Life Expectancy Data Analysis

Jorania F. Alves and Sneha Verma

12/18/2020

Note: Before running the data, change the location of the data

This dataset describes the life expectancy of multiple countries over a period of 15 years. It provides various health and economic factors regarding the country for that year, such as BMI (Body Mass Index), GDP (Gross Domestic Product), Hepatits B immunization coverage, etc. This research project will focus on the central question:

**Do health factors influence the life expectancy of a country?**

The researchers will look at the following predictor variables:

Adult Mortality: Adult mortality rates (probability of dying between 15 and 60 years per 1000 population).

infant deaths: Number of infant deaths per 1000 population.

Alcohol: recorded per capita (15+ years) consumption in litres of pure alochol.

Hepatitis B: Hepatitis B immunization coverage among one-year olds in percentage.

Measles: number of reported cases per 1000 population.

BMI: Average body mass index of entire population (the units were not provided, hence, the researchers will use the most common unit for describing BMI: kilograms per meter squared).

Polio: polio immunization coverage among one-year olds in percentage.

Diphtheria: Diphtheria tetanus toxoid and pertussis immunization coverage among one-year olds in percentage.

thinness 1-19 years: prevalance of thinness among children and adolescents between the ages 10 and 19 in percentage.

thinness 5-9 years: prevalance of thinness among children and adolescents between the ages 5 and 9 in percentage.

Predictor variable: Life expectancy in age (years)

It should be noted that since the units of are complex, interpretations of parameters will be mostly qualitative to indicate the direction of the influence.

## Import the dataset

lifeExpectancy = read.csv('/Users/sneha\_verma/Documents/MATH 327/Project1/Data/Life Expectancy Data.csv')  
  
# Remove any unnecessary variables:  
life\_expectancy0 = subset(lifeExpectancy, select = -c(Country, Year, Status, percentage.expenditure, under.five.deaths, Total.expenditure, HIV.AIDS, GDP, Population, Income.composition.of.resources, Schooling))  
  
# make dataset of complete observations for all columns.  
life\_expectancy = life\_expectancy0[complete.cases(life\_expectancy0), ]  
  
# Rename columns for convenience:  
colnames(life\_expectancy) = c("lifeExpectancy", "adult\_mortality", "infant\_deaths", "alcohol", "hepatitisB", "measles", "bmi", "polio", "diphtheria", "thin10to19", "thin5to9")

# Exploratory analysis

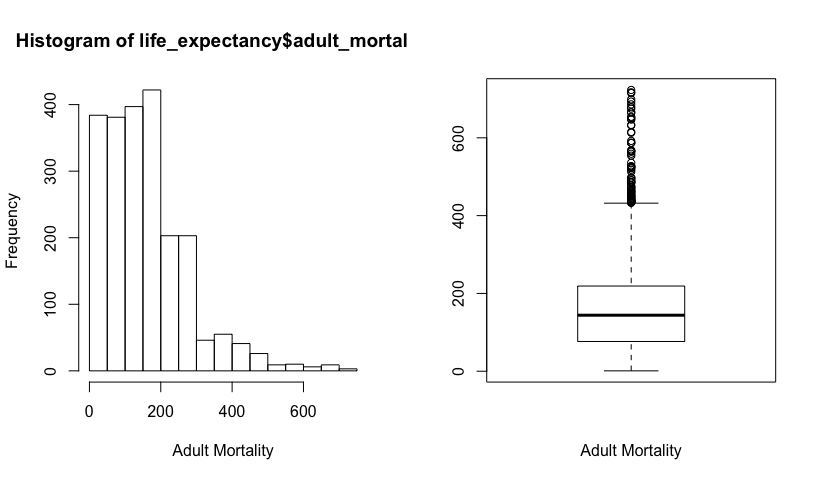
## adult\_mortality

summary(life\_expectancy$adult\_mortality)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 76.5 144.0 160.7 219.0 723.0

This shows that the minimum value of adult mortality is not 0, allowing us to make a log transformation, if necessary.

par(mfrow=c(1,2))  
hist(life\_expectancy$adult\_mortality, xlab = 'Adult Mortality')  
boxplot(life\_expectancy$adult\_mortality, xlab = 'Adult Mortality')



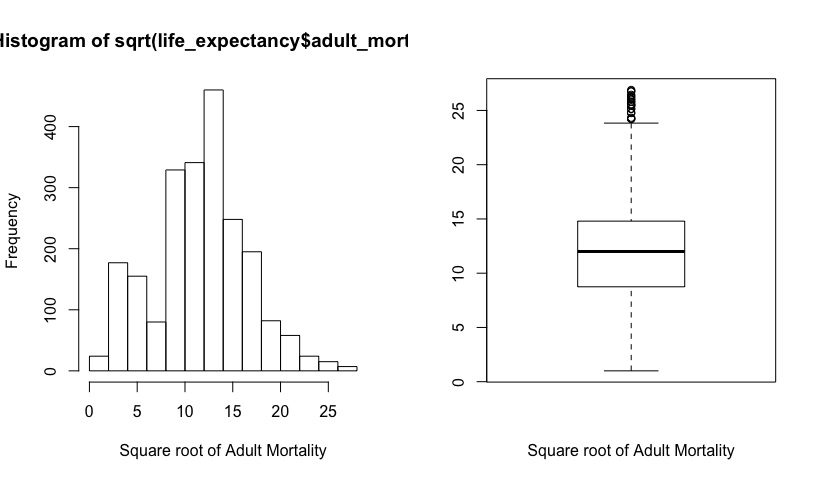
The distribution of adult mortality is clearly right-skewed. Let us try a log transformation to account for the skewness.

par(mfrow=c(1,2))  
hist(log(life\_expectancy$adult\_mortality), xlab = 'Log of Adult Mortality')  
boxplot(log(life\_expectancy$adult\_mortality), xlab = 'Log of Adult Mortality')



The log transformation overadjusts the skewness making the variable left-skewed. Thus, let us try a square-root transformation

par(mfrow=c(1,2))  
hist(sqrt(life\_expectancy$adult\_mortality), xlab = 'Square root of Adult Mortality')  
boxplot(sqrt(life\_expectancy$adult\_mortality), xlab = 'Square root of Adult Mortality')



The distribution of sqrt(adult-mortality) is more symmetric.

### infant\_deaths

summary(life\_expectancy$infant\_deaths)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 3.00 28.01 20.00 1600.00

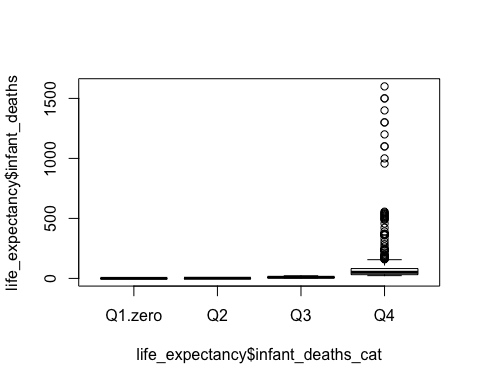
The minimum value of infant deaths is 0, indicating that a log transformation, if necessary, will not be possible since log(0) is undefined.

par(mfrow=c(1,2))  
hist(life\_expectancy$infant\_deaths, xlab = 'Number of infant deaths per 1000')  
boxplot(life\_expectancy$infant\_deaths, xlab = 'Number of infant deaths per 1000')



The distribution of the number of infant deaths per 1000 population is clearly right-skewed. Since this variable has the value zero, let us convert it into a categorical variable where each “bin” is set according to the quartiles found from the summary() function.

# Create a new variable where all values are "Q1.zero"  
life\_expectancy$infant\_deaths\_cat = "Q1.zero"  
  
# Change the new variable to "Q2" for the rows not in category 1  
life\_expectancy$infant\_deaths\_cat[life\_expectancy$infant\_deaths > 0] = "Q2"  
  
# Change the new variable to "Q3" for the rows not in category 1 or 2  
life\_expectancy$infant\_deaths\_cat[life\_expectancy$infant\_deaths > 3.0] = "Q3"  
  
# Change the new variable to "Q4" for the rows not in category 1, 2, or 3  
life\_expectancy$infant\_deaths\_cat[life\_expectancy$infant\_deaths > 22.0] = "Q4"  
  
# visualize the results  
boxplot(life\_expectancy$infant\_deaths ~ life\_expectancy$infant\_deaths\_cat)



# Count number of cases in each category  
table(life\_expectancy$infant\_deaths\_cat)

##   
## Q1.zero Q2 Q3 Q4   
## 613 574 491 517

### alcohol

summary(life\_expectancy$alcohol)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.010 0.690 3.500 4.445 7.540 17.870

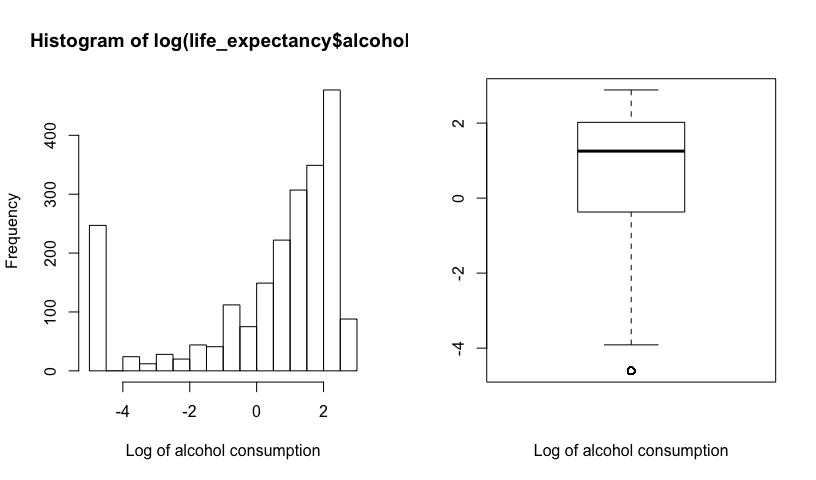
The minimum value of alcohol comsumption is 0.01, indicating that a log transformation, if necessary, will be possible.

par(mfrow=c(1,2))  
hist(life\_expectancy$alcohol, xlab = 'Alcohol consumption')  
boxplot(life\_expectancy$alcohol, xlab = 'Alcohol consumption')



The distribution of alcohol consumption is clearly right-skewed. Let us try a log transformation to account for the skewness.

par(mfrow=c(1,2))  
hist(log(life\_expectancy$alcohol), xlab = 'Log of alcohol consumption')  
boxplot(log(life\_expectancy$alcohol), xlab = 'Log of alcohol consumption')



The log transformation overadjusts the skewness making the variable left-skewed. Thus, let us try a square-root transformation

par(mfrow=c(1,2))  
hist(sqrt(life\_expectancy$alcohol), xlab = 'Log of alcohol consumption')  
boxplot(sqrt(life\_expectancy$alcohol), xlab = 'Log of alcohol consumption')



The distribution of sqrt(alcohol) is more symmetric.

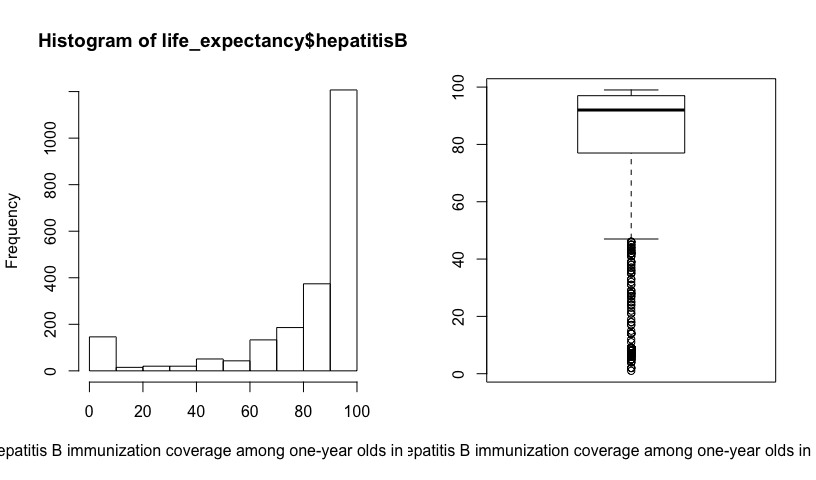
### hepatitsB

summary(life\_expectancy$hepatitisB)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 77.00 92.00 80.91 97.00 99.00

The minimum value of the percentage of one-year olds with HepatitisB immunization coverage is 1%.

par(mfrow=c(1,2))  
hist(life\_expectancy$hepatitisB, xlab = 'Hepatitis B immunization coverage among one-year olds in percent')  
boxplot(life\_expectancy$hepatitisB, xlab = 'Hepatitis B immunization coverage among one-year olds in percent')



The distribution of hepatits B among 1-year old in percent is clearly left-skewed. However, since the variable is in the units of percentage, let us leave the variable untouched and investigate its effects in the model.

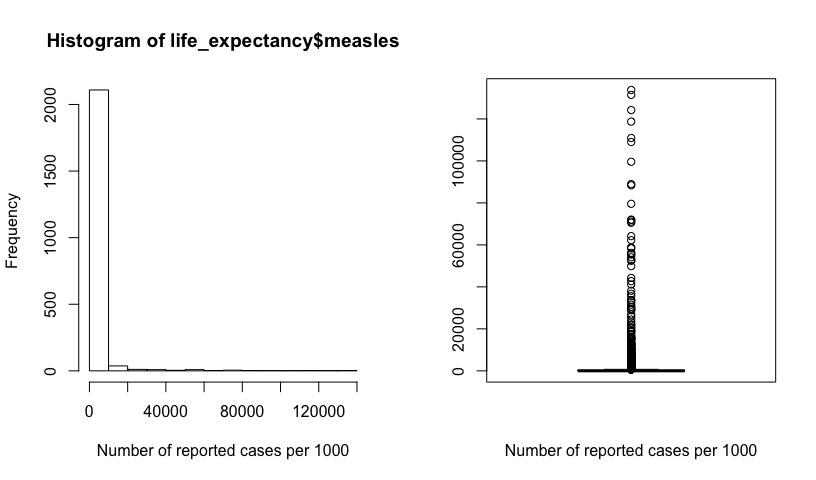
### measles

summary(life\_expectancy$measles)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 13 1995 290 133802

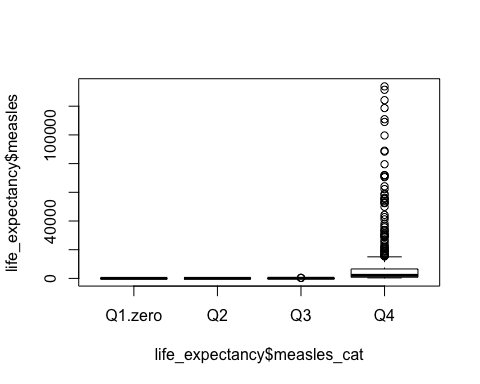
The minimum value of the number of reported cases of measles per 1000 is 0, indicating that a log transformation, if necessary, will not be possible since log(0) is undefined.

par(mfrow=c(1,2))  
hist(life\_expectancy$measles, xlab = 'Number of reported cases per 1000')  
boxplot(life\_expectancy$measles, xlab = 'Number of reported cases per 1000')



The distribution of the number of infant deaths per 1000 population is clearly right-skewed. Since this variable has the value zero, let us convert it into a categorical variable where each “bin” is set according to the quartiles found from the summary() function.

# Create a new variable where all values are "Q1.zero"  
life\_expectancy$measles\_cat = "Q1.zero"  
  
# Change the new variable to "Q2" for the rows not in category 1  
life\_expectancy$measles\_cat[life\_expectancy$measles > 0] = "Q2"  
  
# Change the new variable to "Q3" for the rows not in category 1 or 2  
life\_expectancy$measles\_cat[life\_expectancy$measles > 17.0] = "Q3"  
  
# Change the new variable to "Q4" for the rows not in category 1, 2, or 3  
life\_expectancy$measles\_cat[life\_expectancy$measles > 360.2] = "Q4"  
  
# visualize the results  
boxplot(life\_expectancy$measles ~ life\_expectancy$measles\_cat)



# Count number of cases in each category  
table(life\_expectancy$measles\_cat)

##   
## Q1.zero Q2 Q3 Q4   
## 762 379 548 506

### bmi

summary(life\_expectancy$bmi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.40 21.30 44.70 39.02 56.30 77.10

The minimum value of body mass index is 1.40 kg/m^2, indicating that a log transformation, if necessary, will be possible.

par(mfrow=c(1,2))  
hist(life\_expectancy$bmi, xlab = 'Average body mass index of population')  
boxplot(life\_expectancy$bmi, xlab = 'Average body mass index of population')



The distribution of average bmi across poopulations is bi-modal, which is acceptable for this multiple linear regression analysis.

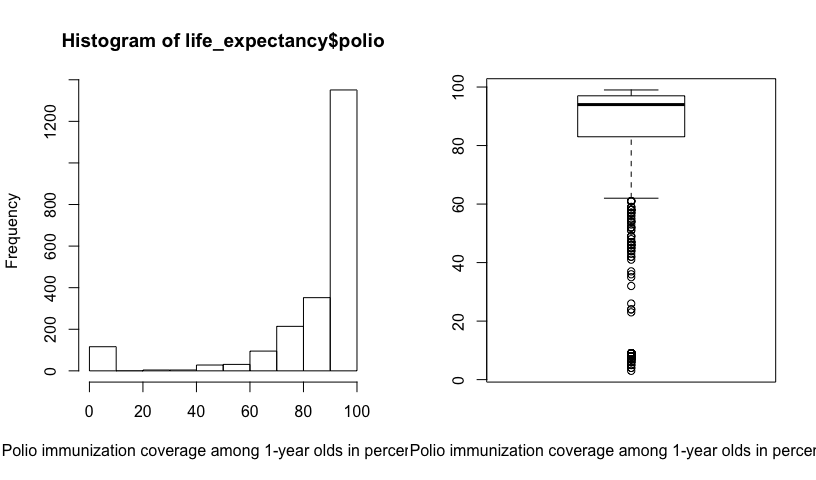
### polio

summary(life\_expectancy$polio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.00 83.00 94.00 84.97 97.00 99.00

The minimum value of one-year olds with polio immunization coverage is 3%, indicating that a log transformation, if necessary, will be possible.

par(mfrow=c(1,2))  
hist(life\_expectancy$polio, xlab = 'Polio immunization coverage among 1-year olds in percentage')  
boxplot(life\_expectancy$polio, xlab = 'Polio immunization coverage among 1-year olds in percentage')



The distribution of polio among 1 year olds cases is clearly left-skewed. However, since the variable is in the units of percentage, let us leave the variable untouched and investigate its effects in the model.

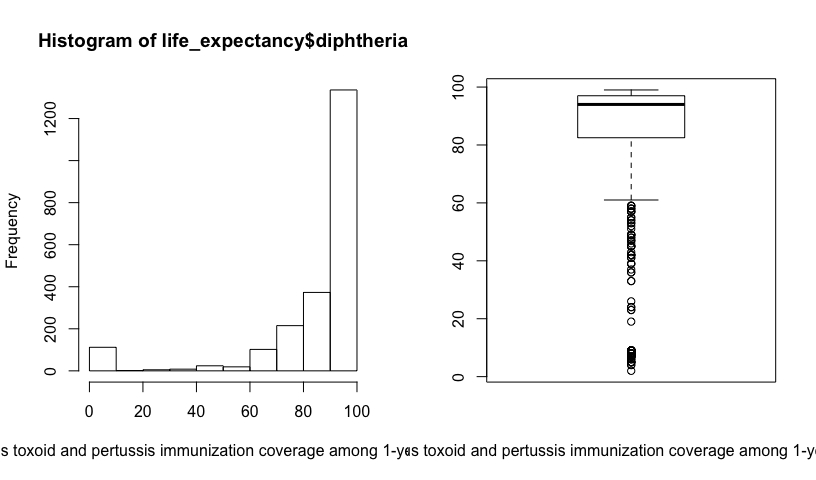
### diphtheria

summary(life\_expectancy$diphtheria)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 82.50 94.00 85.02 97.00 99.00

The minimum value of one-year olds with diphtheria immunization coverage is 2%, indicating that a log transformation, if necessary, will be possible.

par(mfrow=c(1,2))  
hist(life\_expectancy$diphtheria, xlab = 'Diphtheria tetanus toxoid and pertussis immunization coverage among 1-year olds in percentage')  
boxplot(life\_expectancy$diphtheria, xlab = 'Diphtheria tetanus toxoid and pertussis immunization coverage among 1-year olds in percentage')



The distribution of diphtheria cases cases is clearly left-skewed. However, since the variable is in the units of percentage, let us leave the variable untouched and investigate its effects in the model.

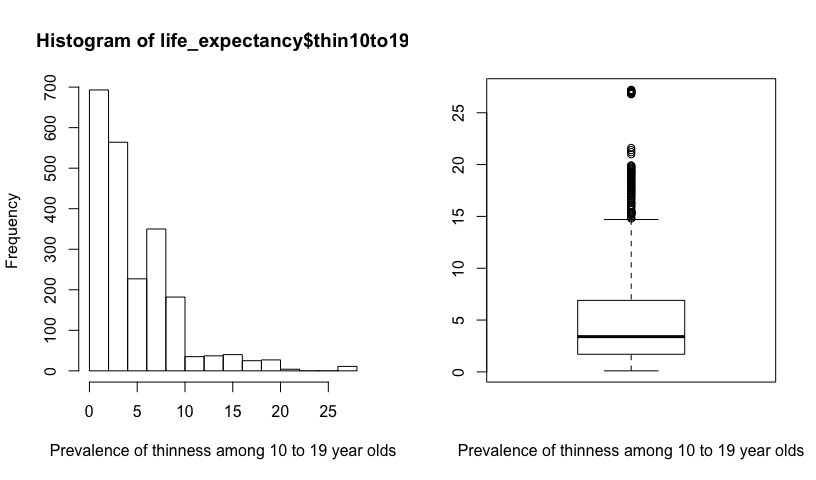
### thin10to19

summary(life\_expectancy$thin10to19)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 1.700 3.400 4.795 6.900 27.200

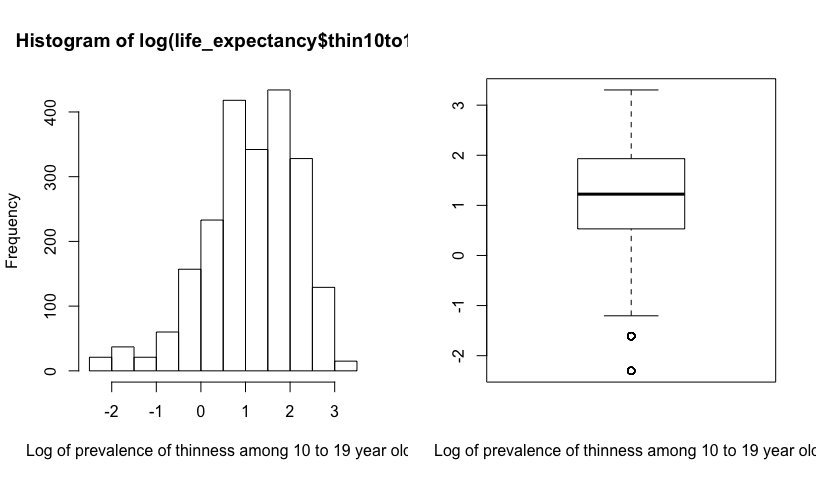
The minimum value of prevalence of thinnes among 10 to 19 year olds is 0.1%, indicating that a log transformation, if necessary, will be possible.

par(mfrow=c(1,2))  
hist(life\_expectancy$thin10to19, xlab = 'Prevalence of thinness among 10 to 19 year olds')  
boxplot(life\_expectancy$thin10to19, xlab = 'Prevalence of thinness among 10 to 19 year olds')



The distribution of the prevalence of thinness among 10 to 19 year olds is clearly right-skewed. Let us try a log transformation to account for the skewness.

par(mfrow=c(1,2))  
hist(log(life\_expectancy$thin10to19), xlab = 'Log of prevalence of thinness among 10 to 19 year olds')  
boxplot(log(life\_expectancy$thin10to19), xlab = 'Log of prevalence of thinness among 10 to 19 year olds')



The log(thin10to19) is more symmetric.

### thin5to9

summary(life\_expectancy$thin5to9)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 1.700 3.400 4.822 6.900 28.200

The minimum value of prevalence of thinnes among 5 to 9 year olds is 0.1%, indicating that a log transformation, if necessary, will be possible.

par(mfrow=c(1,2))  
hist(life\_expectancy$thin5to9, xlab = 'Prevalence of thinness among 5 to 9 year olds')  
boxplot(life\_expectancy$thin5to9, xlab = 'Prevalence of thinness among 5 to 9 year olds')



The distribution of the prevalence of thinness among 5 to 9 year olds is clearly right-skewed. Let us try a log transformation to account for the skewness.

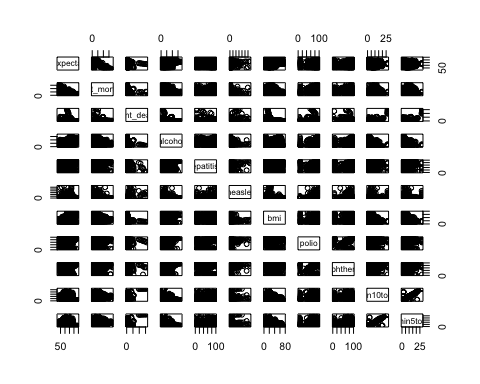
par(mfrow=c(1,2))  
hist(log(life\_expectancy$thin5to9), xlab = 'Log of prevalence of thinness among 5 to 9 year olds')  
boxplot(log(life\_expectancy$thin5to9), xlab = 'Log of prevalence of thinness among 5 to 9 year olds')



The distribution of log(thin5to9) is more symmetric.

## Scatterplot Matrix and Correlations

pairs(life\_expectancy[, 1:11])



Through the scatterplot matrix, it is clear that the relationship between the response variable (life expectancy) and the predictor variables is unclear because of the large number of the observations and the skewness in each variable (which has not been accounted for in this matrix).

Let us now look at pairwise correlations:

corr = cor(life\_expectancy[, 1:11], use = 'complete.obs')  
round(corr, 2)

## lifeExpectancy adult\_mortality infant\_deaths alcohol hepatitisB  
## lifeExpectancy 1.00 -0.72 -0.19 0.37 0.25  
## adult\_mortality -0.72 1.00 0.07 -0.14 -0.16  
## infant\_deaths -0.19 0.07 1.00 -0.10 -0.23  
## alcohol 0.37 -0.14 -0.10 1.00 0.09  
## hepatitisB 0.25 -0.16 -0.23 0.09 1.00  
## measles -0.10 0.03 0.52 -0.06 -0.13  
## bmi 0.51 -0.36 -0.23 0.28 0.15  
## polio 0.35 -0.23 -0.16 0.19 0.48  
## diphtheria 0.36 -0.23 -0.17 0.19 0.59  
## thin10to19 -0.44 0.27 0.44 -0.41 -0.13  
## thin5to9 -0.44 0.28 0.44 -0.39 -0.13  
## measles bmi polio diphtheria thin10to19 thin5to9  
## lifeExpectancy -0.10 0.51 0.35 0.36 -0.44 -0.44  
## adult\_mortality 0.03 -0.36 -0.23 -0.23 0.27 0.28  
## infant\_deaths 0.52 -0.23 -0.16 -0.17 0.44 0.44  
## alcohol -0.06 0.28 0.19 0.19 -0.41 -0.39  
## hepatitisB -0.13 0.15 0.48 0.59 -0.13 -0.13  
## measles 1.00 -0.16 -0.07 -0.08 0.16 0.16  
## bmi -0.16 1.00 0.20 0.18 -0.51 -0.52  
## polio -0.07 0.20 1.00 0.57 -0.16 -0.17  
## diphtheria -0.08 0.18 0.57 1.00 -0.17 -0.16  
## thin10to19 0.16 -0.51 -0.16 -0.17 1.00 0.94  
## thin5to9 0.16 -0.52 -0.17 -0.16 0.94 1.00

Observing the correlation matrix, it can be concluded that most predictor variables are not correlated to each other which shows that this dataset does not display multicollinearity. However, the variables ‘thin5to9’ and ‘thin10to19’ have a high correlation of 0.94 indicating the presence of multicollinearity. Adult mortality, followed by body mass index, are the most correlated with the response varibale, life expectancy.

Looking at the skewness and correlation, the variables indicating the percent of one-year olds with measles immunization coverage and the number of infant deaths should be categorical because they have large skewness with zero as values. Other variables that are left-skewed and have units in percentage have been left untouched to observe its significance in predicting life expectancy of a country.

To deal with the high correlation between variables thin5to9 and thin10to19, let us create a third variable with the average of the first two:

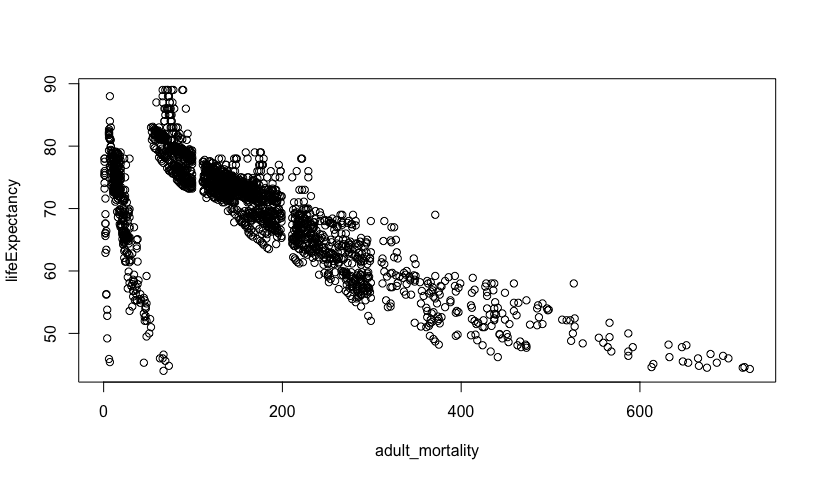
thin = c('thin5to9', 'thin10to19')  
life\_expectancy$thin\_avg = rowMeans(life\_expectancy[thin])

We will use the thin\_avg variable in our first-order model.

## Simple Linear Regression

Let us start with a simple linear regression model of the response variable and the predictor variable with the highest correlation with life expectancy, adult mortality:

plot(data = life\_expectancy, lifeExpectancy ~ adult\_mortality)



simpleFit = lm(data = life\_expectancy, lifeExpectancy ~ adult\_mortality)  
summary(simpleFit)

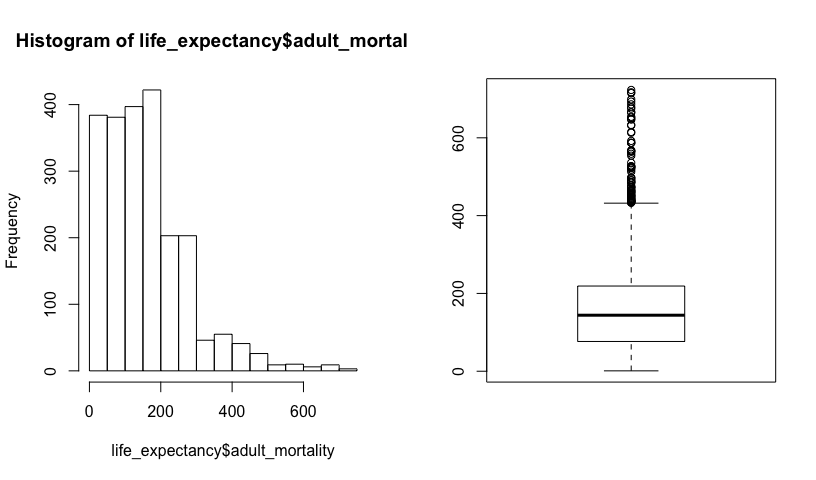
##   
## Call:  
## lm(formula = lifeExpectancy ~ adult\_mortality, data = life\_expectancy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.303 -2.449 1.113 3.524 15.457   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 78.058465 0.210921 370.1 <2e-16 \*\*\*  
## adult\_mortality -0.050737 0.001053 -48.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.904 on 2193 degrees of freedom  
## Multiple R-squared: 0.5144, Adjusted R-squared: 0.5142   
## F-statistic: 2323 on 1 and 2193 DF, p-value: < 2.2e-16

Looking at this plot, two clusters of points are visible indicating the possibility of errors in data entry or some unnatural events that created a large range of values of life expectancy for low adult mortality. We can explore the clusters with the following code, however, since multiple variables are displaying clusters, we will leave the variables untouched and continue with the regression analysis with the originally transformed variables.

# # create cluster 1  
# life\_expectancy$cluster1 = with(life\_expectancy, ifelse((lifeExpectancy < 70 & adult\_mortality < 100) | (lifeExpectancy >= 70 & adult\_mortality < 45), 1, 2))  
# plot(lifeExpectancy ~ adult\_mortality, data = life\_expectancy, col = cluster1)  
# life\_expectancy$adult\_mort2 = ifelse(life\_expectancy$cluster1 == 1, 10, 1) \* life\_expectancy$adult\_mortality  
# plot(lifeExpectancy ~ adult\_mort2, data = life\_expectancy, col = cluster1)  
#   
# # crate cluster 2  
# life\_expectancy$cluster2 = with(life\_expectancy, ifelse((lifeExpectancy < 90 - 0.3 \* adult\_mort2), 3, 4))  
# plot(lifeExpectancy ~ adult\_mort2, data = life\_expectancy, col = cluster2)  
# abline(90, -0.3)  
# life\_expectancy$adult\_mort3 = ifelse(life\_expectancy$cluster2 == 3, 10, 1) \* life\_expectancy$adult\_mort2  
# plot(lifeExpectancy ~ adult\_mort3, data = life\_expectancy, col = cluster2)

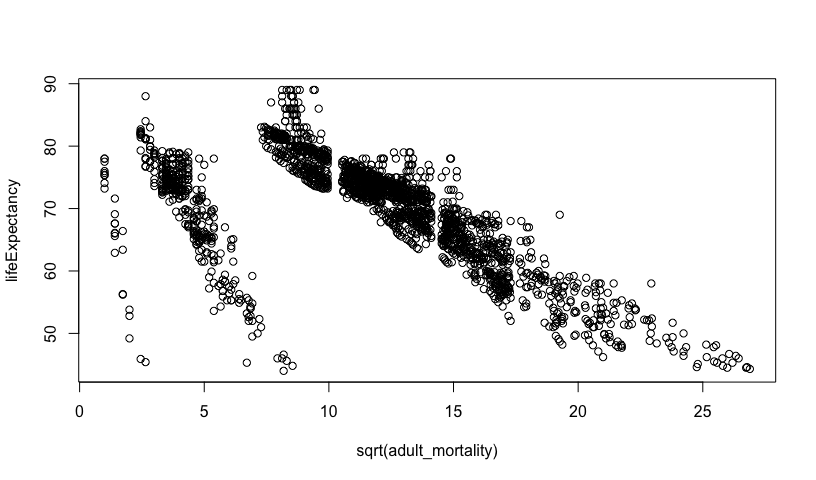
Let us have a second look at the distribution of adult mortality:

par(mfrow=c(1,2))  
hist(life\_expectancy$adult\_mortality)  
boxplot(life\_expectancy$adult\_mortality)



Since the variable is right-skewed, we transformed the variable with a square-root transformation. Let us use the transformed variable to fit a model:

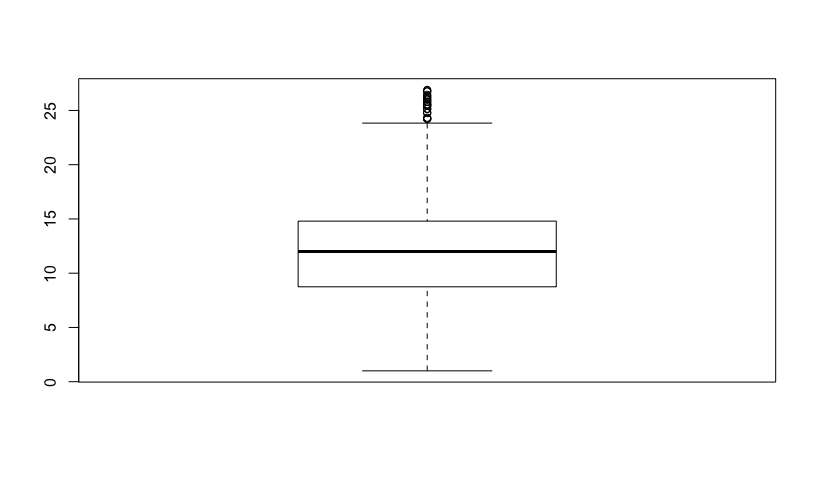
plot(data = life\_expectancy, lifeExpectancy ~ sqrt(adult\_mortality))



hist(sqrt(life\_expectancy$adult\_mortality))



boxplot(sqrt(life\_expectancy$adult\_mortality))



The intial plot shows that there are three clusters in adult\_mortality, not two (as we thought above). The sqrt(adult\_mortality) distribution is more symmetric.

simpleFit\_sqrt = lm(data = life\_expectancy, lifeExpectancy ~ sqrt(adult\_mortality))  
summary(simpleFit\_sqrt)

##   
## Call:  
## lm(formula = lifeExpectancy ~ sqrt(adult\_mortality), data = life\_expectancy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.311 -3.695 1.596 4.475 16.632   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 82.57292 0.37052 222.86 <2e-16 \*\*\*  
## sqrt(adult\_mortality) -1.08173 0.02923 -37.01 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.647 on 2193 degrees of freedom  
## Multiple R-squared: 0.3844, Adjusted R-squared: 0.3842   
## F-statistic: 1370 on 1 and 2193 DF, p-value: < 2.2e-16

Looking at the reggression results, one can observe that intercept is 82.6, indicating that when the square-root of adult mortality is 0, the life expectancy is 82.6 years. The slope is -1.08 indicating that when the square root of adult mortality increases by 1 square-root percentage per 1000 population, the life expectancy decreases by 1.08 years.

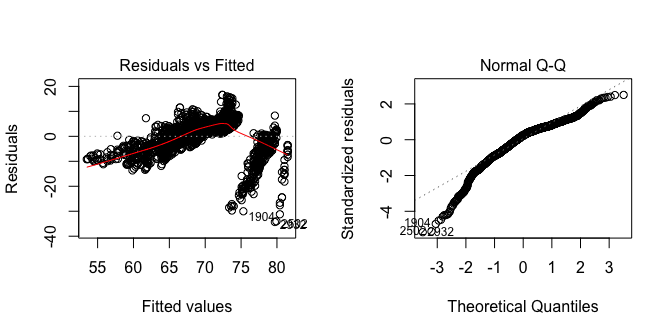
Let us now look at the confidence intervals and the residual plots:

confint(simpleFit\_sqrt)

## 2.5 % 97.5 %  
## (Intercept) 81.846316 83.299524  
## sqrt(adult\_mortality) -1.139054 -1.024413

The confidence interval of the slope is -1.14 to -1.02; this means that the mean life expectancy decreases between 1.02 to 1.14 years with a 1 square-root percent per 1000 increase in adult mortality.

par (mfrow = c(1,2))  
plot (simpleFit\_sqrt, which=1:2)



The residual plot shows non-linearity and non-constant variance. There are two clear clusters of points in the residual v/s fitted plot which were also reflected in the simple linear regression plot. There are more points beneath the 0.0 line and the red trendline shows curvature, indicating non-linearity. The quantile points show a linear trend indicating normally-distributed residuals. However, there is some deviation from the linear pattern at the tails indicating long-tailed residuals, but that does not indicate major deviation from the linear trend. Hence, the residuals follow the normal distribution condition.

## First Order Model

Next, we fit a first-order linear model will all nine predictors, with the transformations decided upon previously:

fit1 = lm(data = life\_expectancy, lifeExpectancy ~ sqrt(adult\_mortality) + infant\_deaths\_cat + hepatitisB + sqrt(alcohol) + measles\_cat + bmi + polio + diphtheria + thin\_avg)  
summary(fit1)

##   
## Call:  
## lm(formula = lifeExpectancy ~ sqrt(adult\_mortality) + infant\_deaths\_cat +   
## hepatitisB + sqrt(alcohol) + measles\_cat + bmi + polio +   
## diphtheria + thin\_avg, data = life\_expectancy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.048 -2.887 0.663 3.372 15.053   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.085e+01 8.374e-01 84.612 < 2e-16 \*\*\*  
## sqrt(adult\_mortality) -7.188e-01 2.595e-02 -27.702 < 2e-16 \*\*\*  
## infant\_deaths\_catQ2 -7.815e-01 3.194e-01 -2.447 0.014477 \*   
## infant\_deaths\_catQ3 -2.530e+00 3.529e-01 -7.169 1.03e-12 \*\*\*  
## infant\_deaths\_catQ4 -5.722e+00 4.274e-01 -13.386 < 2e-16 \*\*\*  
## hepatitisB -1.078e-05 5.786e-03 -0.002 0.998514   
## sqrt(alcohol) 7.350e-01 1.185e-01 6.205 6.54e-10 \*\*\*  
## measles\_catQ2 1.456e-01 3.390e-01 0.429 0.667606   
## measles\_catQ3 7.102e-01 3.155e-01 2.251 0.024483 \*   
## measles\_catQ4 1.523e+00 3.712e-01 4.102 4.24e-05 \*\*\*  
## bmi 7.260e-02 7.082e-03 10.251 < 2e-16 \*\*\*  
## polio 2.260e-02 6.742e-03 3.353 0.000814 \*\*\*  
## diphtheria 4.630e-02 7.323e-03 6.323 3.10e-10 \*\*\*  
## thin\_avg -2.038e-01 3.470e-02 -5.873 4.95e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.317 on 2181 degrees of freedom  
## Multiple R-squared: 0.6084, Adjusted R-squared: 0.606   
## F-statistic: 260.6 on 13 and 2181 DF, p-value: < 2.2e-16

From the regression analysis, the following variables are significant with a 0.001 significance level: the intercept, sqrt(adult\_mort3), infant\_deaths\_cat, sqrt(alcohol), measles\_cat, bmi, polio, and diphtheria. HepatitisB is not significant. The (adjusted) R-squared is 60.6% which implies that this model explains 60.6% of the variation in the response variable. The residual standard error is 5.317 years.

### Interpretation of parameters:

NOTE: Due to the complexity and unclarity of the units, while we will be interpreting the parameter effects with units, we will also interpret it in a qualitative manner by describing the direction of effect.

The intercept 70.85 is the estimated mean life expectancy in age (years), when all the other predictors are 0.

When the square root of the probability of dying between 15 and 60 years per 1000 population increases by 1 square root percent, the life expectancy decreases by 0.7188 years. Hence, when adult mortality increases, the life expectancy decreases, while holding other predictors as constant.

If the number of infant deaths are greater than 0 and less than 3 life expectancy will decrease by 0.7815 years, while holding other predictors constant. If the number of infant deaths are greater than 3 and less than 22 life expectancy will decrease by 2.530 years, while holding other predictors constant. If the number of infant deaths are greater than 22 life expectancy will decrease by 5.722 years, while holding other predictors constant. In overall, it can be seen as the number of infant deaths increase, the life expectancy keeps decreasing.

If hepatitis B immunization coverage among one-year olds increases by 1%, life expectancy will decrease by 1.078e-05 years, while holding other predictors constant.

As alcohol consumption increases by the square root of one square root litre of pure alcohol, the life expectancy increases by 0.735 years, while holding other predictors as constant.

If the number of reported measles cases increases by 0 to 17 cases, the life expectancy will increase by 0.1456 years, while holding other predictors as constant. If the number of reported measles cases increases by 17 to 360.2 cases, the life expectancy will increase by 0.7102 years, while holding other predictors as constant. If the number of reported measles cases increases by more than 360.2 cases, the life expectancy will increase by 1.523 years, while holding other predictors as constant. In general, it appears that as the number of measles cases increases, the life expectancy increases.

An increase in bmi by 1 kg/m^2 will increase life expectancy by 0.0726 years, while keeping other predictors constant.

An increase in polio immunization coverage among 1 year old by 1% will increase life expectancy by 0.02260 years, keeping other predictors constant.

An increase in diphtheria immunization coverage among 1 year old by 1%, will increase life expectancy by 0.0463, keeping other predictors constant.

An increase in the average thinness among 5 to 19 year olds will decrease life expectancy by 0.2038 years, while keeping other predictors constant.

anova(fit1)

## Analysis of Variance Table  
##   
## Response: lifeExpectancy  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sqrt(adult\_mortality) 1 60514 60514 2140.8845 < 2.2e-16 \*\*\*  
## infant\_deaths\_cat 3 21496 7165 253.5002 < 2.2e-16 \*\*\*  
## hepatitisB 1 1283 1283 45.4043 2.044e-11 \*\*\*  
## sqrt(alcohol) 1 3974 3974 140.5982 < 2.2e-16 \*\*\*  
## measles\_cat 3 235 78 2.7719 0.04017 \*   
## bmi 1 5143 5143 181.9525 < 2.2e-16 \*\*\*  
## polio 1 1003 1003 35.4838 2.990e-09 \*\*\*  
## diphtheria 1 1137 1137 40.2316 2.734e-10 \*\*\*  
## thin\_avg 1 975 975 34.4878 4.947e-09 \*\*\*  
## Residuals 2181 61648 28   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The analysis of variance table (ANOVA table) suggests that all variables except the categorical measles variable are significant predictors.

Let us check the variance inflation factor to look for correlation among variables:

library(car)

## Loading required package: carData

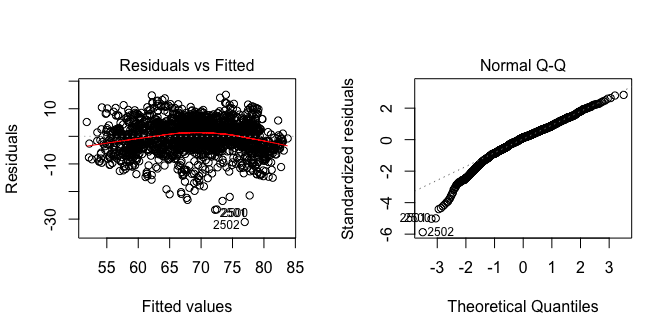
vif(fit1)

## GVIF Df GVIF^(1/(2\*Df))  
## sqrt(adult\_mortality) 1.231999 1 1.109954  
## infant\_deaths\_cat 1.999629 3 1.122427  
## hepatitisB 1.631066 1 1.277132  
## sqrt(alcohol) 1.311486 1 1.145201  
## measles\_cat 1.548812 3 1.075639  
## bmi 1.526217 1 1.235402  
## polio 1.636722 1 1.279344  
## diphtheria 1.890459 1 1.374940  
## thin\_avg 1.700353 1 1.303976

The VIF values of the variables (including the categorical variables) show that none of the variables in the model are correlated with each other, since all values are below 5.

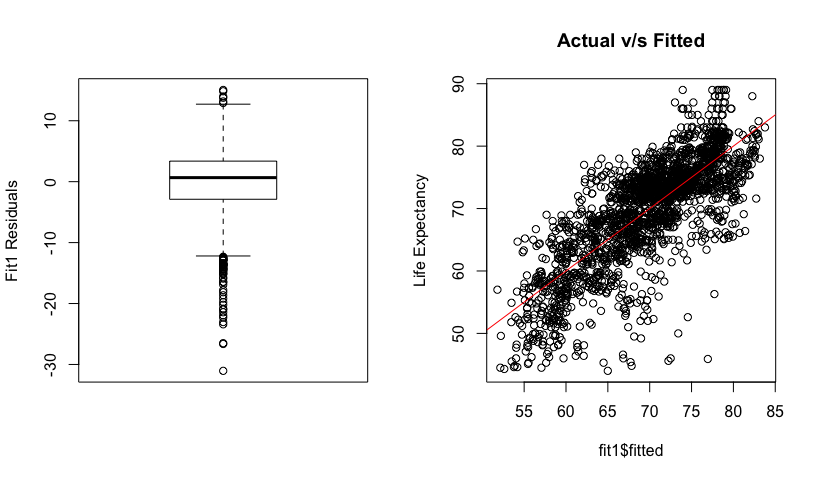
### Residual Analysis

par (mfrow = c(1,2))  
plot (fit1, which=1:2)



The residual v/s fitted value plot is useful for assessing linearity and constant variance conditions of a simple linear model. The residuals are scattered across the plot with high residuals and there appears to bee some curvature in the trendline, however, there does not appear to be an extreme non-linear trend in the residuals. Hence, the residuals are consistent with the linearity condition. Further, the residuals appear to have constant variance with points being scattered above and below the 0.0 line. The normal quantile plot is a scatterplot that displays the observed data v/s the values that would be expected from a normal sample of the same size. In this quantile plot, there does not appear to be any major deviation from a straight line. However, there are some left-skewed residuals but the trend is largely linear indicating that the deviation might be a small variation that can be ignored.

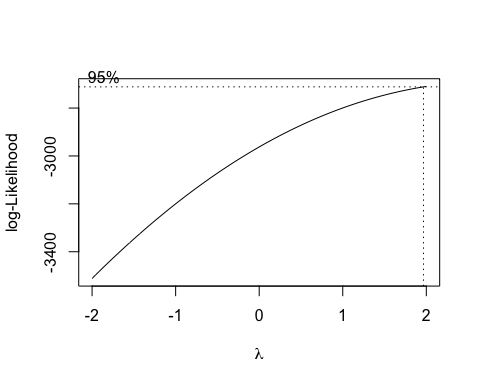
par(mfrow = c(1,2))  
boxplot(fit1$residuals, ylab = 'Fit1 Residuals')  
plot(lifeExpectancy~fit1$fitted, main = "Actual v/s Fitted", ylab = "Life Expectancy", data = life\_expectancy)  
abline(0,1, col = "red")



The box plot shows skewness in either ends of the residuals indicating some deviaation from a normal pattern. The Actual v/s Fitted plot shows a moderate linear trend.

## Box-Cox Analysis

library(MASS)  
boxcox(fit1)



The box-cox analysis suggests a squared power transformation, with lambda = 2, on the response variable.

fit2 = lm(data = life\_expectancy, lifeExpectancy^2 ~ sqrt(adult\_mortality) + infant\_deaths\_cat + hepatitisB + sqrt(alcohol) + measles\_cat + bmi + polio + diphtheria + thin\_avg)  
summary(fit2)

##   
## Call:  
## lm(formula = lifeExpectancy^2 ~ sqrt(adult\_mortality) + infant\_deaths\_cat +   
## hepatitisB + sqrt(alcohol) + measles\_cat + bmi + polio +   
## diphtheria + thin\_avg, data = life\_expectancy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3844.7 -414.7 63.4 442.9 2401.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5051.3667 111.4773 45.313 < 2e-16 \*\*\*  
## sqrt(adult\_mortality) -94.7458 3.4544 -27.427 < 2e-16 \*\*\*  
## infant\_deaths\_catQ2 -95.1125 42.5148 -2.237 0.025377 \*   
## infant\_deaths\_catQ3 -376.0221 46.9719 -8.005 1.92e-15 \*\*\*  
## infant\_deaths\_catQ4 -752.2956 56.9019 -13.221 < 2e-16 \*\*\*  
## hepatitisB -0.2626 0.7703 -0.341 0.733202   
## sqrt(alcohol) 119.7871 15.7695 7.596 4.50e-14 \*\*\*  
## measles\_catQ2 48.1528 45.1290 1.067 0.286089   
## measles\_catQ3 148.8318 41.9993 3.544 0.000403 \*\*\*  
## measles\_catQ4 236.5485 49.4208 4.786 1.81e-06 \*\*\*  
## bmi 9.4200 0.9428 9.992 < 2e-16 \*\*\*  
## polio 3.0310 0.8975 3.377 0.000745 \*\*\*  
## diphtheria 6.2238 0.9749 6.384 2.10e-10 \*\*\*  
## thin\_avg -30.3409 4.6191 -6.569 6.33e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 707.8 on 2181 degrees of freedom  
## Multiple R-squared: 0.6127, Adjusted R-squared: 0.6104   
## F-statistic: 265.4 on 13 and 2181 DF, p-value: < 2.2e-16

From the regression analysis, all variables, except hepatitisB, are significant at the 0.001 level. The (adjusted) R-squared is 61.04% which means that this model explains 61.04% of the variation in the response model. The residual standard error is 707.8 squared years.

anova(fit2)

## Analysis of Variance Table  
##   
## Response: lifeExpectancy^2  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sqrt(adult\_mortality) 1 1071435008 1071435008 2138.9702 < 2.2e-16 \*\*\*  
## infant\_deaths\_cat 3 385944250 128648083 256.8279 < 2.2e-16 \*\*\*  
## hepatitisB 1 20385373 20385373 40.6965 2.164e-10 \*\*\*  
## sqrt(alcohol) 1 92461996 92461996 184.5874 < 2.2e-16 \*\*\*  
## measles\_cat 3 6812518 2270839 4.5334 0.003568 \*\*   
## bmi 1 91136603 91136603 181.9415 < 2.2e-16 \*\*\*  
## polio 1 17930225 17930225 35.7952 2.555e-09 \*\*\*  
## diphtheria 1 20557872 20557872 41.0409 1.821e-10 \*\*\*  
## thin\_avg 1 21612408 21612408 43.1462 6.332e-11 \*\*\*  
## Residuals 2181 1092488238 500912   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The analysis of variance table (ANOVA table) suggests that all variables are significant at the 0.001 level, except the measles\_cat variable which is significant at the 0.01 level (which is higher than the first fitted model).

Let us check the variance inflation factor to check for correlation among variables:

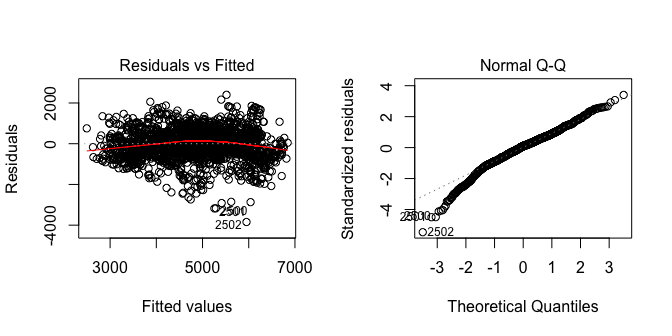
library(car)  
vif(fit2)

## GVIF Df GVIF^(1/(2\*Df))  
## sqrt(adult\_mortality) 1.231999 1 1.109954  
## infant\_deaths\_cat 1.999629 3 1.122427  
## hepatitisB 1.631066 1 1.277132  
## sqrt(alcohol) 1.311486 1 1.145201  
## measles\_cat 1.548812 3 1.075639  
## bmi 1.526217 1 1.235402  
## polio 1.636722 1 1.279344  
## diphtheria 1.890459 1 1.374940  
## thin\_avg 1.700353 1 1.303976

The VIF values of the variables (including the categorical variables) show that none of the variables in the model are correlated with each other, since all values are below 5.

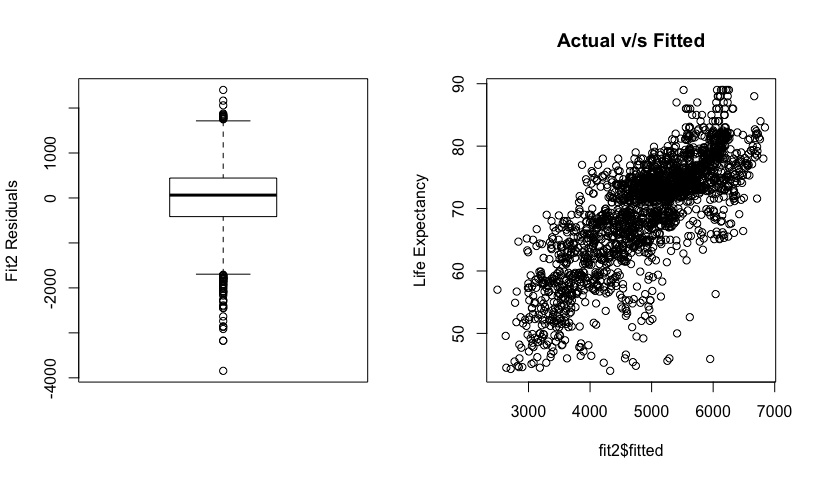
## Residual Analysis

par (mfrow = c(1,2))  
plot (fit2, which=1:2)



The residual v/s fitted value plot is useful for assessing linearity and constant variance conditions of a simple linear model. The residuals are scattered across the plot with high residuals at the bottom of the plot, however, the curvature is less than the previous model hence, there does not appear to be an extreme non-linear trend in the residuals. Thus, the residuals are consistent with the linearity condition. Further, there does not appear to be a major deviation from constant variance with points being scattered above and below the 0.0 line. The normal quantile plot is a scatterplot that displays the observed data v/s the values that would be expected from a normal sample of the same size. In this quantile plot, there does not appear to be any major deviation from a straight line. While there are no right-skewed residuals, there are some left-skewed residuals but the trend is largely linear indicating that the deviation might be a small variation that can be ignored.

par(mfrow = c(1,2))  
boxplot(fit2$residuals, ylab = 'Fit2 Residuals')  
plot(lifeExpectancy~fit2$fitted, main = "Actual v/s Fitted", ylab = "Life Expectancy", data = life\_expectancy)  
abline(0,1, col = "red")



The box plot shows slight skewness at the right end with large skewness on the left end, which was shown in the quantile plot, indicating some deviation from a normal pattern. The Actual v/s Fitted plot shows a stronger linear trend than the originally-fitted model.

## Backward Elimination

Looking at the regression results of the second fitted model, it can be concluded that Hepatitis B is not a significant predictor. Let us remove the variable and refit the model:

fit3 = lm(data = life\_expectancy, lifeExpectancy^2 ~ sqrt(adult\_mortality) + infant\_deaths\_cat + sqrt(alcohol) + measles\_cat + bmi + polio + diphtheria + thin\_avg)  
summary(fit3)

##   
## Call:  
## lm(formula = lifeExpectancy^2 ~ sqrt(adult\_mortality) + infant\_deaths\_cat +   
## sqrt(alcohol) + measles\_cat + bmi + polio + diphtheria +   
## thin\_avg, data = life\_expectancy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3826.1 -413.2 63.3 443.4 2399.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5047.0004 110.7166 45.585 < 2e-16 \*\*\*  
## sqrt(adult\_mortality) -94.7856 3.4518 -27.460 < 2e-16 \*\*\*  
## infant\_deaths\_catQ2 -95.0049 42.5050 -2.235 0.025509 \*   
## infant\_deaths\_catQ3 -375.4832 46.9358 -8.000 2.00e-15 \*\*\*  
## infant\_deaths\_catQ4 -751.7965 56.8715 -13.219 < 2e-16 \*\*\*  
## sqrt(alcohol) 120.2154 15.7161 7.649 3.01e-14 \*\*\*  
## measles\_catQ2 48.0180 45.1182 1.064 0.287323   
## measles\_catQ3 148.8659 41.9907 3.545 0.000401 \*\*\*  
## measles\_catQ4 237.6588 49.3034 4.820 1.53e-06 \*\*\*  
## bmi 9.4118 0.9423 9.989 < 2e-16 \*\*\*  
## polio 2.9698 0.8792 3.378 0.000743 \*\*\*  
## diphtheria 6.0799 0.8786 6.920 5.93e-12 \*\*\*  
## thin\_avg -30.3288 4.6180 -6.567 6.38e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 707.6 on 2182 degrees of freedom  
## Multiple R-squared: 0.6127, Adjusted R-squared: 0.6105   
## F-statistic: 287.6 on 12 and 2182 DF, p-value: < 2.2e-16

From the regression analysis, all variables are significant with a 0.001 significance level. The (adjusted) R-squared is 61.05% which implies that this model explains 61.05% of the variation in the response variable. This is not significantly larger than the R-squared obtained before removing hepatitsB. The residual standard error is 707.6.

### Interpretation of parameters:

NOTE: Due to the complexity and unclarity of the units, we will be interpreting the parameter effects in a qualitative and quantitative manner by focussing on the direction of effect.

The intercept 5047 is the estimated mean life expectancy in years squared, when all the other predictors are 0.

When the square root of the probability of dying between 15 and 60 years per 1000 population increases by 1 square root percent, the life expectancy decreases by 94.79 years squared. Hence, when adult mortality increases, the life expectancy decreases, while holding other predictors constant.

If the number of infant deaths are greater than 0 and less than 3, life expectancy will decrease by 95.0049 years squared, while holding other predictors constant. If the number of infant deaths are greater than 3 and less than 22, life expectancy will decrease by 375.48 years squared, while holding other predictors as constant. If the number of infant deaths are greater than 22, life expectancy will decrease by 751.8 years squared, while holding other predictors constant. In overall, it can be seen as the number of infant deaths increase, the life expectancy will decrease.

As alcohol consumption increases by one square root litre of pure alcohol, the life expectancy increases by 120.22 squared years, while holding other predictors constant. Hence, when alcohol consumption increases, so does life expectancy.

If the number of reported measles cases increases by 0 to 17 cases, the life expectancy will increase by 48.02 squared years, while holding other predictors as constant. If the number of reported measles cases increases by 17 to 360.2 cases, the life expectancy will increase by 148.87 squared years, while holding other predictors as constant. If the number of reported measles cases increases by more than 360.2 cases, the life expectancy will increase by 237.67 squared years, while holding other predictors as constant. In general, it appears that as the number of measles cases increases, the life expectancy increases.

An increase in bmi by 1 kg/m^2 will increase life expectancy by 9.41 squared years, while keeping other predictors constant. Hence, as bmi increases, the life expectancy also increases.

An increase in polio immunization coverage among 1 year old by 1% will increase life expectancy by 2.97 squared years, keeping other predictors constant.

An increase in diphtheria immunization coverage among 1 year old by 1%, will increase life expectancy by 6.08 squared years, keeping other predictors constant.

An increase in the average thinness among 5 to 19 year olds will decrease life expectancy by -30.33 squared years, while keeping other predictors the constant.

The surprising parameters are alcohol and measles because as alcohol comsumption increases and measle cases increases, the life expectancy increases instead of decreasing which is what the researchers expected.

anova(fit3)

## Analysis of Variance Table  
##   
## Response: lifeExpectancy^2  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sqrt(adult\_mortality) 1 1071435008 1071435008 2139.8369 < 2.2e-16 \*\*\*  
## infant\_deaths\_cat 3 385944250 128648083 256.9320 < 2.2e-16 \*\*\*  
## sqrt(alcohol) 1 92742699 92742699 185.2229 < 2.2e-16 \*\*\*  
## measles\_cat 3 5425859 1808620 3.6121 0.01279 \*   
## bmi 1 95030472 95030472 189.7919 < 2.2e-16 \*\*\*  
## polio 1 31727915 31727915 63.3660 2.734e-15 \*\*\*  
## diphtheria 1 24315374 24315374 48.5619 4.222e-12 \*\*\*  
## thin\_avg 1 21596458 21596458 43.1318 6.377e-11 \*\*\*  
## Residuals 2182 1092546455 500709   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The analysis of variance table (ANOVA table) suggests that all variables except the categorical measles variable are significant predictors at the 0.001 level.

Let us check the variance inflation factor to check for correlation among variables:

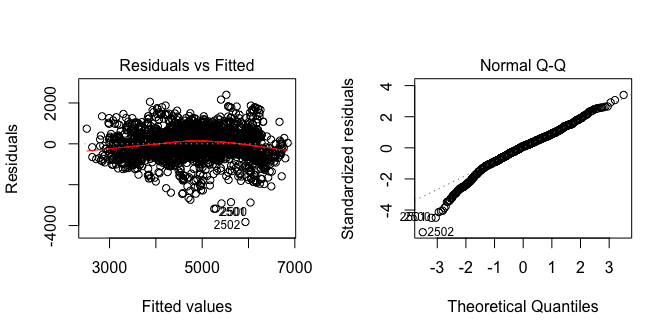
library(car)  
vif(fit3)

## GVIF Df GVIF^(1/(2\*Df))  
## sqrt(adult\_mortality) 1.230588 1 1.109319  
## infant\_deaths\_cat 1.996855 3 1.122168  
## sqrt(alcohol) 1.303162 1 1.141561  
## measles\_cat 1.539060 3 1.074507  
## bmi 1.525217 1 1.234997  
## polio 1.571279 1 1.253507  
## diphtheria 1.536260 1 1.239460  
## thin\_avg 1.700253 1 1.303937

The VIF values of the variables (including the categorical variables) show that none of the variables in the model are correlated with each other, since all values are below 5.

### Residual Analysis

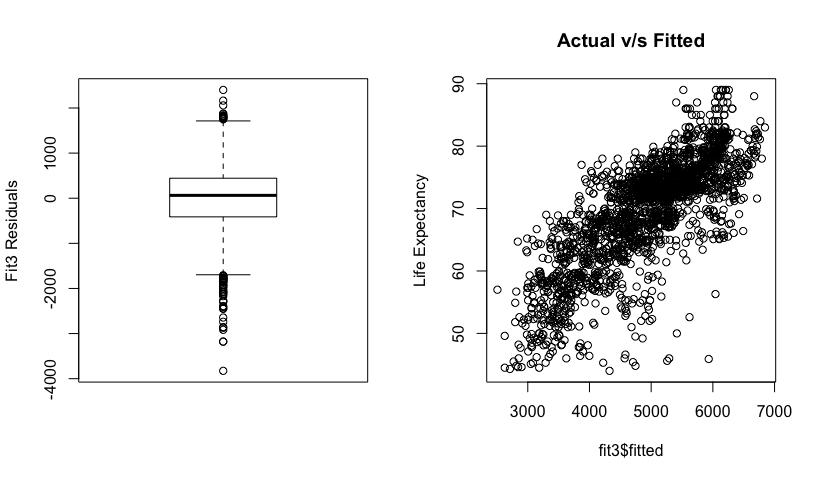
par (mfrow = c(1,2))  
plot (fit3, which=1:2)



The residual v/s fitted value plot is useful for assessing linearity and constant variance conditions of a simple linear model. The residuals are scattered across the plot with high residuals at the bottom of the plot, however, there does not appear to be an extreme non-linear trend in the residuals. Hence, the residuals are consistent with the linearity condition. Further, there does not appear to be a major deviation from constant variance with points being scattered above and below the 0.0 line. The normal quantile plot is a scatterplot that displays the observed data v/s the values that would be expected from a normal sample of the same size. In this quantile plot, there does not appear to be any major deviation from a straight line. While there are no right-skewed residuals, there are some left-skewed residuals but the trend is largely linear indicating that the deviation might be a small variation that can be ignored.

These plots are extremely similar to the plots produced by fit2 indicating that there has not been a large change by the removal of hepatitsB.

par(mfrow = c(1,2))  
boxplot(fit3$residuals, ylab = 'Fit3 Residuals')  
plot(lifeExpectancy~fit3$fitted, main = "Actual v/s Fitted", ylab = "Life Expectancy", data = life\_expectancy)  
abline(0,1, col = "red")



The box plot shows slight skewness at the right end with large skewness on the left end, which was shown in the quantile plot, indicating some deviation from a normal pattern. The Actual v/s Fitted plot shows a similar linear trend as the fit2 model.