e-08   0.508 0.612437
## polio.power4              -2.214e-08  2.328e-08  -0.951 0.343474
## diphtheria.power5          7.798e-11  2.799e-10   0.279 0.781093
## percent.expenditure.sqrt   2.076e+00  8.092e-01   2.566 0.011621 *
## total.expenditure          5.154e-02  1.382e-01   0.373 0.709931
## income.composition.power2  5.902e+01  6.746e+00   8.749 2.67e-14 ***
## gdp.log                   -3.136e-02  2.288e-01  -0.137 0.891243
## alcohol.sqrt              -6.569e-01  4.155e-01  -1.581 0.116708
## bmi                       -6.760e-01  1.876e-01  -3.603 0.000473 ***
## thin59.power0.15          -5.141e+00  5.787e+00  -0.888 0.376322
## thin1019.power0.15         2.153e+00  5.936e+00   0.363 0.717520
## population.log            -9.344e-02  1.278e-01  -0.731 0.466143
## schooling                 -1.110e+00  2.626e-01  -4.228 4.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.035 on 111 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8858
## F-statistic: 56.57 on 18 and 111 DF,  p-value: < 2.2e-16
```

```
# Examine VIFs
vif(fit.2011)
```

```
##                 developed      adult.mortality.sqrt
##                  1.966676                  1.824591
##         infant.deaths.log     under.five.deaths.log
##                 70.621644                 77.091014
##         hivaids.1oversqrt               hepb.power4
##                  3.525550                  4.433998
##              polio.power4         diphtheria.power5
##                  6.773107                  9.042687
##  percent.expenditure.sqrt         total.expenditure
##                  2.209281                  1.630932
## income.composition.power2                   gdp.log
##                 15.416890                  2.146180
##              alcohol.sqrt                       bmi
##                  2.516345                  2.324246
##          thin59.power0.15        thin1019.power0.15
##                 15.964489                 16.194829
##            population.log                 schooling
##                  1.672515                  7.758724
```

```
# Big multicollinearity issues, let's eliminate under.five.deaths
lifeDataFit.df = lifeData.df[,-5]
fit.2011 = lm(life.expectancy~.,data=lifeDataFit.df)
summary(fit.2011)
```

```
##
## Call:
## lm(formula = life.expectancy ~ ., data = lifeDataFit.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.5909 -1.7046 -0.1877  1.9220  6.5733 
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                6.376e+01  6.198e+00  10.288  < 2e-16 ***
## developedTRUE             -2.364e-01  1.053e+00  -0.224 0.822792    
## adult.mortality.sqrt      -3.290e-01  7.596e-02  -4.331 3.25e-05 ***
## infant.deaths.log         -3.546e-01  2.279e-01  -1.556 0.122437    
## hivaids.1oversqrt          1.652e+00  4.425e-01   3.733 0.000299 ***
## hepb.power4                8.609e-09  1.876e-08   0.459 0.647205    
## polio.power4              -2.155e-08  2.319e-08  -0.929 0.354761    
## diphtheria.power5          7.116e-11  2.789e-10   0.255 0.799092    
## percent.expenditure.sqrt   2.070e+00  8.068e-01   2.565 0.011632 *  
## total.expenditure          6.511e-02  1.360e-01   0.479 0.633160    
## income.composition.power2  5.943e+01  6.693e+00   8.879 1.26e-14 ***
## gdp.log                   -3.798e-02  2.280e-01  -0.167 0.867970    
## alcohol.sqrt              -6.756e-01  4.132e-01  -1.635 0.104892    
## bmi                       -6.639e-01  1.861e-01  -3.568 0.000531 ***
## thin59.power0.15          -5.145e+00  5.771e+00  -0.891 0.374598    
## thin1019.power0.15         2.146e+00  5.919e+00   0.362 0.717671    
## population.log            -1.043e-01  1.262e-01  -0.827 0.410039    
## schooling                 -1.113e+00  2.619e-01  -4.249 4.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.027 on 112 degrees of freedom
## Multiple R-squared:  0.9014, Adjusted R-squared:  0.8864 
## F-statistic: 60.21 on 17 and 112 DF,  p-value: < 2.2e-16
```

```r
vif(fit.2011)
```

```
##                 developed      adult.mortality.sqrt 
##                  1.963859                  1.823304 
##         infant.deaths.log         hivaids.1oversqrt 
##                  2.629027                  3.424512 
##               hepb.power4              polio.power4 
##                  4.401909                  6.761315 
##         diphtheria.power5  percent.expenditure.sqrt 
##                  9.028380                  2.208890 
##         total.expenditure income.composition.power2 
##                  1.589003                 15.262448 
##                   gdp.log              alcohol.sqrt 
##                   2.141391                  2.502861
##                        bmi          thin59.power0.15
##                   2.298443                 15.964468
##         thin1019.power0.15            population.log
##                  16.194764                  1.640069
##                  schooling
##                   7.757274
```

```r
# Still multicollinearity issues, let's eliminate thin1019
lifeDataFit.df = lifeDataFit.df[,-17]
fit.2011 = lm(life.expectancy~.,data=lifeDataFit.df)
summary(fit.2011)
```

```
##
## Call:
## lm(formula = life.expectancy ~ ., data = lifeDataFit.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6063 -1.7397 -0.2132  1.8954  6.6918
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.286e+01  6.093e+00  10.317  < 2e-16 ***
## developedTRUE              -3.929e-01  1.034e+00  -0.380 0.704813
## adult.mortality.sqrt       -3.385e-01  7.499e-02  -4.514 1.57e-05 ***
## infant.deaths.log          -4.620e-01  1.870e-01  -2.470 0.014998 *
## hivaids.1oversqrt           1.669e+00  4.414e-01   3.782 0.000250 ***
## hepb.power4                 8.177e-09  1.873e-08   0.437 0.663223
## polio.power4               -2.470e-08  2.284e-08  -1.081 0.281899
## diphtheria.power5           1.077e-10  2.750e-10   0.392 0.696075
## percent.expenditure.sqrt    2.047e+00  8.052e-01   2.542 0.012386 *
## total.expenditure           5.544e-02  1.353e-01   0.410 0.682855
## income.composition.power2   5.844e+01  6.575e+00   8.888 1.14e-14 ***
## gdp.log                    -4.691e-02  2.274e-01  -0.206 0.836910
## alcohol.sqrt               -6.087e-01  4.047e-01  -1.504 0.135332
## bmi                        -6.569e-01  1.856e-01  -3.539 0.000584 ***
## thin59.power0.15           -5.156e+00  5.763e+00  -0.895 0.372844
## thin1019.power0.15          2.324e+00  5.907e+00   0.393 0.694767
## schooling                  -1.109e+00  2.615e-01  -4.241 4.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.022 on 113 degrees of freedom
## Multiple R-squared:  0.9008, Adjusted R-squared:  0.8867
## F-statistic: 64.11 on 16 and 113 DF,  p-value: < 2.2e-16
```

```r
vif(fit.2011)
```

```
##                 developed      adult.mortality.sqrt 
##                  1.900436                  1.781699 
##         infant.deaths.log         hivaids.1oversqrt 
##                  1.775973                  3.416761 
##               hepb.power4              polio.power4 
##                  4.398489                  6.578948 
##         diphtheria.power5  percent.expenditure.sqrt 
##                  8.801779                  2.206243 
##         total.expenditure income.composition.power2 
##                  1.577271                 14.768932 
##                   gdp.log              alcohol.sqrt 
##                  2.136583                  2.407005 
##                       bmi          thin59.power0.15 
##                  2.293662                 15.964375 
##        thin1019.power0.15                 schooling 
##                 16.173308                  7.754989
```

```r
# Still multicollinearity issues, let's eliminate income.composition
lifeDataFit.df = lifeDataFit.df[,-12]
fit.2011 = lm(life.expectancy~.,data=lifeDataFit.df)
summary(fit.2011)
```

```
## 
## Call:
## lm(formula = life.expectancy ~ ., data = lifeDataFit.df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.6469 -1.7410 -0.1566  1.9078  6.7955 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                6.283e+01  6.065e+00  10.359  < 2e-16 ***
## developedTRUE             -4.216e-01  1.021e+00  -0.413 0.680345    
## adult.mortality.sqrt      -3.400e-01  7.429e-02  -4.577 1.21e-05 ***
## infant.deaths.log         -4.667e-01  1.848e-01  -2.525 0.012958 *  
## hivaids.1oversqrt          1.673e+00  4.392e-01   3.808 0.000227 ***
## hepb.power4                8.327e-09  1.863e-08   0.447 0.655845    
## polio.power4              -2.405e-08  2.253e-08  -1.067 0.288026    
## diphtheria.power5          9.839e-11  2.701e-10   0.364 0.716389    
## percent.expenditure.sqrt   2.064e+00  7.976e-01   2.587 0.010926 *  
## total.expenditure          5.346e-02  1.344e-01   0.398 0.691639    
## income.composition.power2  5.808e+01  6.317e+00   9.195 2.08e-15 ***
## alcohol.sqrt              -6.140e-01  4.022e-01  -1.527 0.129597    
## bmi                       -6.622e-01  1.831e-01  -3.617 0.000445 ***
## thin59.power0.15          -5.083e+00  5.728e+00  -0.887 0.376747    
## thin1019.power0.15         2.271e+00  5.876e+00   0.386 0.699864
## schooling                 -1.107e+00  2.602e-01  -4.255 4.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.01 on 114 degrees of freedom
## Multiple R-squared:  0.9007, Adjusted R-squared:  0.8877
## F-statistic: 68.96 on 15 and 114 DF,  p-value: < 2.2e-16
```

```
vif(fit.2011)
```

```
##                 developed      adult.mortality.sqrt
##                  1.865991                  1.763439
##         infant.deaths.log         hivaids.1oversqrt
##                  1.749677                  3.412184
##               hepb.power4              polio.power4
##                  4.391878                  6.454604
##         diphtheria.power5  percent.expenditure.sqrt
##                  8.564352                  2.182938
##         total.expenditure income.composition.power2
##                  1.569307                 13.745362
##              alcohol.sqrt                       bmi
##                  2.397313                  2.249894
##          thin59.power0.15        thin1019.power0.15
##                 15.903271                 16.143176
##                 schooling
##                  7.745590
```

3.7 Appendix G: Eliminating Variables with VIFs for WHO Predictors in Combined Data

```
# Fit using only WHO predictors
fit.whr.who.only.1 = lm(life.expectancy~., data = transformed.who)
summary(fit.whr.who.only.1)
```

```
##
## Call:
## lm(formula = life.expectancy ~ ., data = transformed.who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5434 -1.9928 -0.1763  1.7056  6.8503
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   5.910e+01  6.562e+00   9.006 6.34e-14 ***
## developedTRUE                -3.941e-01  1.108e+00  -0.356 0.722925
## adult.mortality.sqrt         -3.315e-01  8.104e-02  -4.091 9.90e-05 ***
## infant.deaths.shift1.log      1.111e+00  1.401e+00   0.793 0.429825
## under.five.deaths.shift1.log -1.001e+00  1.453e+00  -0.689 0.492993
## hivaids.1oversqrt             1.686e+00  5.100e-01   3.306 0.001401 **
## measles.shift1.log           -1.752e-01  1.235e-01  -1.419 0.159603
## hep.b.power4                  2.259e-08  1.891e-08   1.195 0.235631
## polio.power4                 -8.244e-09  1.945e-08  -0.424 0.672838
## diphtheria                   -2.356e-02  3.079e-02  -0.765 0.446383
## percent.expenditure.sqrt      2.537e+00  9.627e-01   2.635 0.010031 *
## total.expenditure            -1.167e-01  1.577e-01  -0.740 0.461535
## income.composition            6.218e+01  8.721e+00   7.129 3.42e-10 ***
## gdp.per.cap.log              -1.014e-01  2.599e-01  -0.390 0.697526
## alcohol.sqrt                 -4.710e-01  4.673e-01  -1.008 0.316429
## bmi                          -6.085e-01  2.532e-01  -2.404 0.018457 *
## thin.5.9.power0.15           -2.649e+00  1.338e+00  -1.981 0.050932 .
## thin.10.19.power0.15          1.823e+00  1.363e+00   1.337 0.184727
## population.log               -2.131e-02  1.435e-01  -0.148 0.882353
## schooling                    -1.237e+00  3.060e-01  -4.041 0.000118 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 83 degrees of freedom
## Multiple R-squared:  0.9224, Adjusted R-squared:  0.9046
## F-statistic: 51.92 on 19 and 83 DF,  p-value: < 2.2e-16
```

```
# Examine VIFs
vif(fit.whr.who.only.1)
```

```
##                    developed         adult.mortality.sqrt
##                     2.169121                     1.840068
##     infant.deaths.shift1.log under.five.deaths.shift1.log
##                    74.367381                    87.048129
##            hivaids.1oversqrt           measles.shift1.log
##                     3.835125                     1.805815
##                 hep.b.power4                 polio.power4
##                     3.352622                     3.443986
##                   diphtheria     percent.expenditure.sqrt
##                     2.785703                     2.322878
##            total.expenditure           income.composition
##                     1.761361                    23.325390
##              gdp.per.cap.log                 alcohol.sqrt
##                     2.366004                     2.708930
##                          bmi           thin.5.9.power0.15
##                     2.679649                    20.002106
##         thin.10.19.power0.15               population.log
##                    19.627662                     1.507352
##                    schooling
##                     9.353661
```

```
# Big multicollinearity issues, let's eliminate under.five.deaths
fit.whr.who.only.2 = lm(life.expectancy~.-under.five.deaths.shift1.log, data = transformed.who
summary(fit.whr.who.only.2)
```

```
##
## Call:
## lm(formula = life.expectancy ~ . - under.five.deaths.shift1.log,
##     data = transformed.who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5677 -1.9729 -0.1527  1.5666  6.8253
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.849e+01  6.480e+00   9.025 5.29e-14 ***
## developedTRUE            -2.718e-01  1.090e+00  -0.249 0.803722
## adult.mortality.sqrt     -3.342e-01  8.069e-02  -4.142 8.18e-05 ***
## infant.deaths.shift1.log  1.720e-01  3.175e-01   0.542 0.589446
## hivaids.1oversqrt         1.705e+00  5.076e-01   3.359 0.001178 **
## measles.shift1.log       -1.932e-01  1.203e-01  -1.606 0.112093
## hep.b.power4              2.171e-08  1.881e-08   1.154 0.251575
## polio.power4             -8.886e-09  1.937e-08  -0.459 0.647601
## diphtheria               -2.448e-02  3.067e-02  -0.798 0.426922
## percent.expenditure.sqrt  2.486e+00  9.569e-01   2.598 0.011062 *
## total.expenditure        -8.220e-02  1.491e-01  -0.551 0.582838
## income.composition        6.360e+01  8.447e+00   7.530 5.26e-11 ***
## gdp.per.cap.log          -1.112e-01  2.587e-01  -0.430 0.668329
## alcohol.sqrt             -5.103e-01  4.623e-01  -1.104 0.272856
## bmi                      -6.130e-01  2.523e-01  -2.430 0.017240 *
## thin.5.9.power0.15       -2.656e+00  1.333e+00  -1.992 0.049596 *
## thin.10.19.power0.15      1.872e+00  1.357e+00   1.380 0.171238
## population.log           -3.996e-02  1.405e-01  -0.284 0.776860
## schooling                -1.240e+00  3.050e-01  -4.066 0.000107 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.952 on 84 degrees of freedom
## Multiple R-squared:  0.922,  Adjusted R-squared:  0.9052
## F-statistic: 55.13 on 18 and 84 DF,  p-value: < 2.2e-16
```

```
vif(fit.whr.who.only.2)
```

```
##                developed     adult.mortality.sqrt infant.deaths.shift1.log
##                 2.113330                 1.835887                 3.846038
##        hivaids.1oversqrt       measles.shift1.log             hep.b.power4
##                 3.823506                 1.725276                 3.337409
##             polio.power4               diphtheria percent.expenditure.sqrt
##                 3.436065                2.780419                2.309297
##        total.expenditure      income.composition         gdp.per.cap.log
##                 1.583995               22.016338                2.358821
##             alcohol.sqrt                     bmi      thin.5.9.power0.15
##                 2.668491                2.677896               20.000978
##     thin.10.19.power0.15          population.log               schooling
##                19.573011                1.453709                9.350655
```

```
# Still multicollinearity issues, let's eliminate income.composition
fit.whr.who.only.3 = lm(life.expectancy~.-under.five.deaths.shift1.log-income.composition, data
summary(fit.whr.who.only.3)
```

```
##
## Call:
## lm(formula = life.expectancy ~ . - under.five.deaths.shift1.log -
##     income.composition, data = transformed.who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7478 -2.3240  0.1917  1.9613  7.7531
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.640e+01  8.330e+00   6.770 1.56e-09 ***
## developedTRUE            -3.565e-01  1.402e+00  -0.254  0.79994
## adult.mortality.sqrt     -3.284e-01  1.038e-01  -3.163  0.00216 **
## infant.deaths.shift1.log -6.131e-01  3.858e-01  -1.589  0.11573
## hivaids.1oversqrt         3.842e+00  5.415e-01   7.095 3.61e-10 ***
## measles.shift1.log       -8.654e-02  1.537e-01  -0.563  0.57486
## hep.b.power4             -3.244e-08  2.236e-08  -1.451  0.15056
## polio.power4              3.329e-08  2.386e-08   1.395  0.16653
## diphtheria                3.259e-03  3.917e-02   0.083  0.93387
## percent.expenditure.sqrt  4.823e-01  1.183e+00   0.408  0.68440
## total.expenditure        -3.229e-01  1.873e-01  -1.724  0.08838 .
## gdp.per.cap.log           3.209e-01  3.246e-01   0.989  0.32563
## alcohol.sqrt              4.589e-01  5.713e-01   0.803  0.42410
## bmi                       1.152e-01  2.998e-01   0.384  0.70176
## thin.5.9.power0.15       -2.164e+00  1.713e+00  -1.263  0.21005
## thin.10.19.power0.15      5.621e-01  1.731e+00   0.325  0.74617
## population.log            1.052e-01  1.791e-01   0.588  0.55840
## schooling                 3.882e-01  2.767e-01   1.403  0.16430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.797 on 85 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8431
## F-statistic: 33.25 on 17 and 85 DF,  p-value: < 2.2e-16
```

```r
vif(fit.whr.who.only.3)
```

```
##                developed     adult.mortality.sqrt infant.deaths.shift1.log 
##                 2.113105                 1.835720                 3.431195 
##        hivaids.1oversqrt       measles.shift1.log             hep.b.power4 
##                 2.628530                 1.701373                 2.849488 
##             polio.power4               diphtheria percent.expenditure.sqrt 
##                 3.148754                 2.740286                 2.130649 
##        total.expenditure          gdp.per.cap.log             alcohol.sqrt 
##                 1.511142                 2.242752                 2.461666 
##                      bmi       thin.5.9.power0.15     thin.10.19.power0.15 
##                 2.284404                19.952877                19.251066 
##           population.log                schooling 
##                 1.426343                 4.649129
```

```r
# Still multicollinearity issues, let's eliminate thin.5.9
fit.whr.who.only.4 = lm(life.expectancy~.-under.five.deaths.shift1.log-income.composition-thin
summary(fit.whr.who.only.4)
```

```
## 
## Call:
## lm(formula = life.expectancy ~ . - under.five.deaths.shift1.log - 
##     income.composition - thin.5.9.power0.15, data = transformed.who)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -8.139 -2.479  0.024  1.983  7.776 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               5.555e+01  8.332e+00   6.668 2.38e-09 ***
## developedTRUE            -2.559e-01  1.405e+00  -0.182 0.855891    
## adult.mortality.sqrt     -3.513e-01  1.026e-01  -3.425 0.000944 ***
## infant.deaths.shift1.log -6.555e-01  3.857e-01  -1.699 0.092842 .  
## hivaids.1oversqrt         3.702e+00  5.319e-01   6.960 6.35e-10 ***
## measles.shift1.log       -1.018e-01  1.537e-01  -0.662 0.509462    
## hep.b.power4             -3.311e-08  2.243e-08  -1.476 0.143544    
## polio.power4              3.214e-08  2.392e-08   1.343 0.182663    
## diphtheria                6.639e-03  3.921e-02   0.169 0.865958    
## percent.expenditure.sqrt  6.761e-01  1.177e+00   0.575 0.567053    
## total.expenditure        -3.205e-01  1.880e-01  -1.705 0.091841 .  
## gdp.per.cap.log           3.341e-01  3.255e-01   1.026 0.307572    
## alcohol.sqrt              3.545e-01  5.673e-01   0.625 0.533627    
## bmi                       1.350e-01  3.004e-01   0.449 0.654243    
## thin.10.19.power0.15     -1.493e+00  5.925e-01  -2.520 0.013597 *  
## population.log            1.059e-01  1.797e-01   0.589 0.557215    
## schooling                 4.140e-01  2.769e-01   1.495 0.138507    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.81 on 86 degrees of freedom
## Multiple R-squared:  0.8668, Adjusted R-squared:  0.842
## F-statistic: 34.98 on 16 and 86 DF,  p-value: < 2.2e-16
```

```
vif(fit.whr.who.only.4)
```

```
##                developed     adult.mortality.sqrt infant.deaths.shift1.log
##                 2.106288                 1.779663                 3.405306
##        hivaids.1oversqrt       measles.shift1.log             hep.b.power4
##                 2.518917                 1.690799                 2.847856
##             polio.power4               diphtheria percent.expenditure.sqrt
##                 3.144151                 2.727500                 2.094784
##        total.expenditure          gdp.per.cap.log             alcohol.sqrt
##                 1.510977                 2.240417                 2.410196
##                      bmi     thin.10.19.power0.15           population.log
##                 2.278139                 2.240292                 1.426331
##                schooling
##                 4.623678
```

3.8 Appendix H: Eliminating Variables with VIFs for Combined WHO and WHR Predictors

```
fit.who.and.whr.1 = lm(life.expectancy~., data = transformed.who.and.whr)
summary(fit.who.and.whr.1)
```

```
##
## Call:
## lm(formula = life.expectancy ~ ., data = transformed.who.and.whr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9561 -1.4584 -0.1146  1.5857  6.4221
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   5.577e+01  7.500e+00   7.435 1.41e-10 ***
## developedTRUE                 7.622e-01  1.337e+00   0.570 0.570450
## adult.mortality.sqrt         -2.940e-01  8.358e-02  -3.517 0.000744 ***
## infant.deaths.shift1.log      8.294e-01  1.437e+00   0.577 0.565604
## under.five.deaths.shift1.log -9.466e-01  1.482e+00  -0.639 0.524925
## hivaids.1oversqrt             1.924e+00  6.040e-01   3.185 0.002107 **
## measles.shift1.log           -1.291e-01  1.276e-01  -1.012 0.314962
## hep.b.power4                  3.287e-08  2.120e-08   1.551 0.125187
## polio.power4                 -4.906e-09  2.026e-08  -0.242 0.809317
## diphtheria                   -5.084e-02  3.357e-02  -1.515 0.134095
## percent.expenditure.sqrt      2.283e+00  1.049e+00   2.177 0.032661 *  
## total.expenditure            -1.146e-01  1.623e-01  -0.706 0.482464    
## income.composition            5.811e+01  1.146e+01   5.071 2.77e-06 ***
## gdp.per.cap.log              -1.719e-01  2.778e-01  -0.619 0.537937    
## alcohol.sqrt                 -5.042e-01  4.977e-01  -1.013 0.314334    
## bmi                          -6.486e-01  2.967e-01  -2.186 0.031939 *  
## thin.5.9.power0.15           -3.299e+00  1.351e+00  -2.441 0.016990 *  
## thin.10.19.power0.15          2.944e+00  1.413e+00   2.083 0.040626 *  
## population.log               -3.325e-02  1.480e-01  -0.225 0.822832    
## schooling                    -1.129e+00  3.447e-01  -3.275 0.001599 ** 
## social.support               -7.953e-01  3.644e+00  -0.218 0.827845    
## freedom.to.make.life.choices  1.072e+00  3.262e+00   0.329 0.743409    
## generosity                   -1.093e+00  2.632e+00  -0.415 0.679126    
## perceptions.of.corruption    -3.304e+00  2.597e+00  -1.272 0.207317    
## positive.affect               8.730e+00  4.333e+00   2.015 0.047508 *  
## negative.affect               1.326e+01  6.078e+00   2.181 0.032304 *  
## democratic.quality           -2.826e-02  1.115e+00  -0.025 0.979858    
## delivery.quality             -1.417e-01  1.130e+00  -0.125 0.900503    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.938 on 75 degrees of freedom
## Multiple R-squared:  0.9309, Adjusted R-squared:  0.9061 
## F-statistic: 37.45 on 27 and 75 DF,  p-value: < 2.2e-16
```

```r
# Examine VIFs
vif(fit.who.and.whr.1)
```

```
##                    developed         adult.mortality.sqrt 
##                     3.210888                     1.987509 
##     infant.deaths.shift1.log under.five.deaths.shift1.log 
##                    79.520824                    91.916429 
##            hivaids.1oversqrt           measles.shift1.log 
##                     5.463158                     1.959587 
##                 hep.b.power4                 polio.power4 
##                     4.277107                     3.793353 
##                   diphtheria     percent.expenditure.sqrt 
##                     3.361345                     2.800421 
##            total.expenditure           income.composition 
##                     1.895026                    40.897620 
##              gdp.per.cap.log                 alcohol.sqrt 
##                     2.744169                     3.120595 
##                          bmi           thin.5.9.power0.15 
##                     3.738301                    20.733076 
##         thin.10.19.power0.15               population.log 
##                    21.430543                     1.626260 
##                    schooling               social.support 
##                    12.051116                     2.783639 
## freedom.to.make.life.choices                   generosity
##                     2.995516                     2.094476
##    perceptions.of.corruption              positive.affect
##                     2.074353                     3.162948
##              negative.affect           democratic.quality
##                     2.479749                     9.277480
##             delivery.quality
##                    11.498935
```

```
# Big multicollinearity issues, let's eliminate under.five.deaths
fit.who.and.whr.2 = lm(life.expectancy~.-under.five.deaths.shift1.log, data = transformed.who.a
summary(fit.who.and.whr.2)
```

```
## 
## Call:
## lm(formula = life.expectancy ~ . - under.five.deaths.shift1.log, 
##     data = transformed.who.and.whr)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.9820 -1.5592 -0.2994  1.6504  6.3514 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                   5.501e+01  7.378e+00   7.456  1.2e-10 ***
## developedTRUE                 8.609e-01  1.323e+00   0.651 0.517281    
## adult.mortality.sqrt         -2.958e-01  8.320e-02  -3.555 0.000655 ***
## infant.deaths.shift1.log     -6.085e-02  3.495e-01  -0.174 0.862256    
## hivaids.1oversqrt             1.942e+00  6.010e-01   3.232 0.001818 ** 
## measles.shift1.log           -1.455e-01  1.245e-01  -1.169 0.246215    
## hep.b.power4                  3.167e-08  2.103e-08   1.506 0.136288    
## polio.power4                 -5.402e-09  2.017e-08  -0.268 0.789540    
## diphtheria                   -5.095e-02  3.343e-02  -1.524 0.131660    
## percent.expenditure.sqrt      2.247e+00  1.043e+00   2.153 0.034466 *  
## total.expenditure            -8.169e-02  1.533e-01  -0.533 0.595765    
## income.composition            5.946e+01  1.122e+01   5.299  1.1e-06 ***
## gdp.per.cap.log              -1.721e-01  2.767e-01  -0.622 0.535917    
## alcohol.sqrt                 -5.310e-01  4.940e-01  -1.075 0.285836    
## bmi                          -6.514e-01  2.955e-01  -2.204 0.030555 *  
## thin.5.9.power0.15           -3.315e+00  1.346e+00  -2.463 0.016037 *  
## thin.10.19.power0.15          2.991e+00  1.406e+00   2.128 0.036568 *  
## population.log               -5.284e-02  1.442e-01  -0.366 0.715017    
## schooling                    -1.133e+00  3.433e-01  -3.301 0.001471 ** 
## social.support               -1.027e+00  3.612e+00  -0.284 0.777008    
## freedom.to.make.life.choices  9.792e-01  3.246e+00   0.302 0.763697    
## generosity                   -9.233e-01  2.608e+00  -0.354 0.724301    
## perceptions.of.corruption    -3.071e+00  2.562e+00  -1.199 0.234309    
## positive.affect               8.843e+00  4.312e+00   2.051 0.043742 *  
## negative.affect               1.321e+01  6.054e+00   2.182 0.032194 *
## democratic.quality           -5.511e-02  1.110e+00  -0.050 0.960541
## delivery.quality             -1.242e-01  1.125e+00  -0.110 0.912413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.927 on 76 degrees of freedom
## Multiple R-squared:  0.9306, Adjusted R-squared:  0.9068
## F-statistic: 39.18 on 26 and 76 DF,  p-value: < 2.2e-16
```

```
vif(fit.who.and.whr.2)
```

```
##                    developed         adult.mortality.sqrt
##                     3.168025                     1.985237
##     infant.deaths.shift1.log            hivaids.1oversqrt
##                     4.739872                     5.450782
##           measles.shift1.log                 hep.b.power4
##                     1.880188                     4.243291
##                 polio.power4                   diphtheria
##                     3.787795                     3.361245
##     percent.expenditure.sqrt            total.expenditure
##                     2.792087                     1.704407
##           income.composition              gdp.per.cap.log
##                    39.514492                     2.744166
##                 alcohol.sqrt                          bmi
##                     3.098384                     3.737532
##           thin.5.9.power0.15         thin.10.19.power0.15
##                    20.726114                    21.371482
##               population.log                    schooling
##                     1.556350                    12.046996
##               social.support freedom.to.make.life.choices
##                     2.756136                     2.989630
##                   generosity    perceptions.of.corruption
##                     2.073166                     2.033549
##              positive.affect              negative.affect
##                     3.157637                     2.479380
##           democratic.quality             delivery.quality
##                     9.264298                    11.492119
```

```
# Still multicollinearity issues, let's eliminate income.composition
fit.who.and.whr.3 = lm(life.expectancy~.-under.five.deaths.shift1.log-income.composition, data
summary(fit.who.and.whr.3)
```

```
##
## Call:
## lm(formula = life.expectancy ~ . - under.five.deaths.shift1.log -
##     income.composition, data = transformed.who.and.whr)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -7.7973 -1.7640 -0.3038  1.9951  6.8845 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  4.768e+01  8.426e+00   5.659 2.49e-07 ***
## developedTRUE                2.646e-01  1.533e+00   0.173  0.86341    
## adult.mortality.sqrt        -2.967e-01  9.673e-02  -3.068  0.00298 ** 
## infant.deaths.shift1.log    -6.687e-01  3.838e-01  -1.742  0.08549 .  
## hivaids.1oversqrt            3.963e+00  5.400e-01   7.340 1.88e-10 ***
## measles.shift1.log          -1.097e-01  1.446e-01  -0.759  0.45017    
## hep.b.power4                -1.337e-08  2.236e-08  -0.598  0.55163    
## polio.power4                 2.314e-08  2.260e-08   1.024  0.30899    
## diphtheria                  -3.559e-02  3.873e-02  -0.919  0.36097    
## percent.expenditure.sqrt    -1.194e-01  1.096e+00  -0.109  0.91358    
## total.expenditure           -1.874e-01  1.768e-01  -1.060  0.29231    
## gdp.per.cap.log             -1.690e-03  3.195e-01  -0.005  0.99579    
## alcohol.sqrt                 3.762e-01  5.388e-01   0.698  0.48709    
## bmi                          1.408e-01  2.964e-01   0.475  0.63606    
## thin.5.9.power0.15          -3.063e+00  1.564e+00  -1.959  0.05379 .  
## thin.10.19.power0.15         2.370e+00  1.628e+00   1.455  0.14965    
## population.log               5.572e-02  1.659e-01   0.336  0.73791    
## schooling                    1.990e-01  2.718e-01   0.732  0.46638    
## social.support               1.355e+00  4.167e+00   0.325  0.74600    
## freedom.to.make.life.choices  4.344e+00  3.701e+00   1.174  0.24410    
## generosity                  -3.493e+00  2.979e+00  -1.173  0.24458    
## perceptions.of.corruption    5.516e-01  2.870e+00   0.192  0.84810    
## positive.affect              1.096e+01  4.992e+00   2.196  0.03113 *  
## negative.affect              1.332e+01  7.038e+00   1.892  0.06223 .  
## democratic.quality          -2.049e+00  1.215e+00  -1.687  0.09568 .  
## delivery.quality             3.345e+00  1.064e+00   3.145  0.00236 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.403 on 77 degrees of freedom
## Multiple R-squared:  0.9049, Adjusted R-squared:  0.874 
## F-statistic: 29.31 on 25 and 77 DF,  p-value: < 2.2e-16
```

```
vif(fit.who.and.whr.3)
```

```
##                    developed         adult.mortality.sqrt 
##                     3.145117                     1.985227 
##     infant.deaths.shift1.log            hivaids.1oversqrt 
##                     4.229373                     3.255524 
##           measles.shift1.log                 hep.b.power4 
##                     1.874655                     3.550256 
##                 polio.power4                   diphtheria 
##                     3.517626                     3.335965 
##     percent.expenditure.sqrt            total.expenditure
##                     2.280767                     1.675548
##              gdp.per.cap.log                 alcohol.sqrt
##                     2.707119                     2.726280
##                          bmi           thin.5.9.power0.15
##                     2.781266                    20.700168
##          thin.10.19.power0.15               population.log
##                    21.222757                     1.524927
##                     schooling               social.support
##                     5.587684                     2.713481
## freedom.to.make.life.choices                   generosity
##                     2.875231                     2.001468
##    perceptions.of.corruption              positive.affect
##                     1.888721                     3.130528
##              negative.affect           democratic.quality
##                     2.479352                     8.200678
##             delivery.quality
##                     7.600205
```

```
# Still multicollinearity issues, let's eliminate thin.10.19
fit.who.and.whr.4 = lm(life.expectancy~.-under.five.deaths.shift1.log-income.composition-thin.1
summary(fit.who.and.whr.4)
```

```
##
## Call:
## lm(formula = life.expectancy ~ . - under.five.deaths.shift1.log -
##     income.composition - thin.10.19.power0.15, data = transformed.who.and.whr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0482 -1.9198 -0.3051  1.9961  6.7054
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.876e+01  8.453e+00   5.768 1.54e-07 ***
## developedTRUE                 1.381e-01  1.541e+00   0.090  0.92885
## adult.mortality.sqrt         -3.169e-01  9.642e-02  -3.287  0.00152 **
## infant.deaths.shift1.log     -6.668e-01  3.866e-01  -1.725  0.08850 .
## hivaids.1oversqrt             3.765e+00  5.262e-01   7.155 3.99e-10 ***
## measles.shift1.log           -1.277e-01  1.451e-01  -0.880  0.38136
## hep.b.power4                 -1.483e-08  2.250e-08  -0.659  0.51177
## polio.power4                  2.240e-08  2.275e-08   0.984  0.32798
## diphtheria                   -3.050e-02  3.884e-02  -0.785  0.43476
## percent.expenditure.sqrt     -7.149e-03  1.101e+00  -0.006  0.99484
## total.expenditure            -1.862e-01  1.780e-01  -1.046  0.29875
## gdp.per.cap.log               1.388e-02  3.216e-01   0.043  0.96569
## alcohol.sqrt                  2.713e-01  5.377e-01   0.505  0.61527
## bmi                           1.708e-01  2.978e-01   0.574  0.56783
## thin.5.9.power0.15           -9.440e-01  5.750e-01  -1.642  0.10467
## population.log                6.099e-02  1.671e-01   0.365  0.71607
## schooling                     2.300e-01  2.729e-01   0.843  0.40199
## social.support                1.073e+00  4.192e+00   0.256  0.79865
## freedom.to.make.life.choices  4.102e+00  3.723e+00   1.102  0.27392
## generosity                   -3.124e+00  2.990e+00  -1.045  0.29927
## perceptions.of.corruption     9.114e-01  2.880e+00   0.316  0.75250
## positive.affect               9.571e+00  4.935e+00   1.939  0.05606 .
## negative.affect               1.070e+01  6.854e+00   1.562  0.12241
## democratic.quality           -2.101e+00  1.223e+00  -1.719  0.08964 .
## delivery.quality              3.377e+00  1.071e+00   3.153  0.00230 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.427 on 78 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.8722
## F-statistic: 30.01 on 24 and 78 DF,  p-value: < 2.2e-16
```

```r
vif(fit.who.and.whr.4)
```

```
##                    developed         adult.mortality.sqrt
##                     3.135002                     1.944328
##     infant.deaths.shift1.log            hivaids.1oversqrt
##                     4.229327                     3.047790
##           measles.shift1.log                 hep.b.power4
##                     1.860951                     3.543128
##                 polio.power4                   diphtheria
##                     3.515820                     3.308723
##     percent.expenditure.sqrt            total.expenditure
##                     2.269482                     1.675511
##              gdp.per.cap.log                 alcohol.sqrt
##                     2.704085                     2.677489
##                          bmi           thin.5.9.power0.15
##                     2.767793                     2.759434
##               population.log                    schooling
##                     1.524203                     5.553360
##               social.support freedom.to.make.life.choices
##                     2.707633                     2.869463
##                   generosity    perceptions.of.corruption
##                     1.986940                     1.874709
##              positive.affect              negative.affect
##                     3.015956                     2.317875
##           democratic.quality             delivery.quality
##                     8.193412                     7.597046
```

```r
# Use stepwise
fit.who.and.whr.5 = stepAIC(fit.who.and.whr.4, direction = "both", trace = 0)
summary(fit.who.and.whr.5)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log +
##     hivaids.1oversqrt + diphtheria + thin.5.9.power0.15 + schooling +
##     freedom.to.make.life.choices + generosity + positive.affect +
##     negative.affect + democratic.quality + delivery.quality,
##     data = transformed.who.and.whr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0547 -1.8094 -0.3603  1.9808  7.1545
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  51.24135    4.70366  10.894  < 2e-16 ***
## adult.mortality.sqrt         -0.29899    0.08633  -3.463 0.000819 ***
## infant.deaths.shift1.log     -0.85131    0.27731  -3.070 0.002831 **
## hivaids.1oversqrt             3.89643    0.46865   8.314 9.15e-13 ***
## diphtheria                   -0.03641    0.02280  -1.597 0.113759
## thin.5.9.power0.15           -0.97420    0.50039  -1.947 0.054669 .
## schooling                     0.38327    0.20998   1.825 0.071271 .
## freedom.to.make.life.choices  4.59028    3.25857   1.409 0.162375
## generosity                   -3.87749    2.44477  -1.586 0.116239
## positive.affect              11.73560    3.95178   2.970 0.003821 **
## negative.affect              11.75441    5.18549   2.267 0.025799 *
## democratic.quality           -1.95436    1.02945  -1.898 0.060841 .
## delivery.quality              3.13232    0.89753   3.490 0.000750 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.282 on 90 degrees of freedom
## Multiple R-squared:  0.8966, Adjusted R-squared:  0.8828
## F-statistic: 65.04 on 12 and 90 DF,  p-value: < 2.2e-16
```

```
# Eliminate variables one by one
fit.who.and.whr.6 = update(fit.who.and.whr.5,.~.-freedom.to.make.life.choices)
summary(fit.who.and.whr.6)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log +
##     hivaids.1oversqrt + diphtheria + thin.5.9.power0.15 + schooling +
##     generosity + positive.affect + negative.affect + democratic.quality +
##     delivery.quality, data = transformed.who.and.whr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8137 -1.8348 -0.2155  1.7391  7.3285
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              53.00468    4.55850  11.628  < 2e-16 ***
## adult.mortality.sqrt     -0.28502    0.08622  -3.306 0.001357 **
## infant.deaths.shift1.log -0.75942    0.27098  -2.802 0.006197 **
## hivaids.1oversqrt         3.89721    0.47118   8.271 1.05e-12 ***
## diphtheria               -0.03430    0.02287  -1.500 0.137201
## thin.5.9.power0.15       -0.98049    0.50307  -1.949 0.054374 .
## schooling                 0.38455    0.21111   1.822 0.071804 .
## generosity               -3.20513    2.41066  -1.330 0.186984
## positive.affect          13.90966    3.65758   3.803 0.000258 ***
## negative.affect          10.02391    5.06505   1.979 0.050835 .
## democratic.quality       -1.63001    1.00878  -1.616 0.109595
## delivery.quality          3.22106    0.90015   3.578 0.000557 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.3 on 91 degrees of freedom
## Multiple R-squared:  0.8943, Adjusted R-squared:  0.8816
## F-statistic: 70.02 on 11 and 91 DF,  p-value: < 2.2e-16
```

```r
fit.who.and.whr.7 = update(fit.who.and.whr.6,.~.-generosity)
summary(fit.who.and.whr.7)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log +
##     hivaids.1oversqrt + diphtheria + thin.5.9.power0.15 + schooling +
##     positive.affect + negative.affect + democratic.quality +
##     delivery.quality, data = transformed.who.and.whr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4644 -2.3775 -0.0019  1.8355  7.0837
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              54.08978    4.50352  12.011  < 2e-16 ***
## adult.mortality.sqrt     -0.28118    0.08653  -3.250 0.001615 **
## infant.deaths.shift1.log -0.77412    0.27189  -2.847 0.005439 **
## hivaids.1oversqrt         3.91114    0.47302   8.268 9.93e-13 ***
## diphtheria               -0.03143    0.02287  -1.375 0.172623
## thin.5.9.power0.15       -1.03002    0.50378  -2.045 0.043751 *
## schooling                 0.40202    0.21158   1.900 0.060546 .
## positive.affect          11.91328    3.34905   3.557 0.000595 ***
## negative.affect           9.76979    5.08252   1.922 0.057670 .
## democratic.quality       -1.45826    1.00464  -1.452 0.150037
## delivery.quality          2.94112    0.87882   3.347 0.001186 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.313 on 92 degrees of freedom
## Multiple R-squared:  0.8923, Adjusted R-squared:  0.8806
## F-statistic:  76.2 on 10 and 92 DF,  p-value: < 2.2e-16
```

```
fit.who.and.whr.8 = update(fit.who.and.whr.7,.~.-diphtheria)
summary(fit.who.and.whr.8)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log +
##     hivaids.1oversqrt + thin.5.9.power0.15 + schooling + positive.affect +
##     negative.affect + democratic.quality + delivery.quality,
##     data = transformed.who.and.whr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0518  -2.2933   0.0499   1.7431   7.3124
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              52.02169    4.26499  12.197  < 2e-16 ***
## adult.mortality.sqrt     -0.28840    0.08678  -3.323 0.001273 **
## infant.deaths.shift1.log -0.77821    0.27317  -2.849 0.005402 **
## hivaids.1oversqrt         3.80243    0.46859   8.115 1.95e-12 ***
## thin.5.9.power0.15       -1.00911    0.50595  -1.994 0.049027 *
## schooling                 0.38128    0.21204   1.798 0.075405 .
## positive.affect          11.74635    3.36281   3.493 0.000733 ***
## negative.affect           9.80883    5.10668   1.921 0.057822 .
## democratic.quality       -1.38069    1.00784  -1.370 0.173999
## delivery.quality          2.80115    0.87706   3.194 0.001917 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.329 on 93 degrees of freedom
## Multiple R-squared:  0.8901, Adjusted R-squared:  0.8794
## F-statistic: 83.66 on 9 and 93 DF,  p-value: < 2.2e-16
```

```
fit.who.and.whr.9 = update(fit.who.and.whr.8,.~.-democratic.quality)
summary(fit.who.and.whr.9)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log +
##     hivaids.1oversqrt + thin.5.9.power0.15 + schooling + positive.affect +
##     negative.affect + delivery.quality, data = transformed.who.and.whr)
```

```
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.0981  -2.1722  -0.1013   1.8822   7.1146 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)              51.79418    4.28159  12.097  < 2e-16 ***
## adult.mortality.sqrt     -0.27207    0.08636  -3.150 0.002187 ** 
## infant.deaths.shift1.log -0.69496    0.26756  -2.597 0.010904 *  
## hivaids.1oversqrt         4.05895    0.43154   9.406 3.37e-15 ***
## thin.5.9.power0.15       -0.91187    0.50328  -1.812 0.073201 .  
## schooling                 0.30625    0.20580   1.488 0.140074    
## positive.affect          11.83925    3.37777   3.505 0.000701 ***
## negative.affect          10.59698    5.09778   2.079 0.040365 *  
## delivery.quality          1.93946    0.61409   3.158 0.002134 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.345 on 94 degrees of freedom
## Multiple R-squared:  0.8878, Adjusted R-squared:  0.8783 
## F-statistic: 93.02 on 8 and 94 DF,  p-value: < 2.2e-16
```

```
fit.who.and.whr.10 = update(fit.who.and.whr.9,.~.-schooling)
summary(fit.who.and.whr.10)
```

```
## 
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log + 
##     hivaids.1oversqrt + thin.5.9.power0.15 + positive.affect + 
##     negative.affect + delivery.quality, data = transformed.who.and.whr)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -10.029  -2.142   0.008   1.928   6.937 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)              55.06152    3.69916  14.885  < 2e-16 ***
## adult.mortality.sqrt     -0.27783    0.08682  -3.200 0.001869 ** 
## infant.deaths.shift1.log -0.81818    0.25605  -3.195 0.001895 ** 
## hivaids.1oversqrt         4.30802    0.40030  10.762  < 2e-16 ***
## thin.5.9.power0.15       -0.98892    0.50379  -1.963 0.052577 .  
## positive.affect          12.23272    3.38886   3.610 0.000492 ***
## negative.affect          11.20283    5.11387   2.191 0.030922 *  
## delivery.quality          2.29020    0.57067   4.013 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.366 on 95 degrees of freedom
## Multiple R-squared:  0.8852, Adjusted R-squared:  0.8767
## F-statistic: 104.7 on 7 and 95 DF,  p-value: < 2.2e-16
```

```
fit.who.and.whr.11 = update(fit.who.and.whr.10,.~.-thin.5.9.power0.15)
summary(fit.who.and.whr.11)
```

```
##
## Call:
## lm(formula = life.expectancy ~ adult.mortality.sqrt + infant.deaths.shift1.log +
##     hivaids.1oversqrt + positive.affect + negative.affect + delivery.quality,
##     data = transformed.who.and.whr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.591  -1.986  -0.257   2.330   6.954
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               52.5070     3.5138  14.943  < 2e-16 ***
## adult.mortality.sqrt      -0.3008     0.0873  -3.445 0.000847 ***
## infant.deaths.shift1.log  -0.9627     0.2488  -3.868 0.000200 ***
## hivaids.1oversqrt          4.4092     0.4028  10.946  < 2e-16 ***
## positive.affect           13.7794     3.3446   4.120 8.04e-05 ***
## negative.affect           13.9296     4.9942   2.789 0.006373 **
## delivery.quality           2.6422     0.5497   4.806 5.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.416 on 96 degrees of freedom
## Multiple R-squared:  0.8805, Adjusted R-squared:  0.8731
## F-statistic: 117.9 on 6 and 96 DF,  p-value: < 2.2e-16
```

```
vif(fit.who.and.whr.11)
```

```
##     adult.mortality.sqrt infant.deaths.shift1.log        hivaids.1oversqrt
##                 1.604683                 1.764358                 1.798164
##          positive.affect          negative.affect         delivery.quality
##                 1.394716                 1.238993                 2.014922
```