W271 - Section 3 - Lab 3

Nicholas Chen, David Skarbrevik, Rama Thamman, Johnny Yeo

December 9, 2016

Research Question of Interest

Is alcohol consumption a valid proxy/indirect indicator of a thriving country (i.e. physically healthy and/or financially healthy)?

While there may be many obvious indicators of physical and financial health (such as rates of certain disease, quality of air, economic stability, etc.), we explore alcohol consumption as a possible novel and unexpected proxy indicator of physical/financial country health.

In media, wine consumption in particular is sometimes cited as having health benefits (http://www.telegraph.co.uk/news/health/11066516/Glassofwinewithdinnerhelpsyoulivelongerscientistsclaim.html). We are in no way positing that our analysis may support the belief that wine (or more generally, alcohol) consumption leads directly to health benefits, instead we are only interested in seeing whether alcohol consumption may be a reasonable novel, indirect indicator of health and/or financial success in a country. Our intuition is that countries with more alcohol consumption are likely more gregarious (more social events) and have more disposable income and that these properties may in turn be indirect indicators of a healthy/wealthy country. At any rate, if our analysis below supports our intuition it would not in any way prove an exact mechanism of action, as this would need to be the topic of future research.
Data sources

Main data source: https://followthedata.wordpress.com/2010/03/15/food-and-health-data-set/

Documentation for this dataset or it's original source could no longer be found. To offset the lack of knowledege/validation about how the data in this dataset was gathered, a more documented dataset was cross-examined to validate our dataset of interest.

Dataset used for validation: http://data.worldbank.org/data-catalog/health-nutrition-and-population-statist

Data from World Bank.

Justification for using this dataset (pros and cons)

The most interesting thing about our chosen dataset is the sheer volume of variables. There are many unusual/interesting variables and data is apparently complete for 86 countries (although bottom-coding may have occured in certain instances). There are many variables relating to health such as obesity occurance, mean cholesterol, and mean blood pressure. This makes analysis via linear model building a natural choice. More specifically, we are interested in comparing alcohol/wine consumption to life expectancy (proxy for physical health) and GNP (proxy for financial health). All of these variables are captured in this dataset. Unfortunately, as previously stated, the method for how this data was gathered is not documented and although we try to validate using more established datasets, this certainly hurts the real-world applicability of our analysis.

Exploratory Data Analysis

Univariate Analysis

```
# load packages
library(rJava)
library(xlsxjars)
library(xlsx)
library(ggplot2)
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(lmtest)
library(car)
library(sandwich)

setwd("~/Desktop/W271_Lab3/Data")

# load data of interest
diet.data <- read.csv("diet-forcsv - Sheet 1.csv")

# load data for validation purposes
data.validation.country.mapping <- read.xlsx("Data from Third Parties for Validation.xlsx", sheetName =
data.validation.life_expect <- read.xlsx("Data from Third Parties for Validation.xlsx", sheetName = "Li
data.validation.growth_rate <- read.xlsx("Data from Third Parties for Validation.xlsx", sheetName = "Po



#Missing values check
na.check = sapply(diet.data, function(x) sum(is.na(x))) # check specifically for NA values
if(sum(na.check) == 0)
{
        cat("No NA values in this data.")
} else {
na.check
cat("There are a total of", sum(na.check), "NAs in the data.")
}
```

```
## No NA values in this data.
```

```
cat("Number of rows: ",nrow(diet.data))
```

```
## Number of rows:  86
```

```r
cat("Number of complete cases: ",nrow(diet.data[complete.cases(diet.data),]))
```

```
## Number of complete cases:  86
```

```r
#There are no missing values
```

Independent Variables

Wine consumption - Calories consumed by wine per person per day (120 Calories = about 1 glass of wine)

```r
#Summary statistics for variables of interest
summary(diet.data$Wine..kcal.day.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    2.00   15.38   18.00  100.00
```

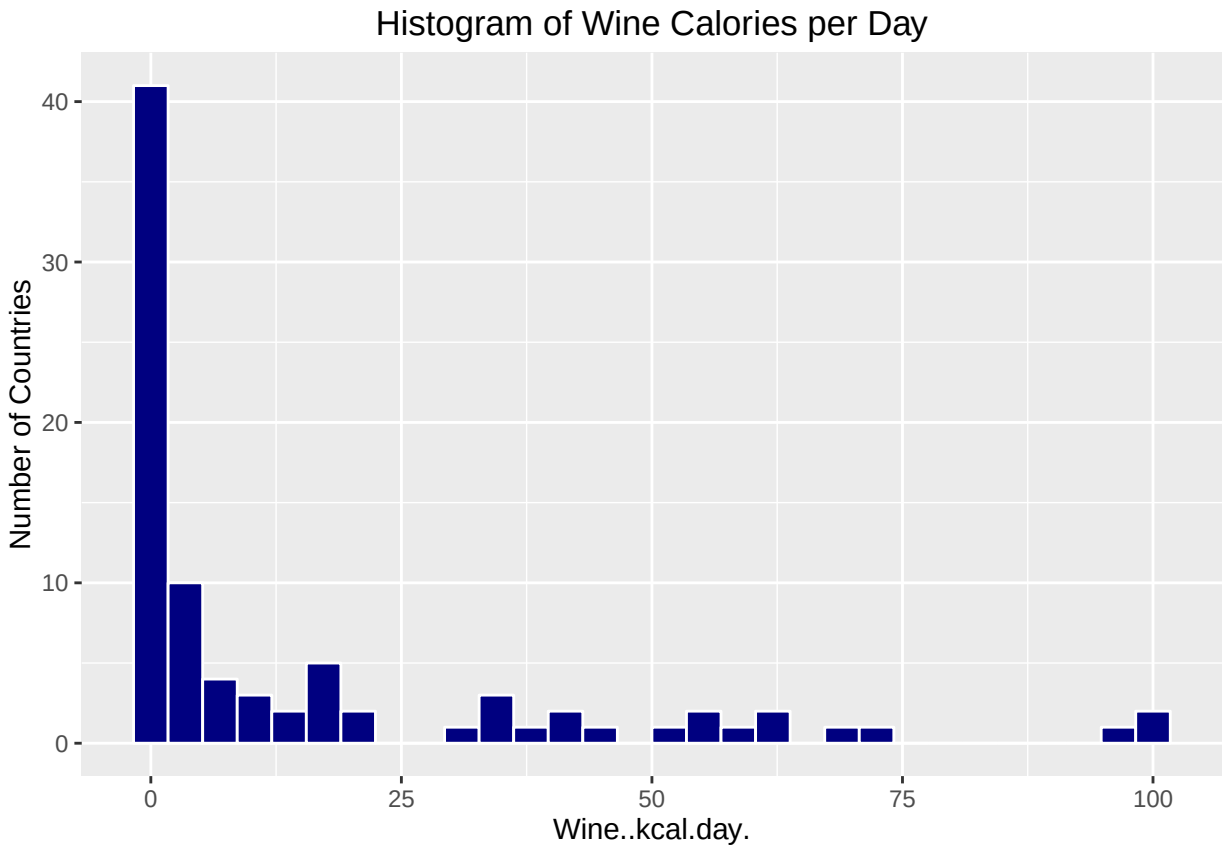
```r
sum(diet.data$Wine..kcal.day. == 0)
```

```
## [1] 32
```

There are 32 countries with zero wine consumption. This could be because of bottom coding to cover for null values.

```r
wine.hist <- ggplot(data = diet.data, aes(x = Wine..kcal.day.))
wine.hist + geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of Wine Calories per Da
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Alcohol consumption - Calories consumed per person per day

```
#Alcoholic beverages calories per day
summary(diet.data$Alcoholic.Beverages..kcal.day.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   22.00   68.50   88.79  146.80  285.00
```

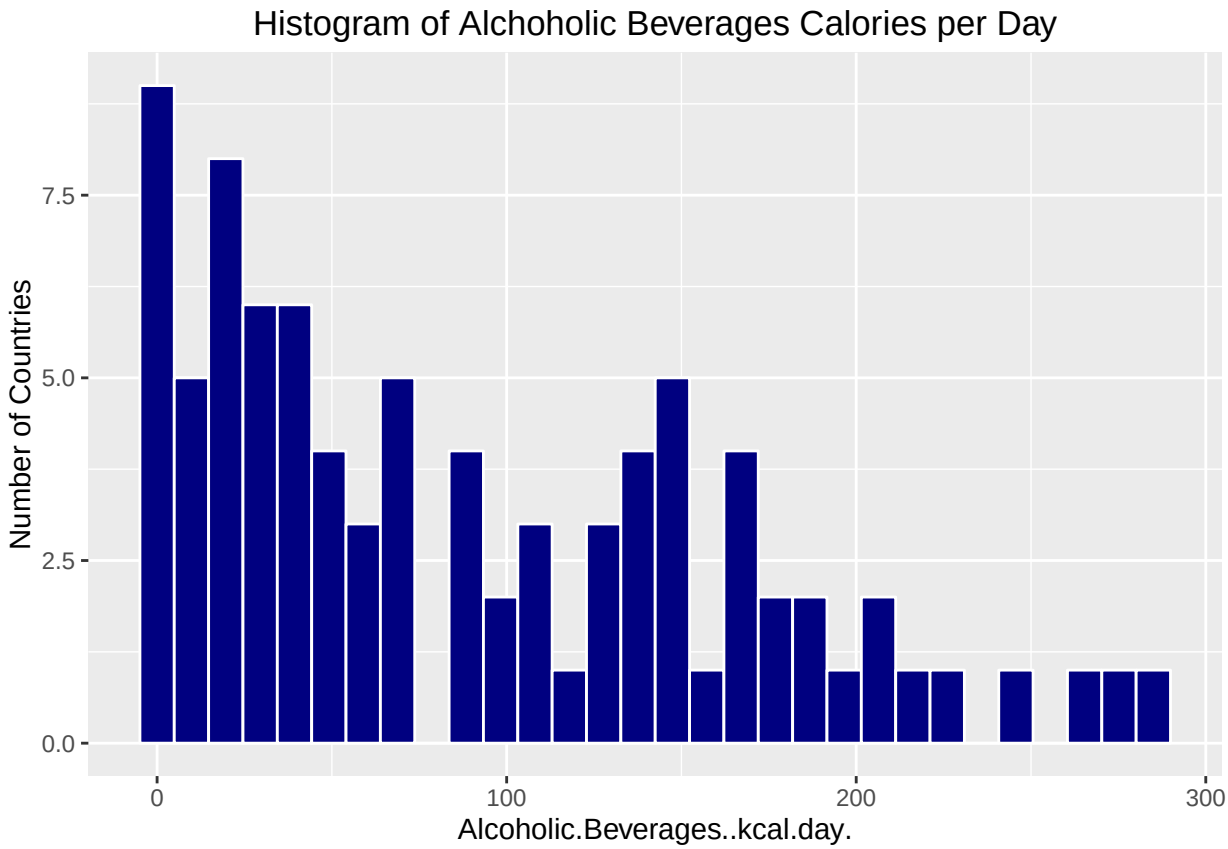
```
sum(diet.data$Alcoholic.Beverages..kcal.day. == 0)
```

```
## [1] 5
```

Like wine, there are a lot of countries with zero or very little consumption of alcoholic beverages.

```
Alcoholic.bevs.cals.hist <- ggplot(data = diet.data, aes(x = Alcoholic.Beverages..kcal.day.))
Alcoholic.bevs.cals.hist + geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of Alcho
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#life expectancy at birth in years (for both male/female)
summary(diet.data$Life.expectancy.at.birth..years..both.sexes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   38.50   67.00   71.50   69.58   77.88   82.00
```

Systolic blood pressure (mean), adults aged 15 and above, men (mmHg)

```
# Justification for using only male blood pressure (highly correlated)
cor(diet.data$Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.,diet.data$Systolic.blood.pr
```

```
## [1] 0.7598489
```

```
summary(diet.data$Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.)
```

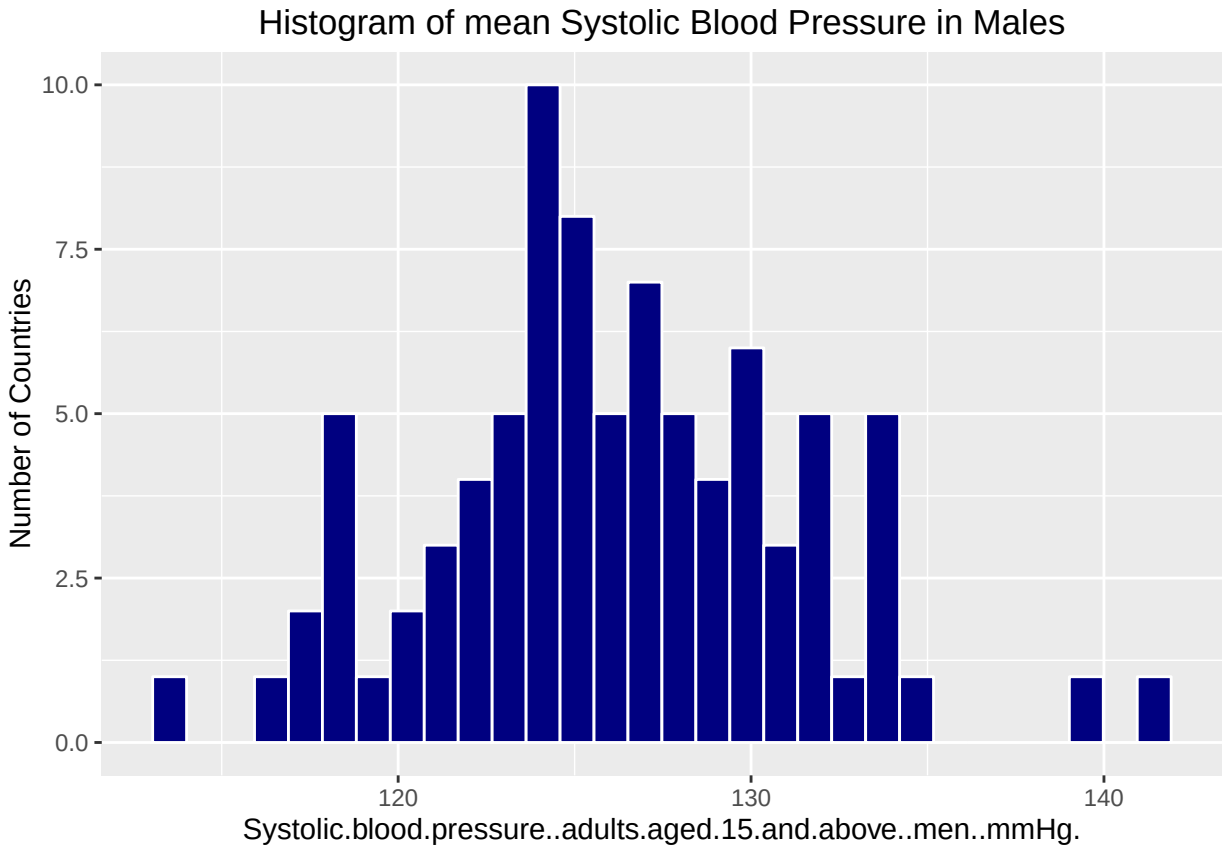
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   113.9   123.3   125.9   126.2   129.5   141.8
```

```
sum(diet.data$Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. == 0)
```

```
## [1] 0
```

```
BloodPressure.mean.hist <- ggplot(data = diet.data, aes(x = Systolic.blood.pressure..adults.aged.15.and
BloodPressure.mean.hist + geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of mean S
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

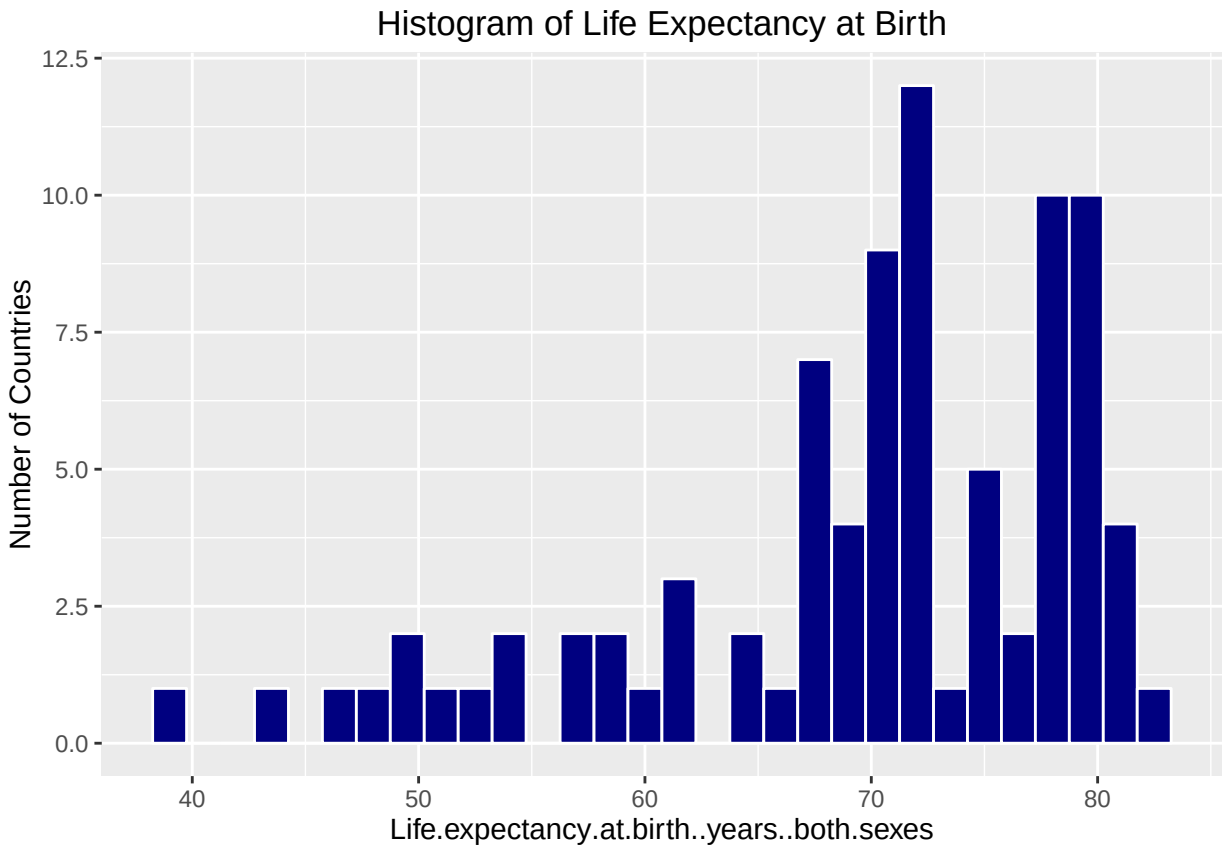


Dependent/Outcome Variables

Life expectancy in years (from birth, both male/female)

```
life.expect.all.hist <- ggplot(data = diet.data, aes(x = Life.expectancy.at.birth..years..both.sexes))
life.expect.all.hist + geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of Life Expe
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The life expectancy variable shows a negative skew (because no one lives to be 160).

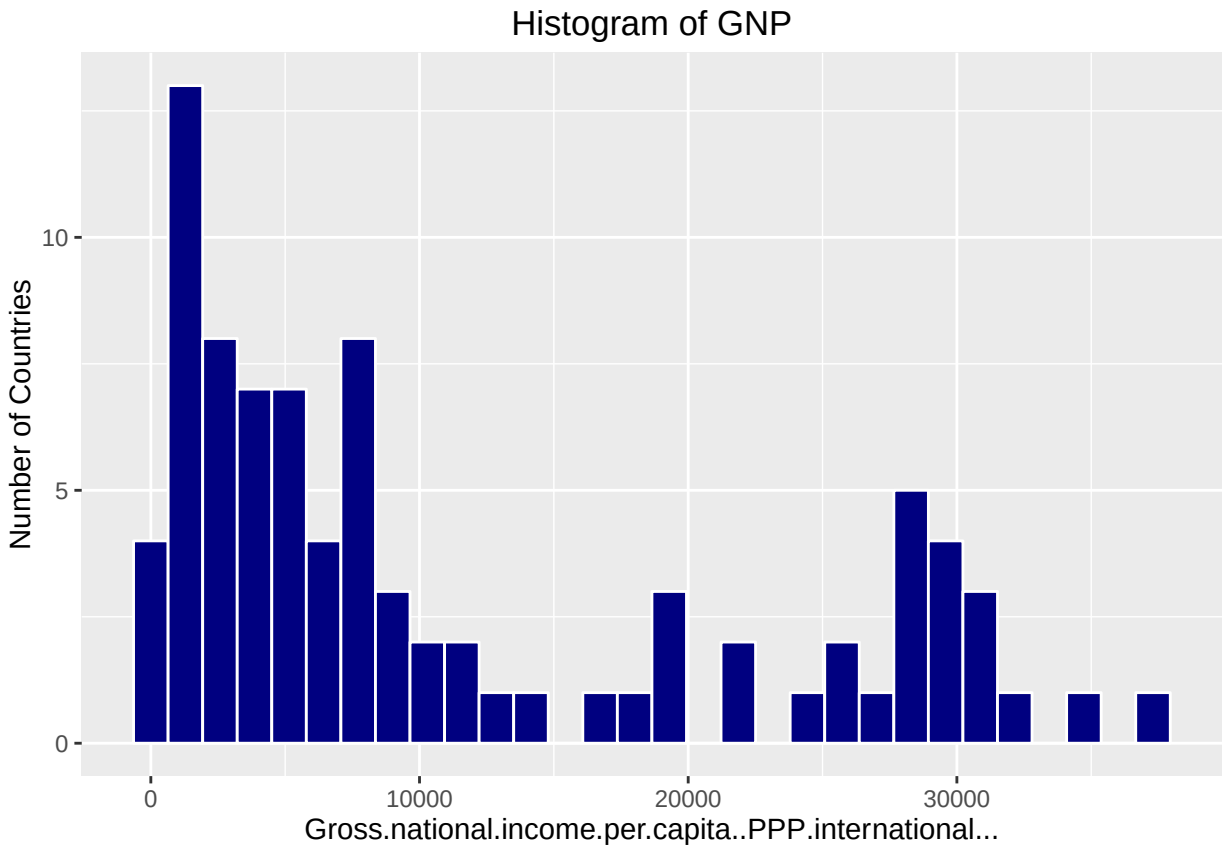
Gross National Product (GNP)

```
#GNP per capita
summary(diet.data$Gross.national.income.per.capita..PPP.international...)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     240    2852    6890   11390   19220   37540
```

```
#GNP histogram
GNP.hist <- ggplot(data = diet.data, aes(x = Gross.national.income.per.capita..PPP.international...))
GNP.hist + geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of GNP") + labs(y = "Nu
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



GNP also shows skew (this is normally expected for GNP).

Multivariate Analysis

```
#Correlations
#correlation between wine consumption and alcoholic beverage consumption per day and life expectancy at
cor(diet.data$Wine..kcal.day., diet.data$Life.expectancy.at.birth..years..both.sexes)
```

```
## [1] 0.4963627
```

```
cor(diet.data$Alcoholic.Beverages..kcal.day., diet.data$Life.expectancy.at.birth..years..both.sexes)
```

```
## [1] 0.5054063
```

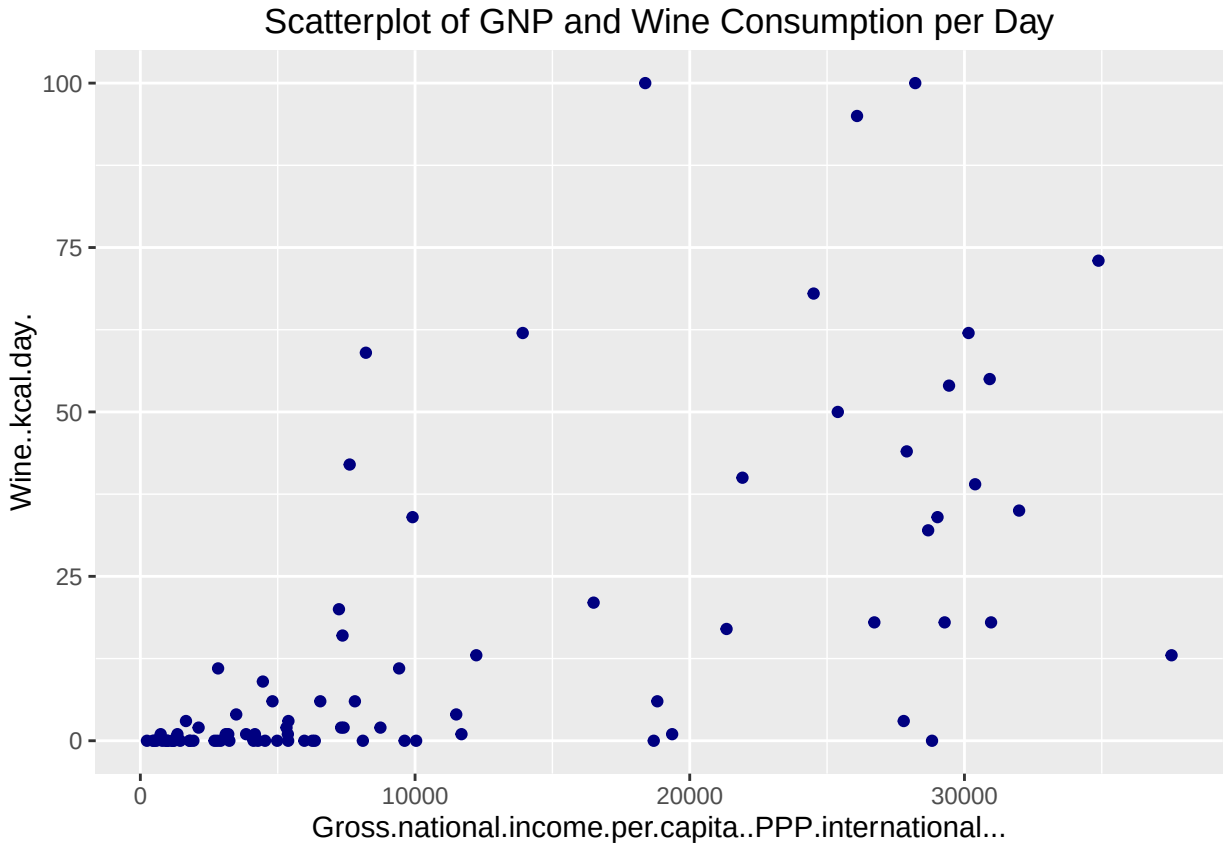
```
#look at correlation between wine / alcohol consumption and GNP to see if the above result appears to b
#There are very high correlations between wine and alcohol consumption with GNP, both being above 0.6.
cor(diet.data$Gross.national.income.per.capita..PPP.international..., diet.data$Wine..kcal.day.)
```

```
## [1] 0.6433328
```

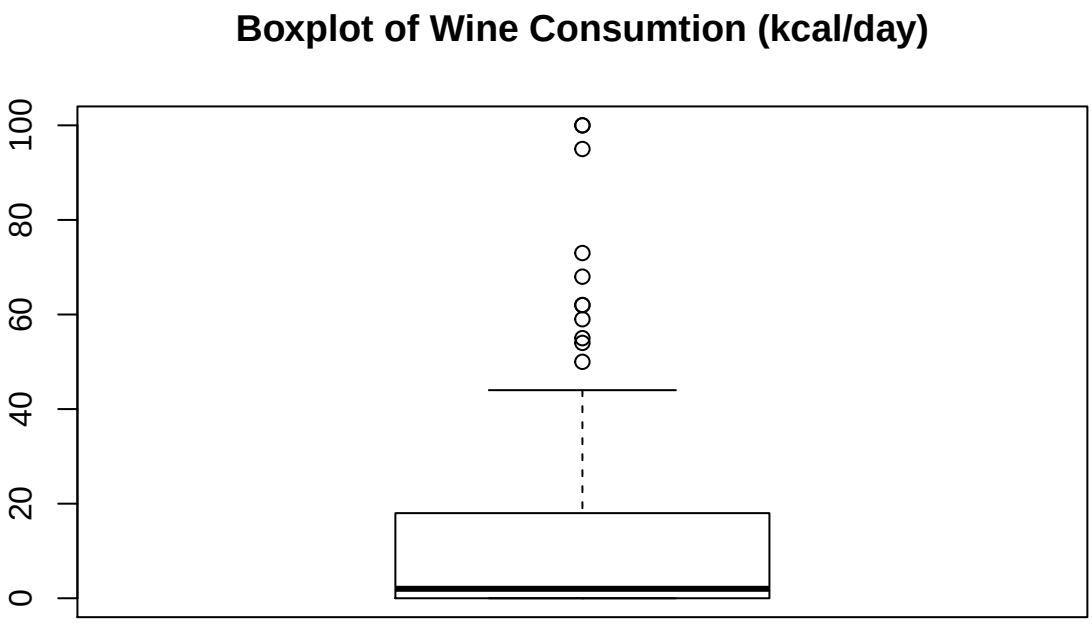
```
cor(diet.data$Gross.national.income.per.capita..PPP.international..., diet.data$Alcoholic.Beverages..kca
```

```
## [1] 0.6576478
```

```
#diet.data$Alcoholic.Beverages..kcal.day. > 0
wine.gnp.scatter <- ggplot(data = diet.data, aes(x = Gross.national.income.per.capita..PPP.international
wine.gnp.scatter + geom_point(colour = "navy") + ggtitle("Scatterplot of GNP and Wine Consumption per Da
```




```
#further analysis of correlation between wine / alcohol consumption and life expectancy at birth
i = 52
wine.box <- boxplot(diet.data[i], main = "Boxplot of Wine Consumtion (kcal/day)")
```




```
df <- cbind(diet.data[i], diet.data$Countries, diet.data$Life.expectancy.at.birth..years..both.sexes)
names(df) <- c("wine_consumption", "countries", "life_expectancy")
ordered_df <- df[order(df[1]),]
ordered_df[ordered_df$wine_consumption > wine.box$stats[5],]
```

```
##    wine_consumption   countries life_expectancy
## 33               50      Greece            79.0
## 7                54     Belgium            78.5
## 4                55     Austria            79.0
## 2                59   Argentina            74.5
## 20               62     Denmark            78.0
## 36               62     Hungary            72.5
## 73               68       Spain            80.0
## 76               73 Switzerland            81.0
## 42               95       Italy            80.0
## 28              100      France            80.0
## 64              100    Portugal            78.0
```

```
#Given the boxplot, these are the countries with "outlier-level" wine consumption, and their life expec
#Every country with high wine consumption has a life expectancy of over 70.
#It is important to also notice, however, that all of these countries (minus Argentina) are a part of E
#where wine consumption is on average higher than the rest of the world.
#Given these results, despite the high correlation, it's hard to tell whether we see any good indicatio
```

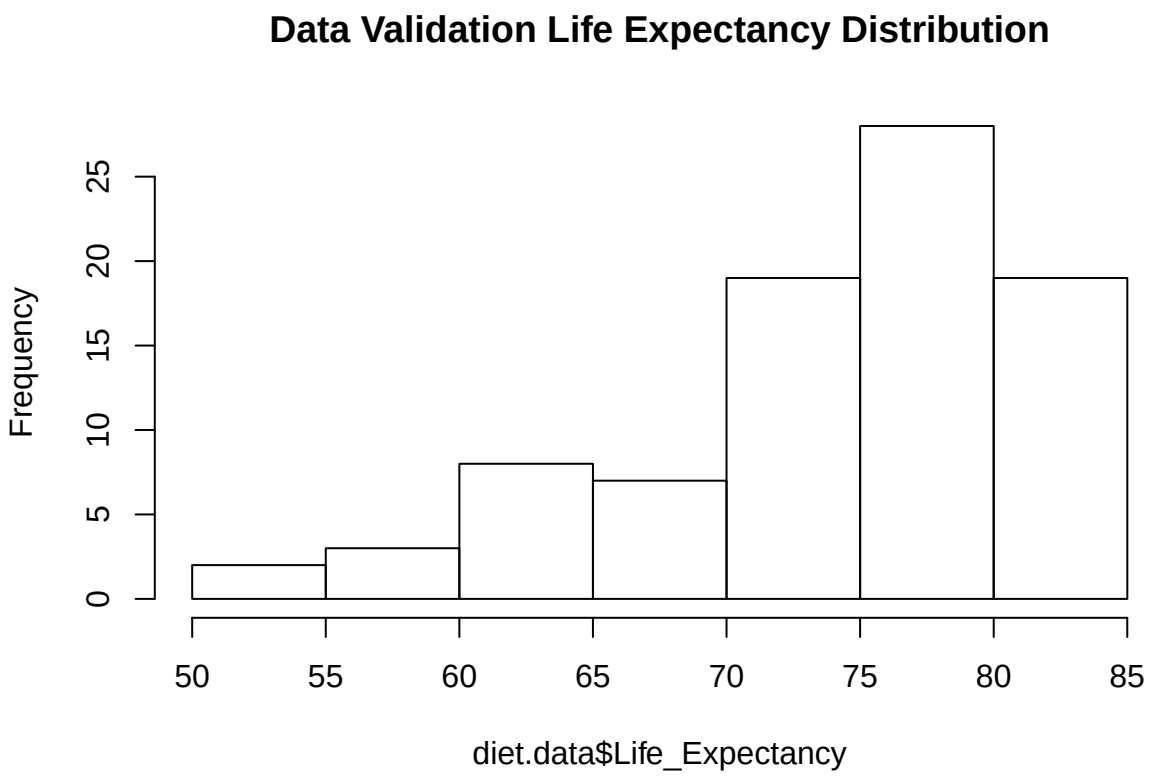
Data Validation

```
#Merge country code with validation datasets.
data.validation.growth_rate <- merge(data.validation.growth_rate, data.validation.country.mapping[,c("P
```

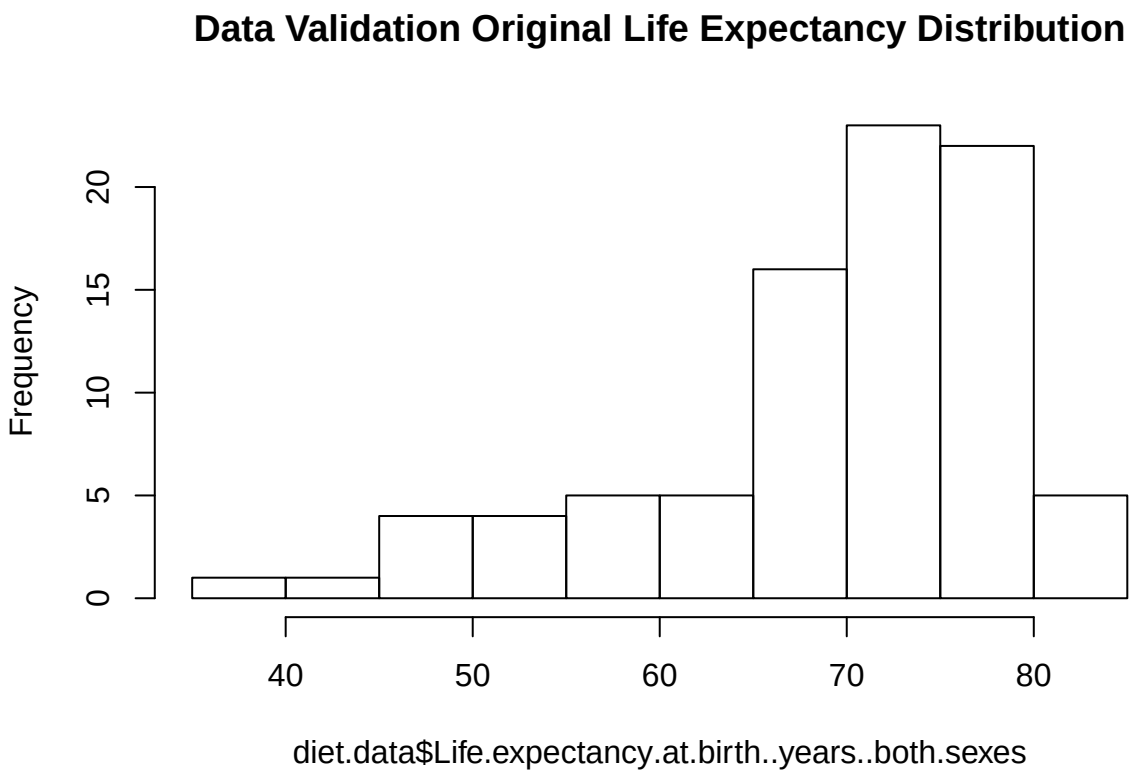
```r
data.validation.life_expect <- merge(data.validation.life_expect, data.validation.country.mapping[,c("C

#Merge validating data into the main country dataset.
diet.data <- merge(diet.data, data.validation.growth_rate[,c("Country_main_dataset", "Growth_rate_2000"
diet.data <- merge(diet.data, data.validation.life_expect[,c("Country_main_dataset", "Life_Expectancy")]

#Now compare data validation sources to main dataset
#Life expectancy
diet.data$Life_Expectancy_pct_diff <- (diet.data$Life.expectancy.at.birth..years..both.sexes - diet.data

summary(diet.data$Life_Expectancy_pct_diff)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.511700 -0.087700 -0.047750 -0.073340 -0.028300 -0.002548
```

```r
hist(diet.data$Life_Expectancy, main = "Data Validation Life Expectancy Distribution")
```




```r
hist(diet.data$Life.expectancy.at.birth..years..both.sexes, main = "Data Validation Original Life Expec
```




```
hist(diet.data$Life_Expectancy_pct_diff, main = "Percent Difference Life Expectancy")
```

Life expectancy in the original dataset appears to be systematically lower than the 2016 life expectancies downloaded from the CIA factbook.

This makes sense given that we believe the life expectancy in the original data to be from an earlier period, likely 2000 - 2005 based on the other variables, and that we expect life expectancy to increase over time.

```
Growth.rate.examination <- diet.data[,c("Countries", "Growth_rate_2000", "Growth_rate_2005", "Growth_ra

Growth.rate.examination$Growth_rate_pct_diff_2000 <- (Growth.rate.examination$Population.annual.growth.
Growth.rate.examination$Growth_rate_pct_diff_2005 <- (Growth.rate.examination$Population.annual.growth.
Growth.rate.examination$Growth_rate_pct_diff_2010 <- (Growth.rate.examination$Population.annual.growth.

#Summary statistics of each growth rate
summary(Growth.rate.examination$Population.annual.growth.rate....)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.100   0.500   1.050   1.151   1.775   4.800
```

```
summary(Growth.rate.examination$Growth_rate_2000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.0440  0.4296  1.1660  1.2580  1.9680  5.5940
```

```
summary(Growth.rate.examination$Growth_rate_2005)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.3040  0.4397  1.1470  1.2400  1.6810 11.9800
```

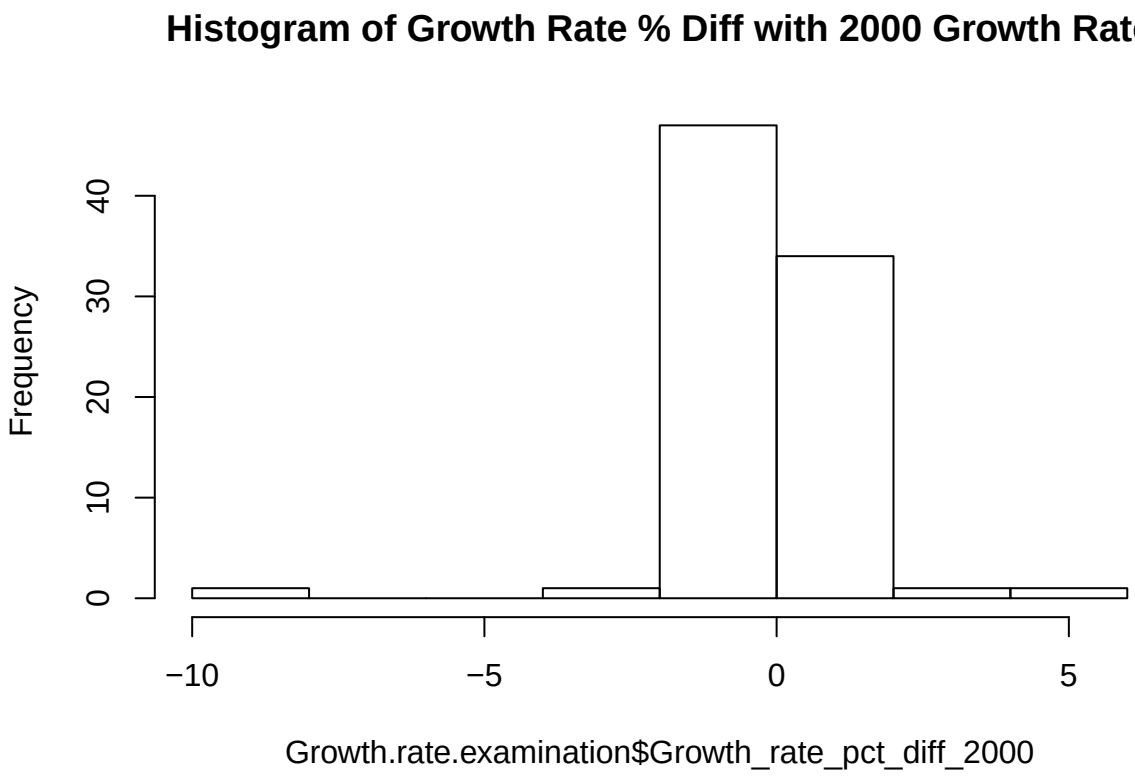
```
summary(Growth.rate.examination$Growth_rate_2010)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.3160  0.3746  1.0960  1.1720  1.6160  7.7870
```

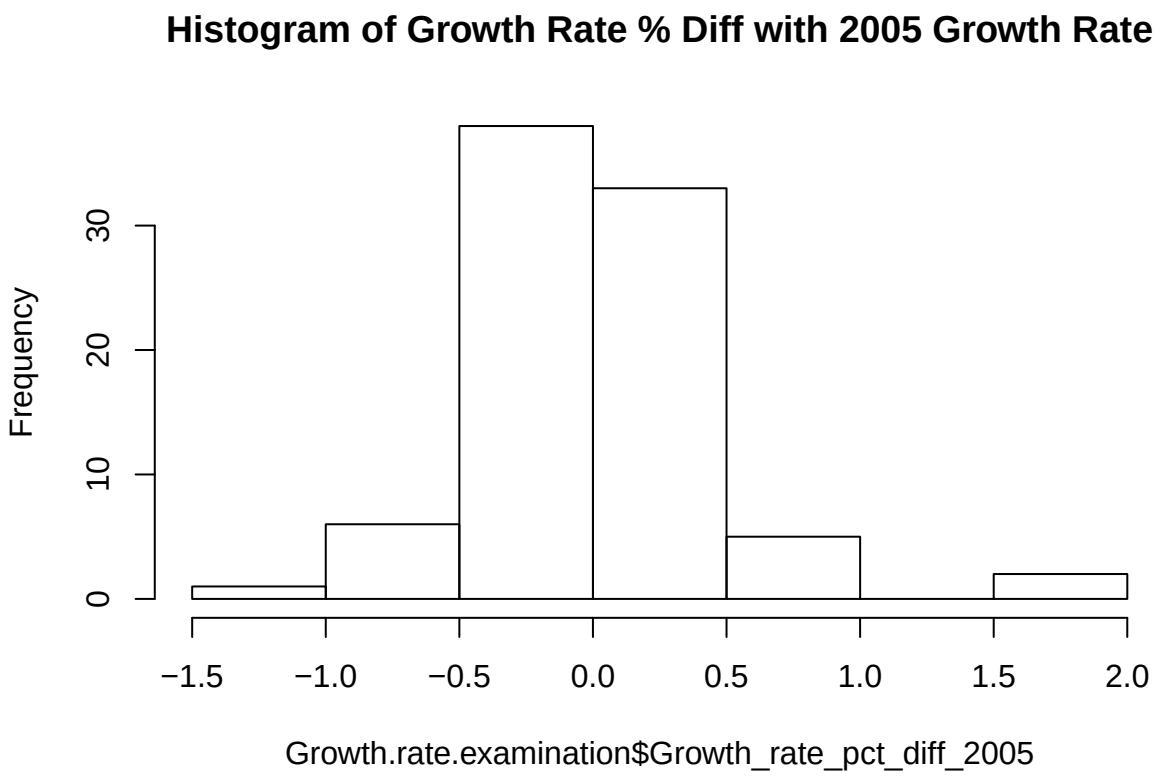
```
#Histograms of percent difference with each known year growth rate
summary(Growth.rate.examination$Growth_rate_pct_diff_2000)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -9.44300 -0.15450 -0.02783      Inf  0.14770      Inf
```

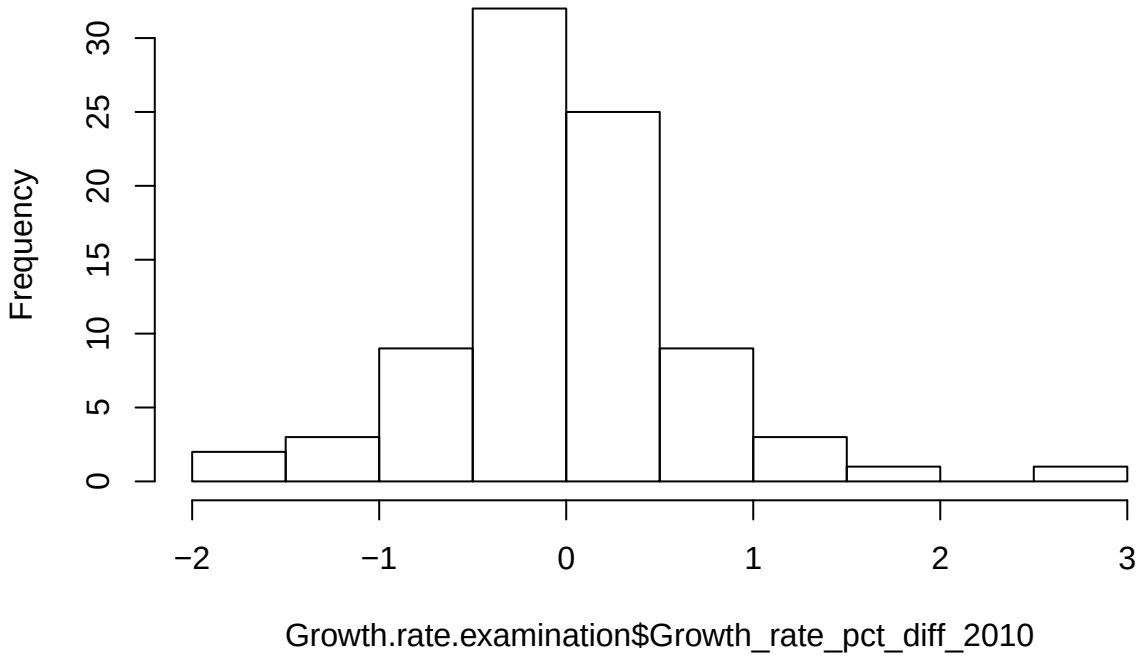
```
hist(Growth.rate.examination$Growth_rate_pct_diff_2000, main = "Histogram of Growth Rate % Diff with 20
```



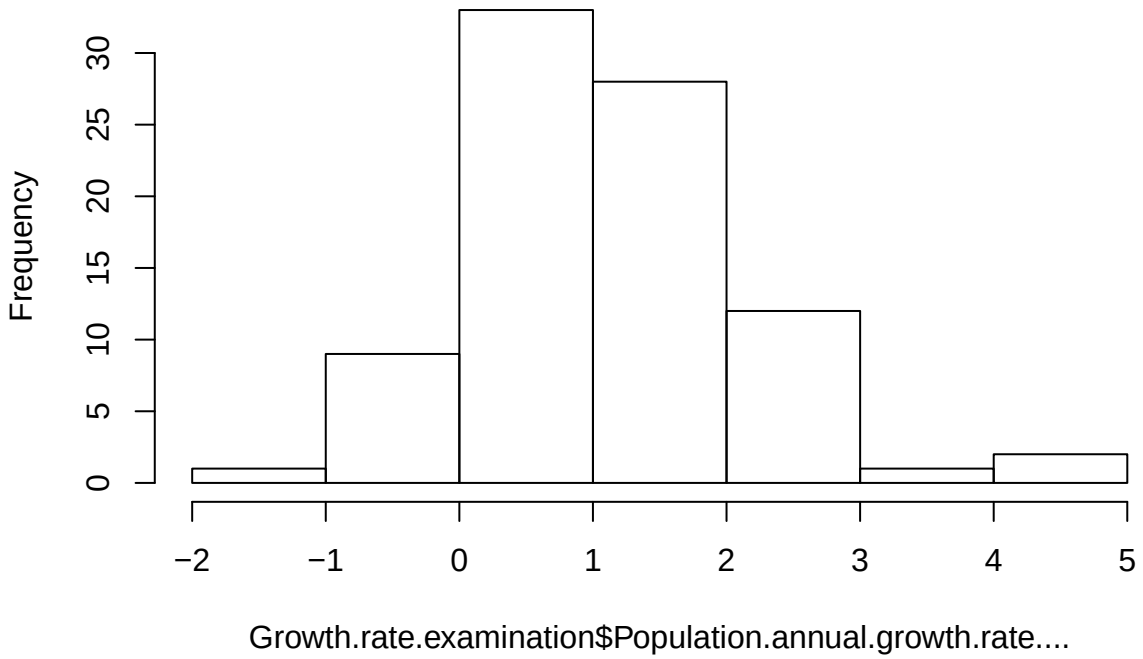

```
hist(Growth.rate.examination$Growth_rate_pct_diff_2005, main = "Histogram of Growth Rate % Diff with 20
```



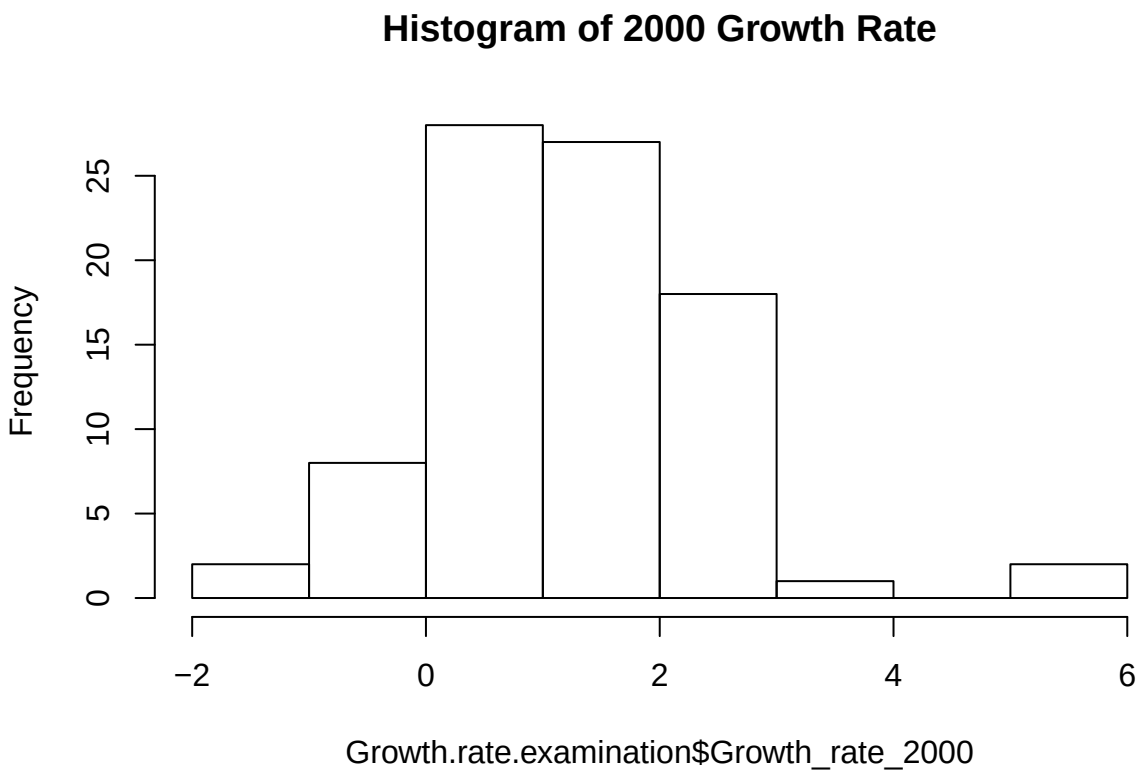
```r
hist(Growth.rate.examination$Growth_rate_pct_diff_2010, main = "Histogram of Growth Rate % Diff with 201
```



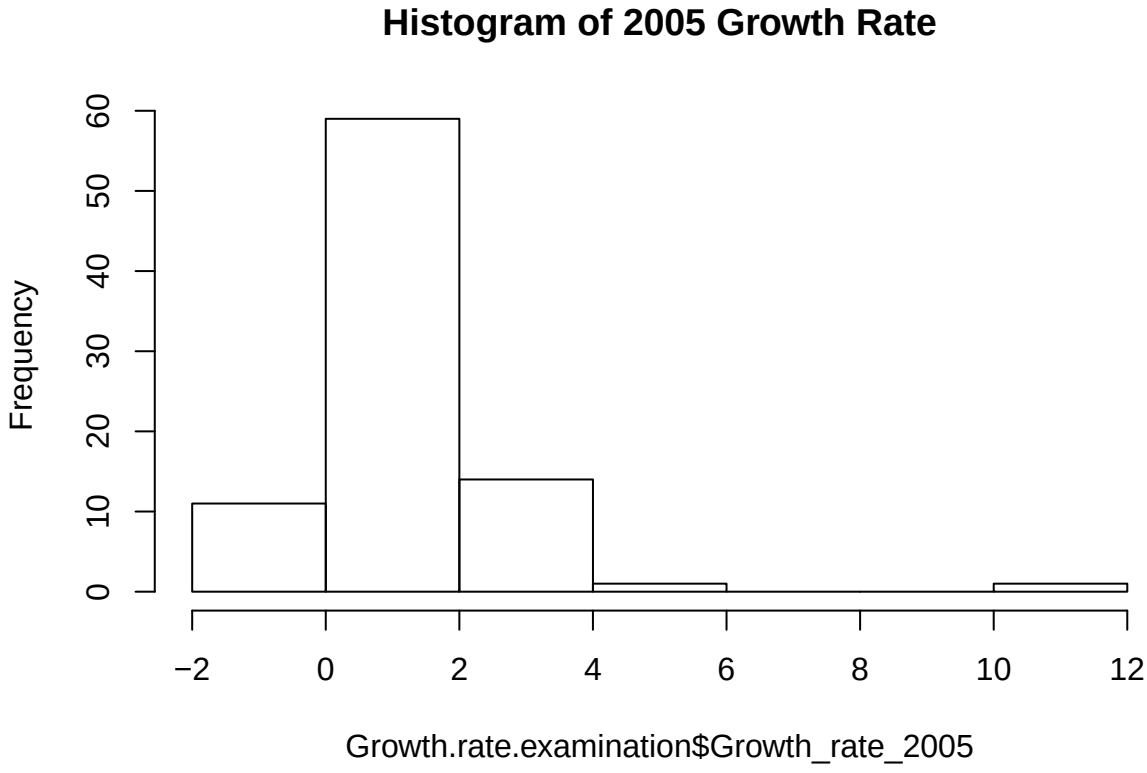

```r
#Histograms of each growth rate
hist(Growth.rate.examination$Population.annual.growth.rate...., main = "Histogram of Original Growth Ra
```



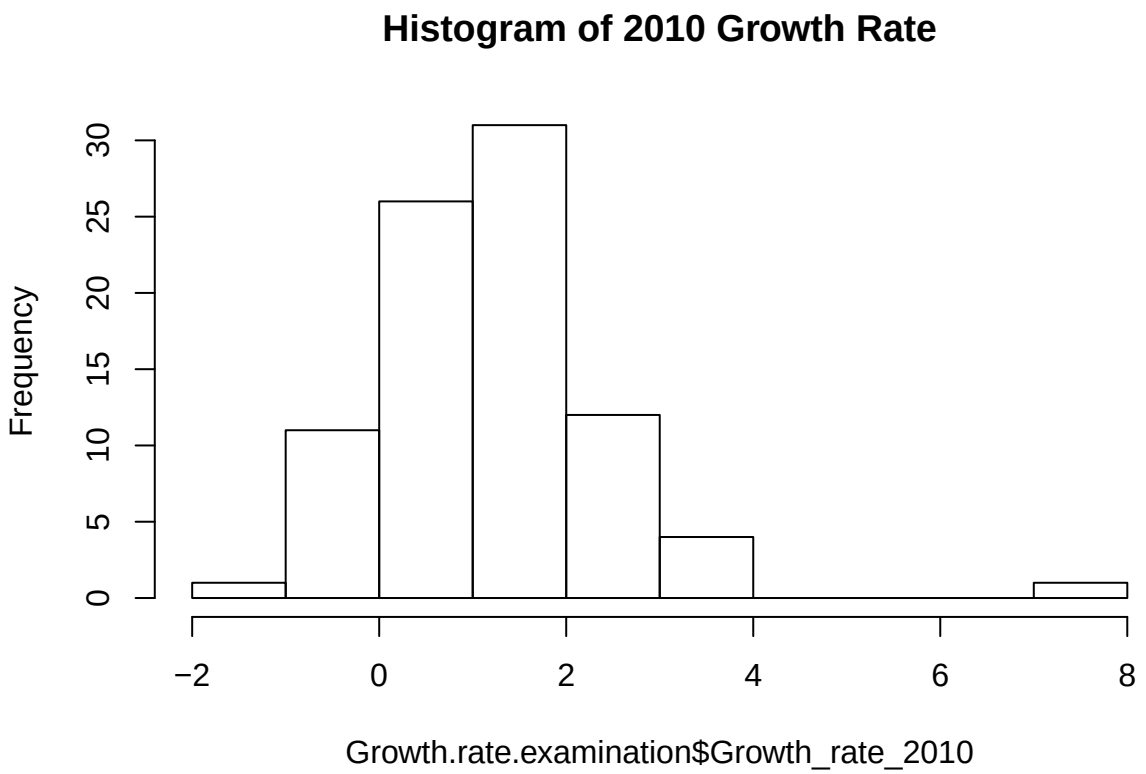
```r
hist(Growth.rate.examination$Growth_rate_2000, main = "Histogram of 2000 Growth Rate")
```




```r
hist(Growth.rate.examination$Growth_rate_2005, main = "Histogram of 2005 Growth Rate")
```



```r
hist(Growth.rate.examination$Growth_rate_2010, main = "Histogram of 2010 Growth Rate")
```




```r
#Correlation between main dataset growth rate and year 2000 growth rate
cor(Growth.rate.examination$Population.annual.growth.rate...., Growth.rate.examination$Growth_rate_2010)
```

```
## [1] 0.8620627
```

The population growth rate distribution from the original dataset looks the most similar to the 2000 population growth rate.

This makes sense and is a good sign of data validation given that other variables appear to be measures of this time period.

Linear Model Building

Null Hypotheses: Alcohol/wine consumption has no impact on GNP or life expectancy

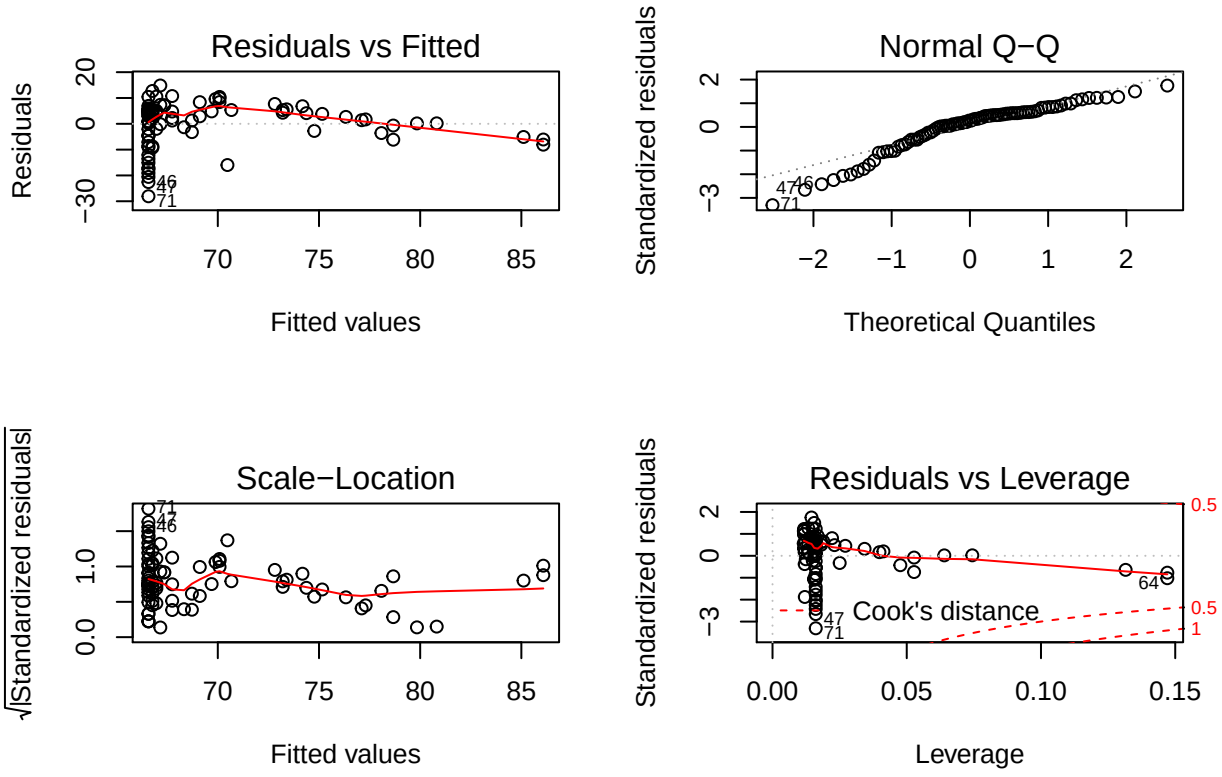
Model 1 - parsimonious model - life expectancy ~ wine

```r
wine.model.1 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Wine..kcal.day., data = diet.data)
summary(wine.model.1)
```

```
##
```

```
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ Wine..kcal.day.,
##     data = diet.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.074  -4.327   2.243   5.182  14.841
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     66.57425    1.08633   61.28  < 2e-16 ***
## Wine..kcal.day.  0.19510    0.03723    5.24 1.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.56 on 84 degrees of freedom
## Multiple R-squared:  0.2464, Adjusted R-squared:  0.2374
## F-statistic: 27.46 on 1 and 84 DF,  p-value: 1.172e-06
```

```
par(mfrow=c(2,2))
plot(wine.model.1)
```




```
bptest(wine.model.1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  wine.model.1
## BP = 3.5231, df = 1, p-value = 0.06052
```

```r
durbinWatsonTest(wine.model.1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2439961      1.511177    0.02
##  Alternative hypothesis: rho != 0
```

```r
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(wine.model.1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     66.574249   1.197593  55.590 < 2.2e-16 ***
## Wine..kcal.day.  0.195098   0.030565   6.383 9.092e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first model shows that wine consumption at the country level has quite a strong relationship with life expectancy.

The coefficient estimate for wine is .195 which is statistically significant at the $p < .001$ level. The statistical significance of the estimate holds when heteroskedasticity robust standard errors are used.

The wine consumption variable is measured in calories, so the interpretation of this coefficient is that one additional calorie of wine consumption per day across the population is associated with a 0.19 year increase in life expectancy.

A glass of wine has about 120 calories so this coeficcient indicates that on average a population that drinks one additional glass of wine per day is expected to have a life expectancy of about 22.8 years greater, all else equal.

However, this interpretation relies on the assumption that there is a linear relationship between average wine consumption and population life expectancy which may or may not be true.

The diagnostic residuals vs. fitted values plot shows that heteroskedasticity may be a problem. Part of this result is caused by the fact that there are so many countries where average wine consumption is zero.

As a result, we may want to use the generalized alcohol consumption variable that has fewer observations of zero.

The Breusch pagan test confirms that heteroskedasticity of errors is borderline problematic.

The Durbin Watson test also gives a statistically significant result which means we should reject the null hypothesis of the test that the errors are not correlated. This is a bit of a strange result that we may want to look into further.

Our theoretical foundation could also support the use of the generalized alcohol consumption variable as the main independent variable in the model as it may be able to extend our hypothesis to cultures where wine consumption is not common, but instead other alcoholic beverages are consumed at group meals.

Despite the statistically significant coefficient estimate, there is by no means any evidence of any casual relationship between wine consumption and life expectancy at this point.

It is interesting to see that there is a relationship of some sort between the two variables, but this could be just a result of two variables affected caused by a third variable, or simply a phenomena due to chance, or any other reasonable scenario that can be thought up at this point.

Model 1.1 - Sensitivity analysis - Healthy life expectancy ~ wine

This analysis is to test if Healthy life expectancy is a proxy for Life expectancy.

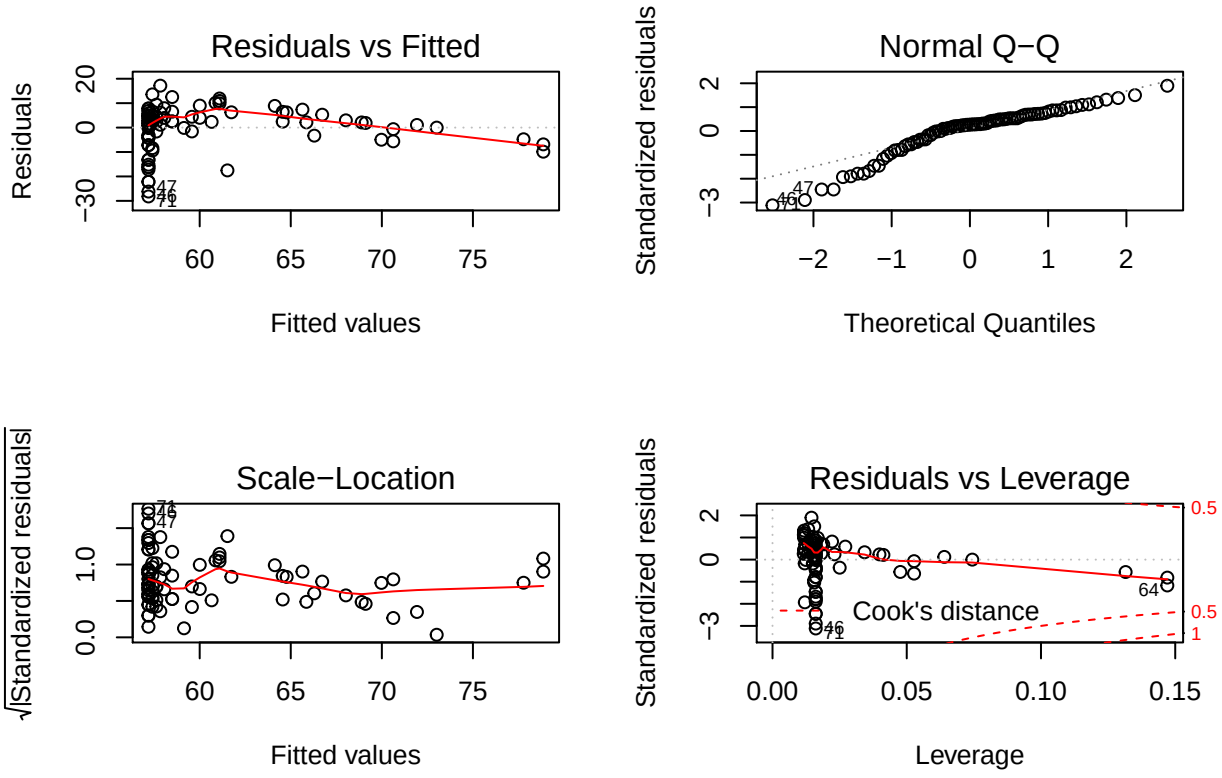
```
#There is a high correlation between Healthy life expectancy and Life expectancy at birth
cor(diet.data$Healthy.life.expectancy..HALE..at.birth..years..both.sexes, diet.data$Life.expectancy.at.
```

```
## [1] 0.9862559
```

```
#Start with a simple linear regression and build up from there comparing models along the way.
wine.model.1.1 <- lm(diet.data$Healthy.life.expectancy..HALE..at.birth..years..both.sexes ~ Wine..kcal.
summary(wine.model.1.1)
```

```
##
## Call:
## lm(formula = diet.data$Healthy.life.expectancy..HALE..at.birth..years..both.sexes ~
##     Wine..kcal.day., data = diet.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.189  -3.965   2.410   5.677  17.161
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     57.18897    1.15745  49.409  < 2e-16 ***
## Wine..kcal.day.  0.21674    0.03967   5.464 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.121 on 84 degrees of freedom
## Multiple R-squared:  0.2622, Adjusted R-squared:  0.2534
## F-statistic: 29.85 on 1 and 84 DF,  p-value: 4.666e-07
```

```
par(mfrow=c(2,2))
plot(wine.model.1.1)
```




```
bptest(wine.model.1.1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  wine.model.1.1
## BP = 3.0372, df = 1, p-value = 0.08138
```

```
durbinWatsonTest(wine.model.1.1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2332932      1.532319   0.028
##  Alternative hypothesis: rho != 0
```

Outcome of the analysis is very similar to Model #1. This validates the data Healty life expectancy and Life expectancy are consistent.

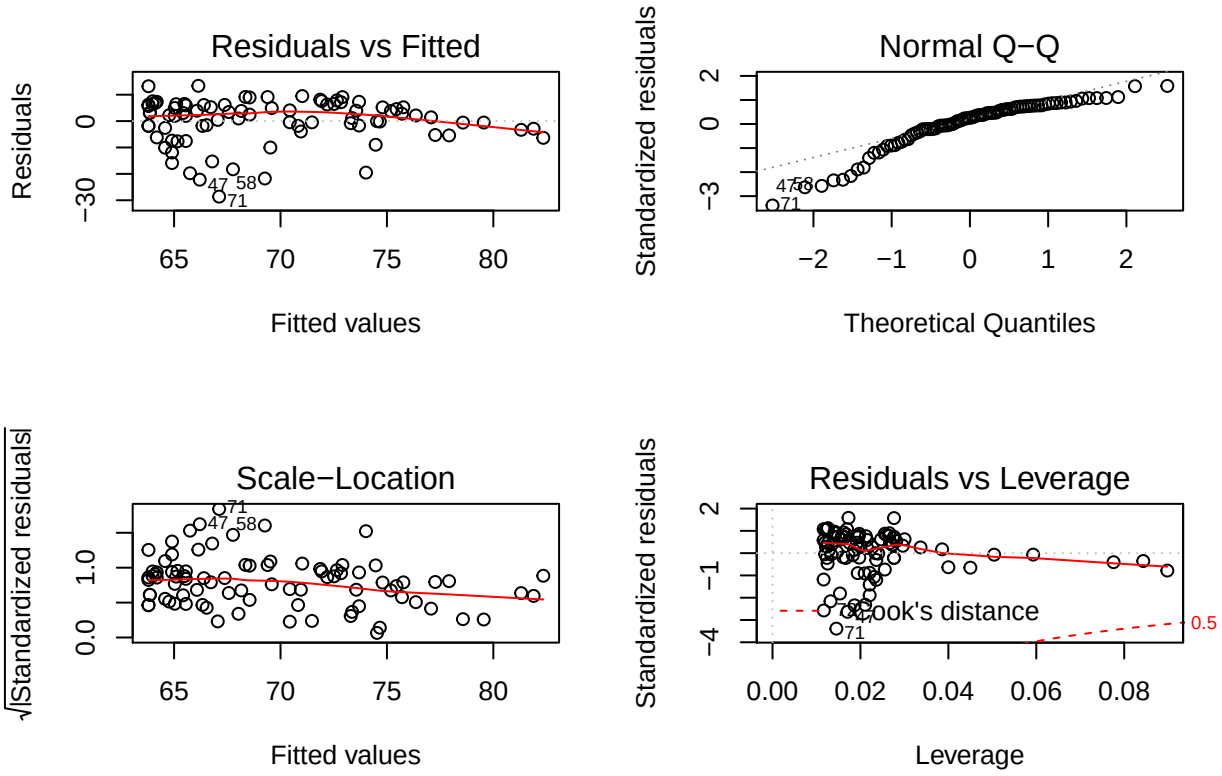
Model 2 - parsimonious model using alcohol consumption- life expectancy ~ alcohol

```
alc.model.1 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Alcoholic.Beverages..kcal.day., data = c
summary(alc.model.1)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ Alcoholic.Beverages..kcal.day.,
```

```
##     data = diet.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.616  -2.813   2.184   6.118  13.361
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    63.79606    1.41455  45.100  < 2e-16 ***
## Alcoholic.Beverages..kcal.day.  0.06509    0.01213   5.368 6.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.509 on 84 degrees of freedom
## Multiple R-squared:  0.2554, Adjusted R-squared:  0.2466
## F-statistic: 28.82 on 1 and 84 DF,  p-value: 6.936e-07
```

```
par(mfrow=c(2,2))
plot(alc.model.1)
```




```
bptest(alc.model.1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  alc.model.1
## BP = 2.308, df = 1, p-value = 0.1287
```

```
durbinWatsonTest(alc.model.1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2947401      1.407349    0.01
##  Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                    63.796064   1.469330  43.419 < 2.2e-16 ***
## Alcoholic.Beverages..kcal.day.  0.065091   0.009508   6.846 1.163e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient estimate for alcohol consumption is .065 indicating that for a country where average daily alcohol consumption across the population is 1 calorie higher is expected to have a higher life expectancy by .065 years, holding all else equal.

This coefficient is statistically significant at $p < .001$ level using heteroskedasticity robust errors.

Again, the diagnostic residuals vs. fitted values plot shows that heteroskedasticity may continue be a problem.

The Breusch-Pagan test however yields a non-statistically significant result which means that we fail to reject the null hypothesis that the variance of the errors is stable across levels of fitted values.

The Durbin-Watson test again shows that the errors are correlated. We should be sure to keep an eye on this after adding controls to the model.

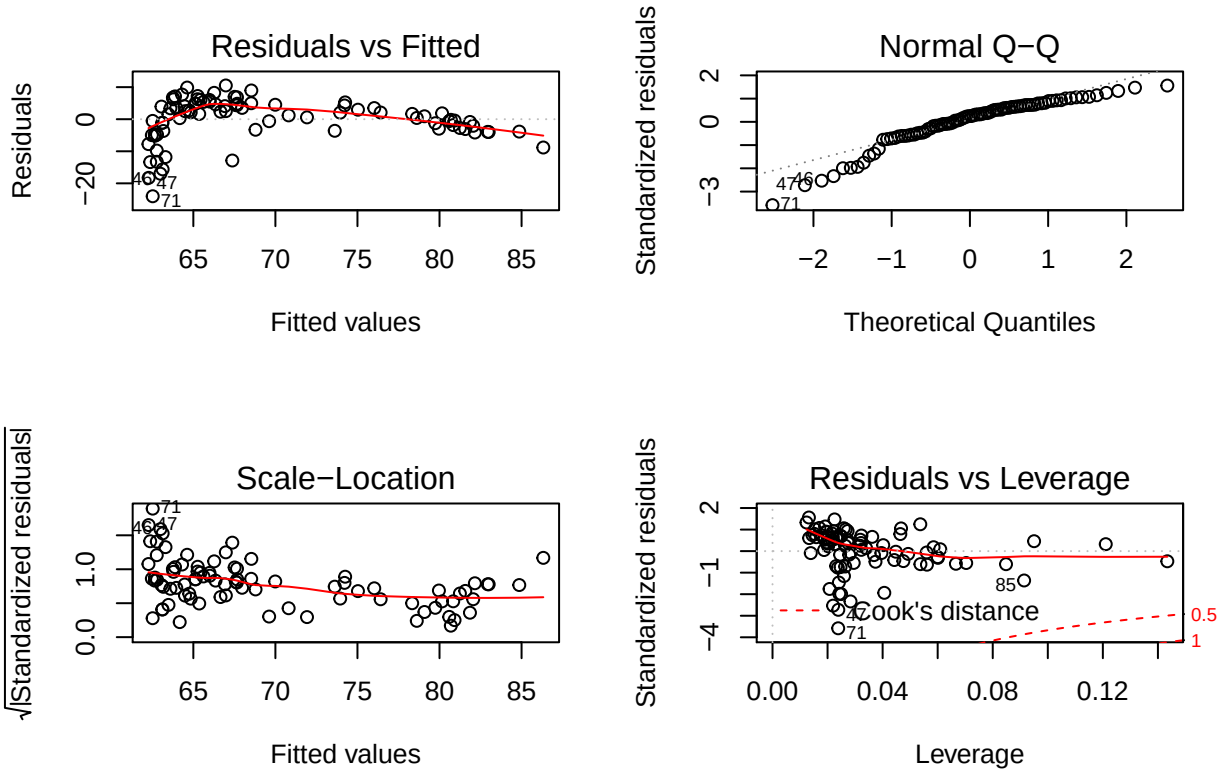
Model 3 - alcohol consumption with control for GNP - life expectancy ~ alcohol + GNP

```
alc.model.2 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Alcoholic.Beverages..kcal.day. + Gross.
summary(alc.model.2)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ Alcoholic.Beverages..kcal.day. +
##     Gross.national.income.per.capita..PPP.international..., data = diet.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.025  -3.235   1.654   4.613  10.521
##
## Coefficients:
##                                                    Estimate
## (Intercept)                                       6.191e+01
## Alcoholic.Beverages..kcal.day.                    5.980e-03
## Gross.national.income.per.capita..PPP.international... 6.266e-04
##                                                   Std. Error t value
```

```
## (Intercept)                                                1.160e+00  53.357
## Alcoholic.Beverages..kcal.day.                             1.284e-02   0.466
## Gross.national.income.per.capita..PPP.international...     8.950e-05   7.000
##                                                           Pr(>|t|)
## (Intercept)                                                < 2e-16 ***
## Alcoholic.Beverages..kcal.day.                               0.643
## Gross.national.income.per.capita..PPP.international...     6.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.787 on 83 degrees of freedom
## Multiple R-squared:  0.5318, Adjusted R-squared:  0.5206
## F-statistic: 47.15 on 2 and 83 DF,  p-value: 2.095e-14
```

```
par(mfrow=c(2,2))
plot(alc.model.2)
```




```
bptest(alc.model.2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  alc.model.2
## BP = 9.4047, df = 2, p-value = 0.009074
```

```
durbinWatsonTest(alc.model.2)
```

```
##  lag Autocorrelation D-W Statistic p-value
```

```
##    1       0.3447243      1.300288   0.002
##  Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                                    Estimate
## (Intercept)                                       6.1906e+01
## Alcoholic.Beverages..kcal.day.                    5.9803e-03
## Gross.national.income.per.capita..PPP.international... 6.2655e-04
##                                                   Std. Error t value
## (Intercept)                                       1.3157e+00 47.0512
## Alcoholic.Beverages..kcal.day.                    9.5706e-03  0.6249
## Gross.national.income.per.capita..PPP.international... 8.4144e-05  7.4462
##                                                    Pr(>|t|)
## (Intercept)                                       < 2.2e-16 ***
## Alcoholic.Beverages..kcal.day.                       0.5338
## Gross.national.income.per.capita..PPP.international...  8.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model drastically changes the impact of alcoholic beverage consumption on life expectancy.

The coefficient estimate of the impact of alcoholic beverage consumption on life expectancy decreases to .006 and is no longer statistically significant.

Heteroskedasticity of errors and correlation continue to be a problem.

The residuals vs. fitted plot also seems to show a violation of the zero conditional mean assumption.

The presence of heteroskedasticity of errors and a violation of the zero conditional mean assumption may indicate a non-linear relationship in the population.

Including wealth as a control seems to have pulled away what had seemed to be a strong linear relationship between alcohol consumption and life expectancy.

This is a reasonable result, as it seems that wealth would be a key driver for both alcohol consumption and life expectancy.

Adding the wealth control, therefore, reveals that alcohol consumption in an of itself may not be as strongly relate with life expectancy as previously suspected.

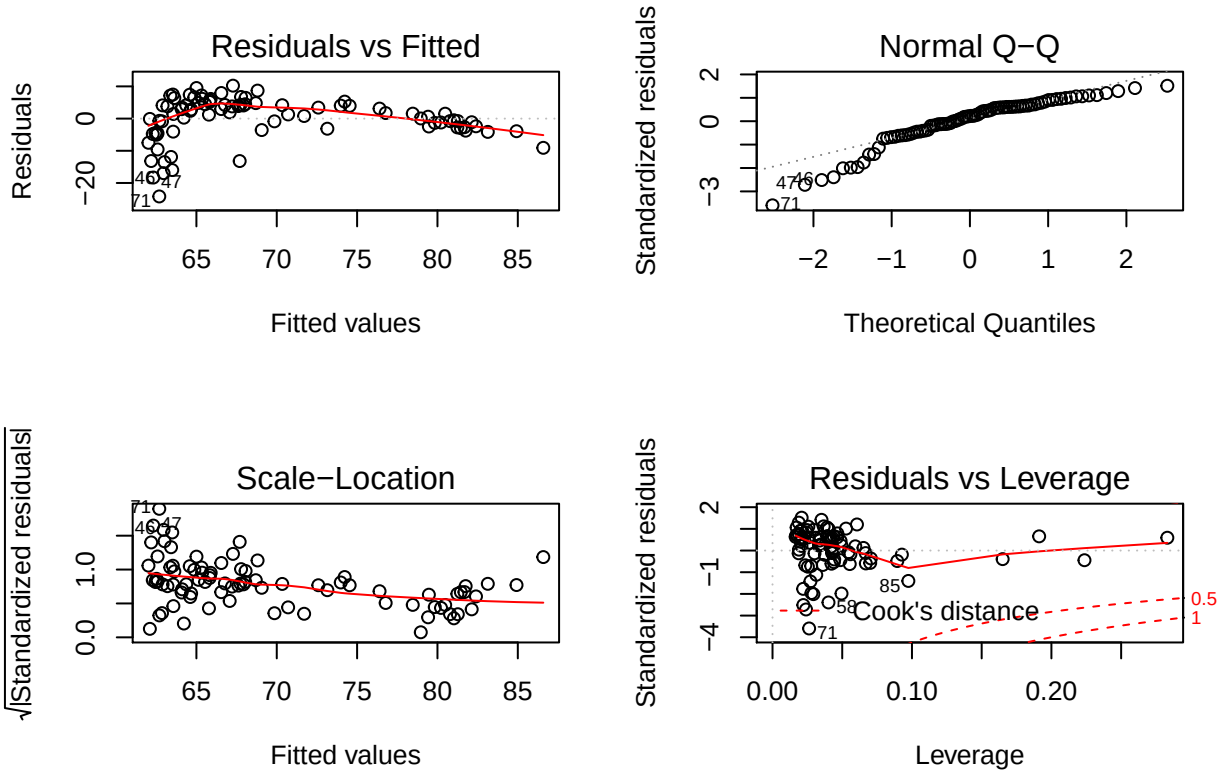
Model 4 - non-linear alcohol consumption with control for GNP - life expectancy ~ alcohol^2 + alcohol + GNP

```
alc.model.3 <- lm(Life.expectancy.at.birth..years..both.sexes ~ I(Alcoholic.Beverages..kcal.day.^2) + Al
summary(alc.model.3)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ I(Alcoholic.Beverages..kcal.day.^2) +
```

```
##     Alcoholic.Beverages..kcal.day. + Gross.national.income.per.capita..PPP.international...,
##     data = diet.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.190  -2.902   1.476   4.250  10.224
##
## Coefficients:
##                                                    Estimate
## (Intercept)                                       6.149e+01
## I(Alcoholic.Beverages..kcal.day.^2)              -6.214e-05
## Alcoholic.Beverages..kcal.day.                    2.053e-02
## Gross.national.income.per.capita..PPP.international...  6.236e-04
##                                                   Std. Error t value
## (Intercept)                                        1.433e+00  42.913
## I(Alcoholic.Beverages..kcal.day.^2)                1.253e-04  -0.496
## Alcoholic.Beverages..kcal.day.                     3.206e-02   0.640
## Gross.national.income.per.capita..PPP.international...  9.010e-05   6.921
##                                                   Pr(>|t|)
## (Intercept)                                        < 2e-16 ***
## I(Alcoholic.Beverages..kcal.day.^2)                  0.621
## Alcoholic.Beverages..kcal.day.                       0.524
## Gross.national.income.per.capita..PPP.international... 9.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.819 on 82 degrees of freedom
## Multiple R-squared:  0.5332, Adjusted R-squared:  0.5162
## F-statistic: 31.23 on 3 and 82 DF,  p-value: 1.457e-13
```

```
par(mfrow=c(2,2))
plot(alc.model.3)
```




```
bptest(alc.model.3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  alc.model.3
## BP = 10.581, df = 3, p-value = 0.01422
```

```
durbinWatsonTest(alc.model.3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3383441      1.312686   0.002
##  Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                                    Estimate
## (Intercept)                                      6.1493e+01
## I(Alcoholic.Beverages..kcal.day.^2)             -6.2138e-05
## Alcoholic.Beverages..kcal.day.                   2.0532e-02
## Gross.national.income.per.capita..PPP.international...  6.2364e-04
##                                                  Std. Error t value
## (Intercept)                                      1.3573e+00 45.3068
```

```
## I(Alcoholic.Beverages..kcal.day.^2)                     9.7465e-05 -0.6375
## Alcoholic.Beverages..kcal.day.                          2.7666e-02  0.7422
## Gross.national.income.per.capita..PPP.international...  8.7053e-05  7.1639
##                                                         Pr(>|t|)
## (Intercept)                                            < 2.2e-16 ***
## I(Alcoholic.Beverages..kcal.day.^2)                       0.5255
## Alcoholic.Beverages..kcal.day.                            0.4601
## Gross.national.income.per.capita..PPP.international...  3.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including a non-linear effect of alcohol consumption in the model does not improve the problems with heteroskedasticity and correlation of errors.

The alcoholic beverage consumption coefficient estimates are still not significant.

The residuals vs. fitted values plot shows heteroskedasticity of errors and the Breusch-Pagan test confirms the errors are heteroskedastic.

Therefore, we need to be sure to use heteroskedasticity robust standard errors to assess statistical significance of the coefficient estimates in the model.

The Durbin-Watson test shows that correlation remains a problem.

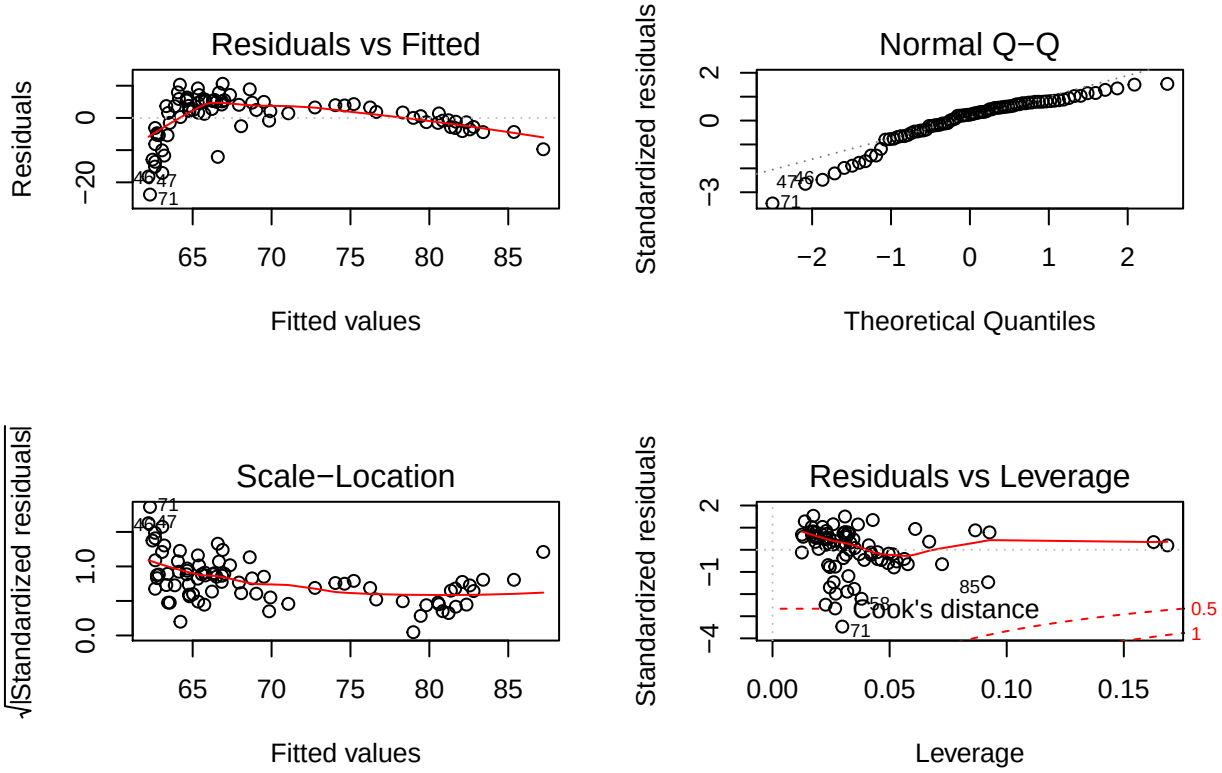
Model 5 - log transformation of alcohol consumption with control for GNP - life expectancy ~ log(alcohol) + GNP

```
#First, remove observations of zero alcoholic beverage consumption so can implement a log transformatio
diet.data.2 <- diet.data[diet.data$Alcoholic.Beverages..kcal.day. > 0, ]

#Estimate the model
alc.model.4 <- lm(Life.expectancy.at.birth..years..both.sexes ~ log(Alcoholic.Beverages..kcal.day.) + Gr
summary(alc.model.4)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ log(Alcoholic.Beverages..kcal.day.) +
##     Gross.national.income.per.capita..PPP.international..., data = diet.data.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.795  -3.122   1.686   4.987  10.601
##
## Coefficients:
##                                                          Estimate
## (Intercept)                                             6.293e+01
## log(Alcoholic.Beverages..kcal.day.)                    -2.492e-01
## Gross.national.income.per.capita..PPP.international...  6.802e-04
##                                                         Std. Error t value
## (Intercept)                                              2.728e+00  23.070
## log(Alcoholic.Beverages..kcal.day.)                      7.659e-01  -0.325
## Gross.national.income.per.capita..PPP.international...   8.863e-05   7.675
##                                                         Pr(>|t|)
## (Intercept)                                              < 2e-16 ***
```

```
## log(Alcoholic.Beverages..kcal.day.)                          0.746
## Gross.national.income.per.capita..PPP.international... 4.01e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.966 on 78 degrees of freedom
## Multiple R-squared:  0.5266, Adjusted R-squared:  0.5145
## F-statistic: 43.39 on 2 and 78 DF,  p-value: 2.15e-13
```

```
par(mfrow=c(2,2))
plot(alc.model.4)
```




```
bptest(alc.model.4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  alc.model.4
## BP = 11.328, df = 2, p-value = 0.003468
```

```
durbinWatsonTest(alc.model.4)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2942603      1.402137    0.01
##  Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.4, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                                    Estimate
## (Intercept)                                       6.2935e+01
## log(Alcoholic.Beverages..kcal.day.)              -2.4921e-01
## Gross.national.income.per.capita..PPP.international...  6.8021e-04
##                                                   Std. Error t value
## (Intercept)                                       2.0382e+00 30.8776
## log(Alcoholic.Beverages..kcal.day.)               6.5485e-01 -0.3806
## Gross.national.income.per.capita..PPP.international...  9.4874e-05  7.1696
##                                                   Pr(>|t|)
## (Intercept)                                      < 2.2e-16 ***
## log(Alcoholic.Beverages..kcal.day.)                 0.7046
## Gross.national.income.per.capita..PPP.international...  3.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including a log transformation for alcoholic beverage consumption does not fix the problem of heteroskedasticity of errors as evidenced by the residuals vs. fitted values plot and the Breusch Pagan test.

The Durbin Watson test also shows that correlation of errors remains a problem.

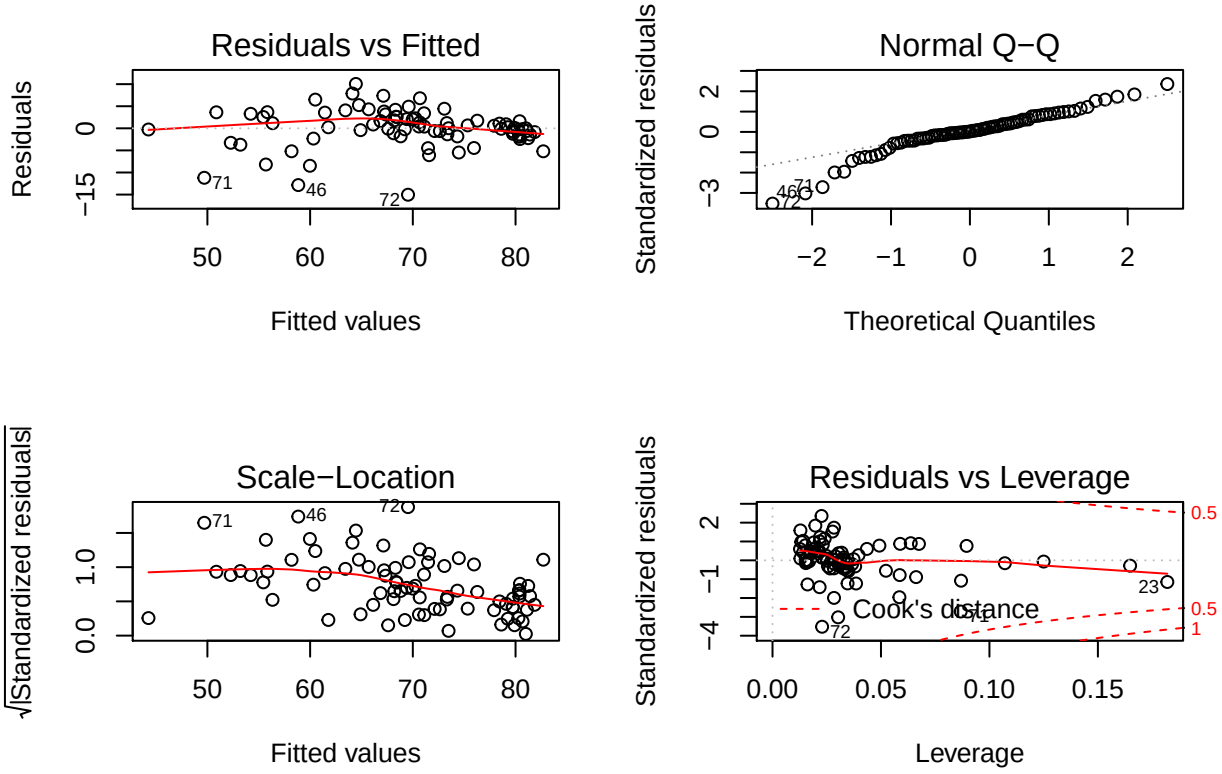
Model 6 - log transformation of alcohol consumption with control for log transformation of GNP - life expectancy ~ log(alcohol) + log(GNP)

```
#Estimate the model
alc.model.5 <- lm(Life.expectancy.at.birth..years..both.sexes ~ log(Alcoholic.Beverages..kcal.day.) + l
summary(alc.model.5)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ log(Alcoholic.Beverages..kcal.day.) +
##     log(Gross.national.income.per.capita..PPP.international...),
##     data = diet.data.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.033  -1.412   0.098   2.553  10.043
##
## Coefficients:
##                                                    Estimate
## (Intercept)                                          5.2317
## log(Alcoholic.Beverages..kcal.day.)                 -1.2710
## log(Gross.national.income.per.capita..PPP.international...)   7.9595
##                                                   Std. Error
## (Intercept)                                          3.4695
## log(Alcoholic.Beverages..kcal.day.)                  0.4674
```

```
## log(Gross.national.income.per.capita..PPP.international...)     0.4766
##                                                                t value
## (Intercept)                                                      1.508
## log(Alcoholic.Beverages..kcal.day.)                             -2.719
## log(Gross.national.income.per.capita..PPP.international...)     16.702
##                                                                Pr(>|t|)
## (Intercept)                                                     0.13562
## log(Alcoholic.Beverages..kcal.day.)                             0.00806 **
## log(Gross.national.income.per.capita..PPP.international...)     < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.314 on 78 degrees of freedom
## Multiple R-squared:  0.8185, Adjusted R-squared:  0.8138
## F-statistic: 175.8 on 2 and 78 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(alc.model.5)
```




```
bptest(alc.model.5)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  alc.model.5
## BP = 9.9923, df = 2, p-value = 0.006764
```

```
durbinWatsonTest(alc.model.5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1371907      1.694322    0.14
##  Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.5, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                                                 Estimate
## (Intercept)                                                      5.23166
## log(Alcoholic.Beverages..kcal.day.)                             -1.27103
## log(Gross.national.income.per.capita..PPP.international...)      7.95946
##                                                                 Std. Error
## (Intercept)                                                        4.13913
## log(Alcoholic.Beverages..kcal.day.)                                0.49582
## log(Gross.national.income.per.capita..PPP.international...)        0.58735
##                                                                 t value
## (Intercept)                                                      1.2640
## log(Alcoholic.Beverages..kcal.day.)                             -2.5635
## log(Gross.national.income.per.capita..PPP.international...) 13.5515
##                                                                 Pr(>|t|)
## (Intercept)                                                      0.21001
## log(Alcoholic.Beverages..kcal.day.)                              0.01229 *
## log(Gross.national.income.per.capita..PPP.international...)      < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using a log transformation on GNP and alcohol consumption makes sense because each of these variables is positively skewed.

After making these transformations, the Durbin-Watson test shows that correlated errors seems to have been solved.

The residuals vs. fitted values plot shows that heteroskedasticity of errors continues to be a problem. The Breusch-Pagan test confirms this result.

Using heteroskedasticity robust standard errors, the coefficients are both statistically significant.

The coefficient estimate on log(alcohol consumption) is -1.271 which is statistically significant ($p = .012$ using heteroskedasticity robust errors).

The interpretation of this coefficient is that a one percent increase in alcohol consumption corresponds with a decrease of 1.271 years in life expectancy while holding GNP equal.

While this model outputs a statistically significant coefficient estimate of alcohol consumption, now its relationship with life expectancy is reversed, making it fairly suspect that there is a real meaningful relationship between the two variables.

In contrast, the wealth control, the GNP, still retained a similar relationship with life expectancy, which is consistent with its coefficient estimates in previous models.