

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 its original source could no longer be found. To offset the lack of knowledge/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-statistics>

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 occurred in certain instances). There are many variables relating to health such as obesity occurrence, 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 = "Li")
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
```

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

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

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

Independent Variables

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

```
#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
```

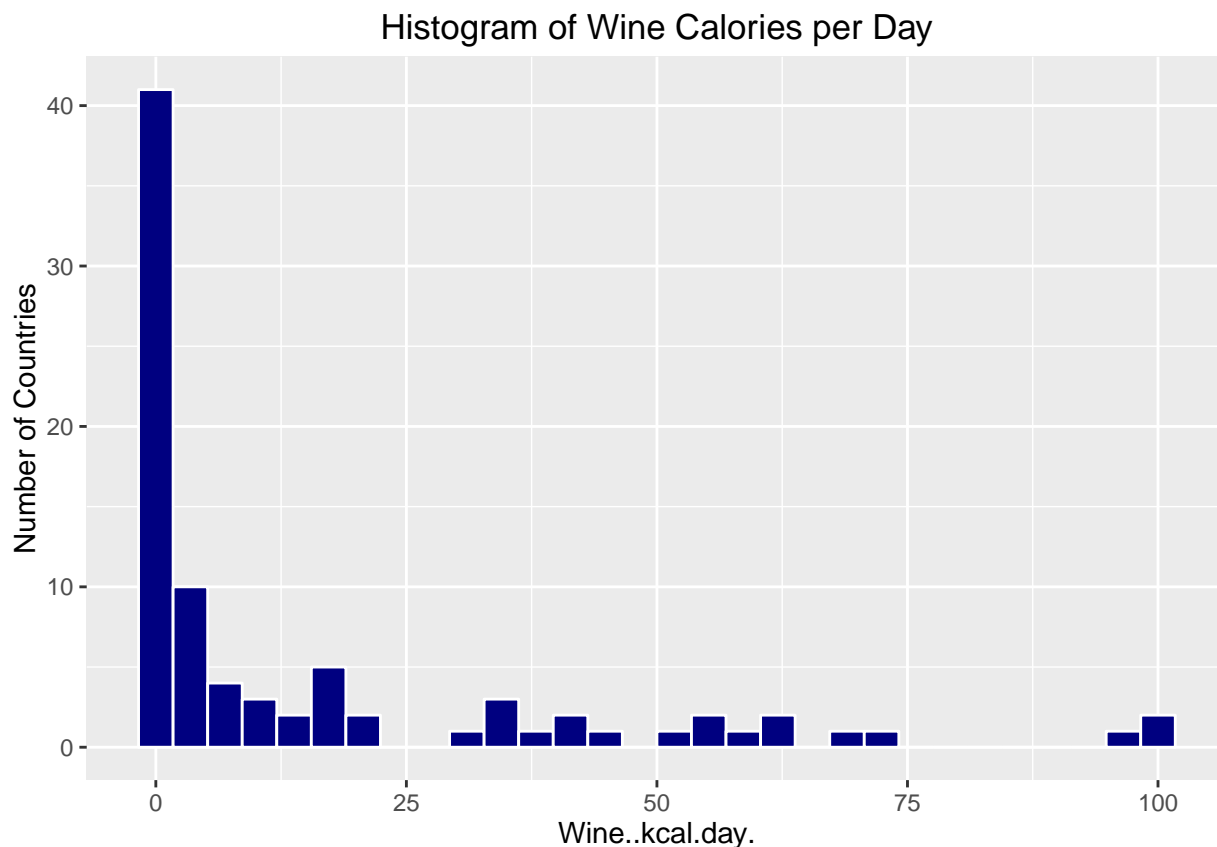
```
sum(diet.data$Wine..kcal.day. == 0)
```

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

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

```
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 D
```

```
## `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.

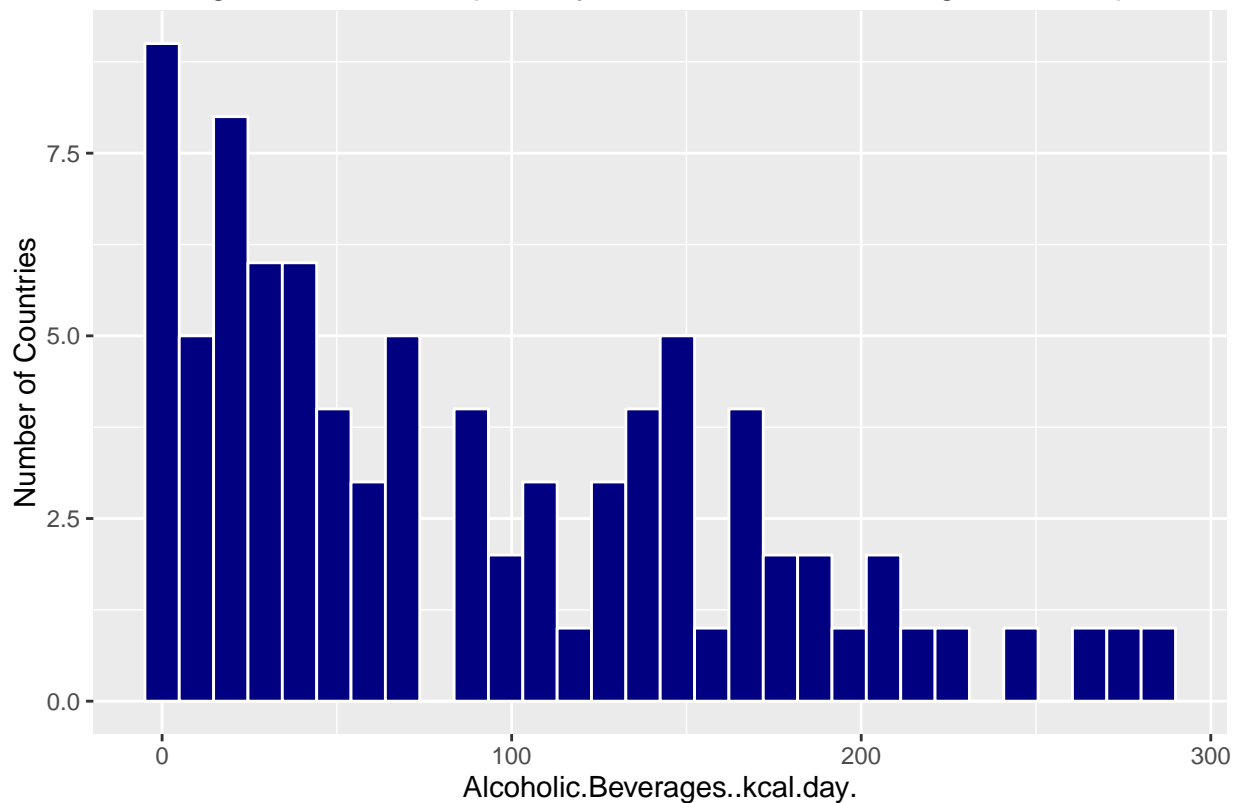
```
diet.data$pure
```

```
## NULL
```

```
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 Calor
```

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

Histogram of Calories per Day from Alcoholic Beverage Consumption



```
#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.)
```

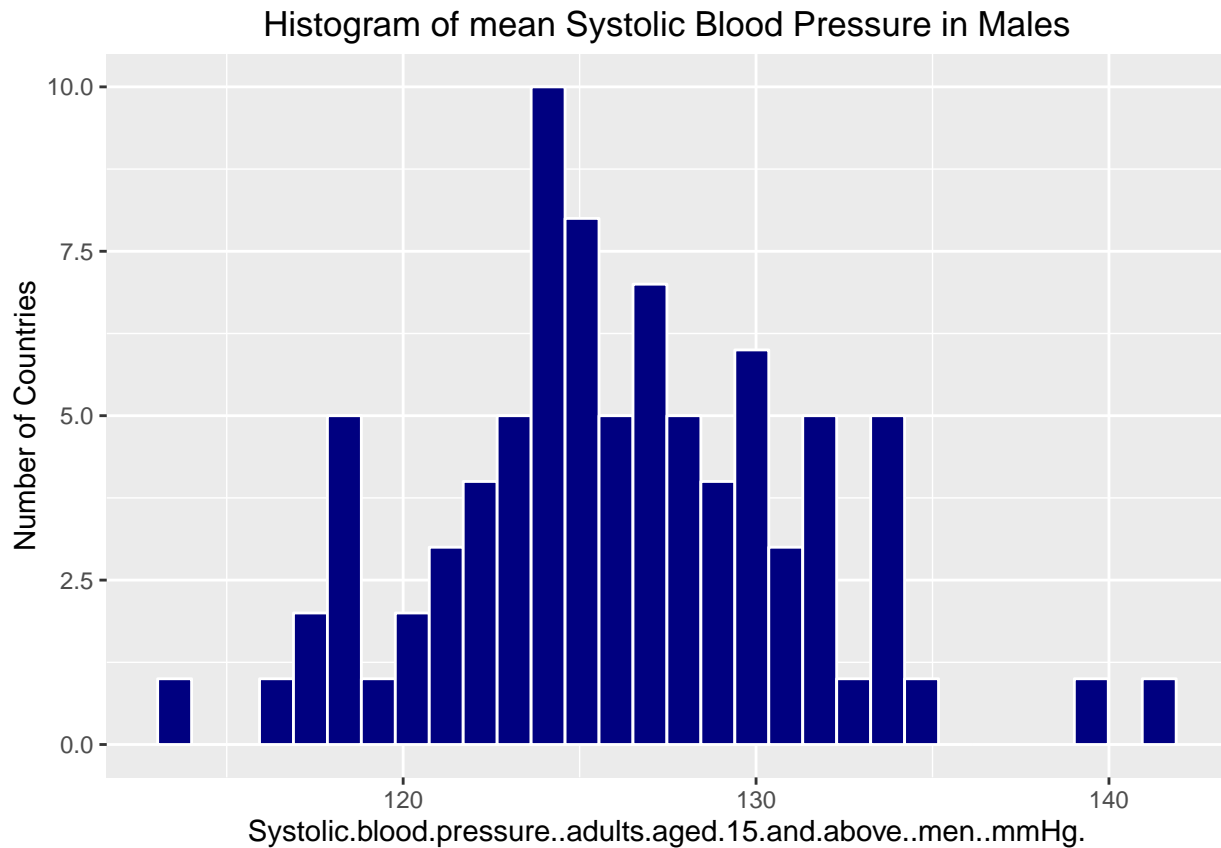
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   113.9   123.3   125.9   126.2   129.5   141.8
```

```
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.above..men..mmHg)) +
  geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of mean Systolic Blood Pressure in Males")
```

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



No countries seem to be missing or bottom coded. Values seem normally distributed.

Obesity - prevalence in percentage

```
# Justification for using only male blood pressure (highly correlated)
cor(diet.data$Obesity.prevalence..men..., diet.data$Obesity.prevalence..female...)
```

```
## [1] 0.6956928
```

```
summary(diet.data$Obesity.prevalence..men...)
```

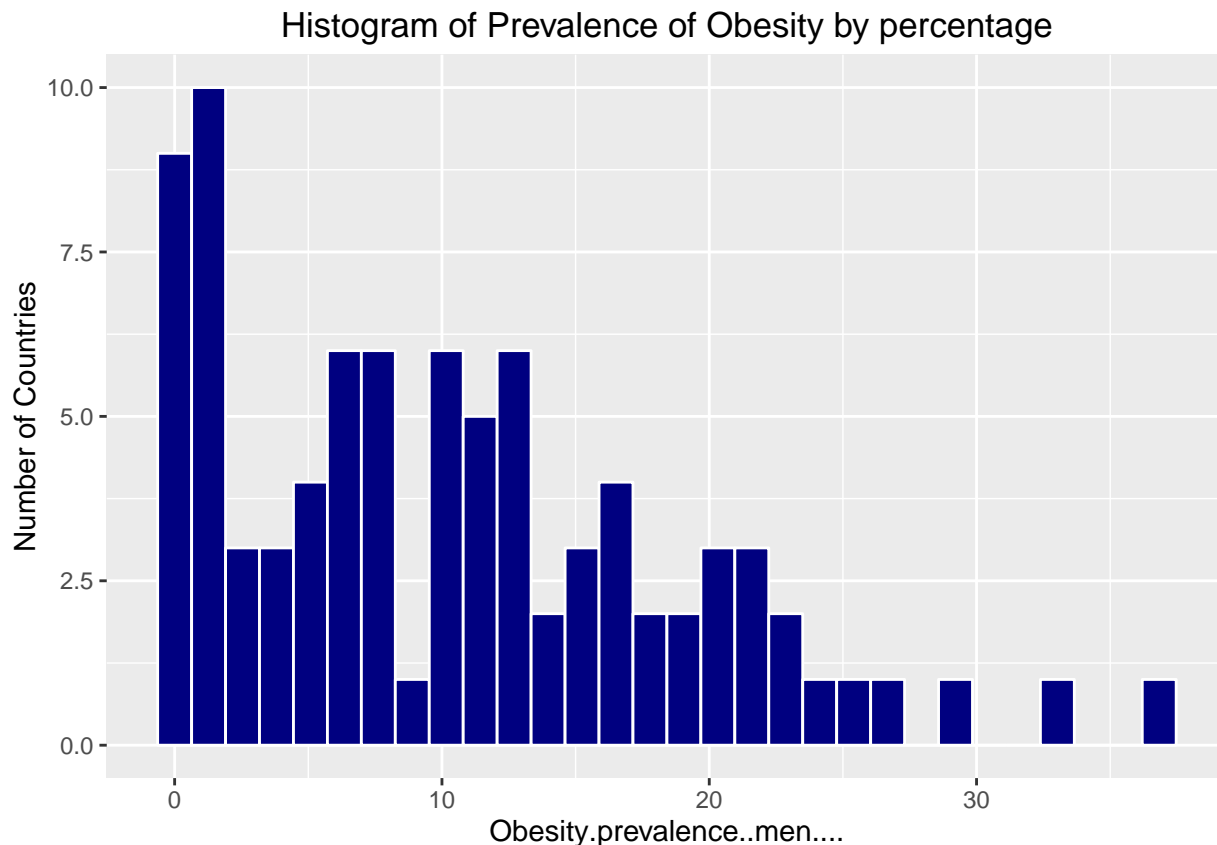
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.100  2.992   9.600  10.520  15.970  36.930
```

```
sum(diet.data$Obesity.prevalence..men... == 0)
```

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

```
Obesity.pcnt.hist <- ggplot(data = diet.data, aes(x = Obesity.prevalence..men....))
Obesity.pcnt.hist + geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of Prevalence of Obesity by percentage")

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



No countries missing or bottom coded for obesity. This variable appears to have a positive skew. Might want to consider an alternative specification for it in models. Also worth noting that while obesity would not be considered a sign of health in an individual, at the country level it may be an indicator of physical health (we defined physical health by life expectancy) because it may correlate with financial wealth and other positive indicators of an otherwise physically healthy life.

Cholesterol - total mean (mg/dl) 2005

```
# Justification for using only male blood pressure (highly correlated)
cor(diet.data$Mean.total.cholesterol..men..mg.dl...2005, diet.data$Mean.total.cholesterol..female..mg.dl...2005)

## [1] 0.9054735

diet.data[diet.data$Mean.total.cholesterol..men..mg.dl...2005>230,c("Countries","Mean.total.cholesterol..men..mg.dl...2005")]

##           Countries Mean.total.cholesterol..men..mg.dl...2005
## 6              Belarus                                234.0
## 16             Colombia                                245.7
```

```
## 18 Cyprus 234.0
## 79 Trinidad and Tobago 234.0
```

```
summary(diet.data$Mean.total.cholesterol..men..mg.dl...2005)
```

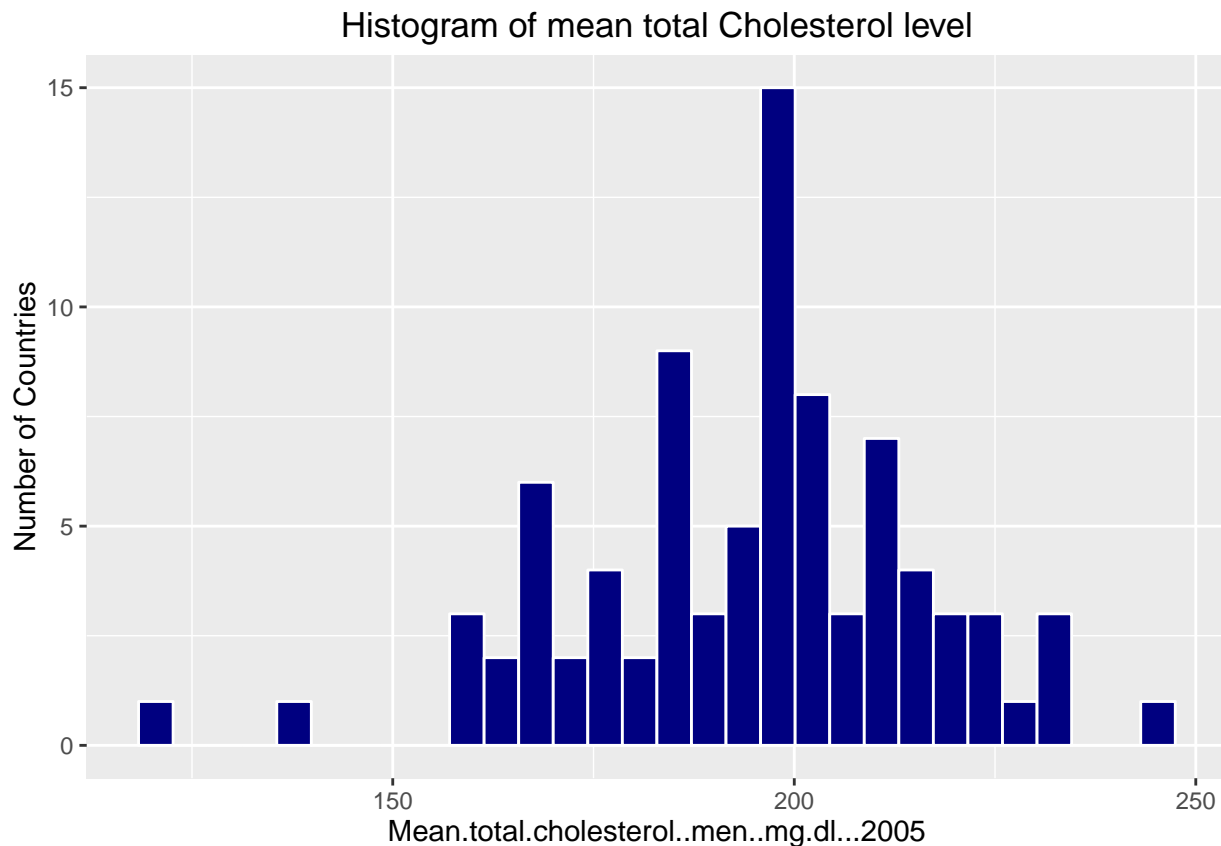
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    120.9   183.3   198.9   194.7   209.6   245.7
```

```
sum(diet.data$Mean.total.cholesterol..men..mg.dl...2005 == 0)
```

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

```
Cholesterol.mean.hist <- ggplot(data = diet.data, aes(x = Mean.total.cholesterol..men..mg.dl...2005))
Cholesterol.mean.hist + geom_histogram(fill = "navy", colour = "white") + ggtitle("Histogram of mean to
```

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



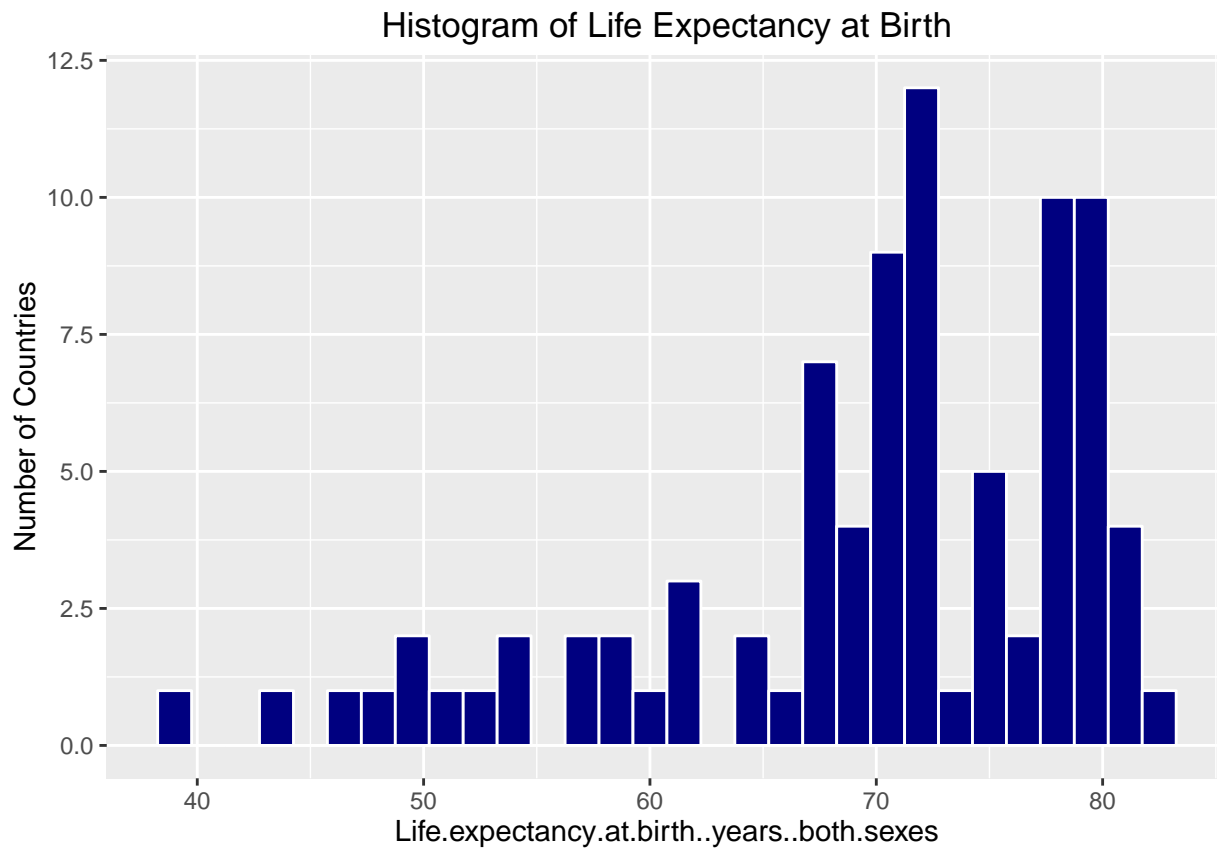
No missing/bottom coded countries. Variable is mostly normally distributed.

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 Exp
```

```
## `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).

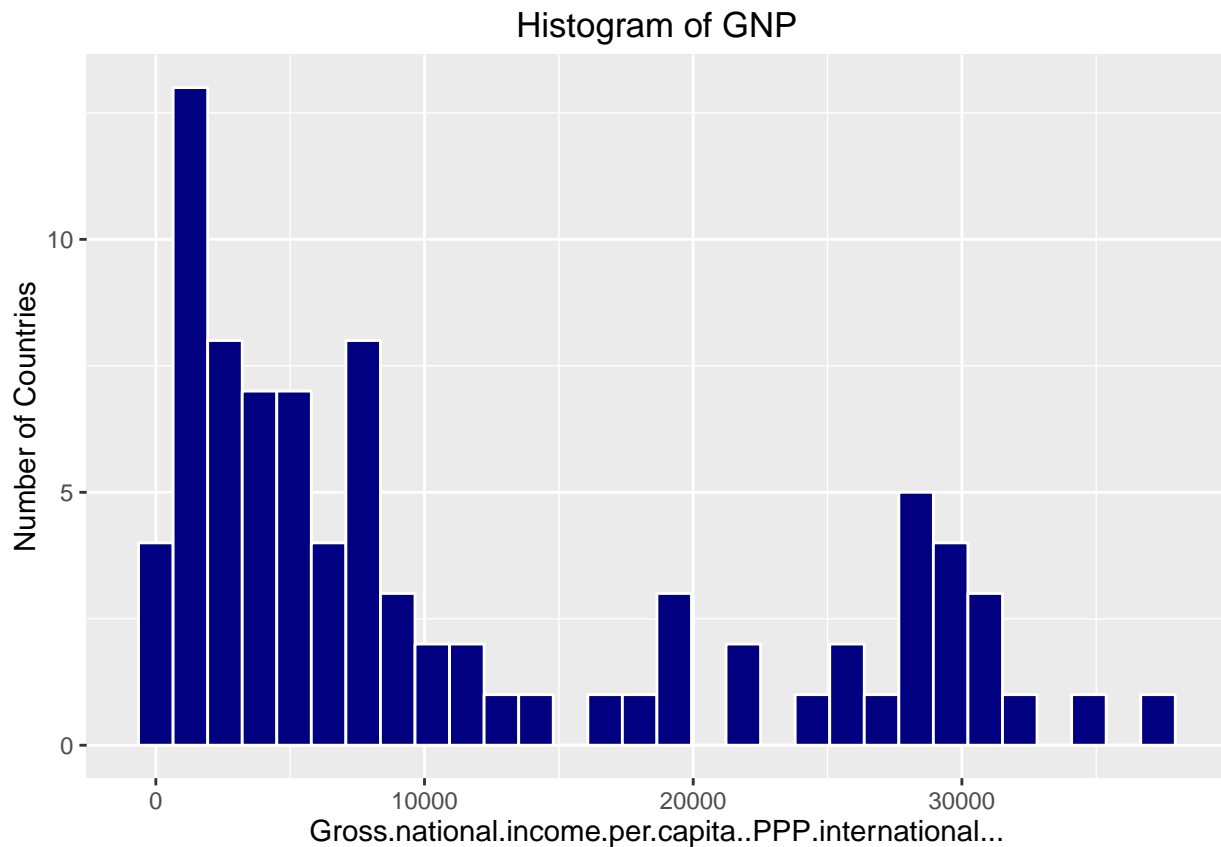
Gross National Product (GNP)

```
#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 = "Num
```

```
## `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 birth
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

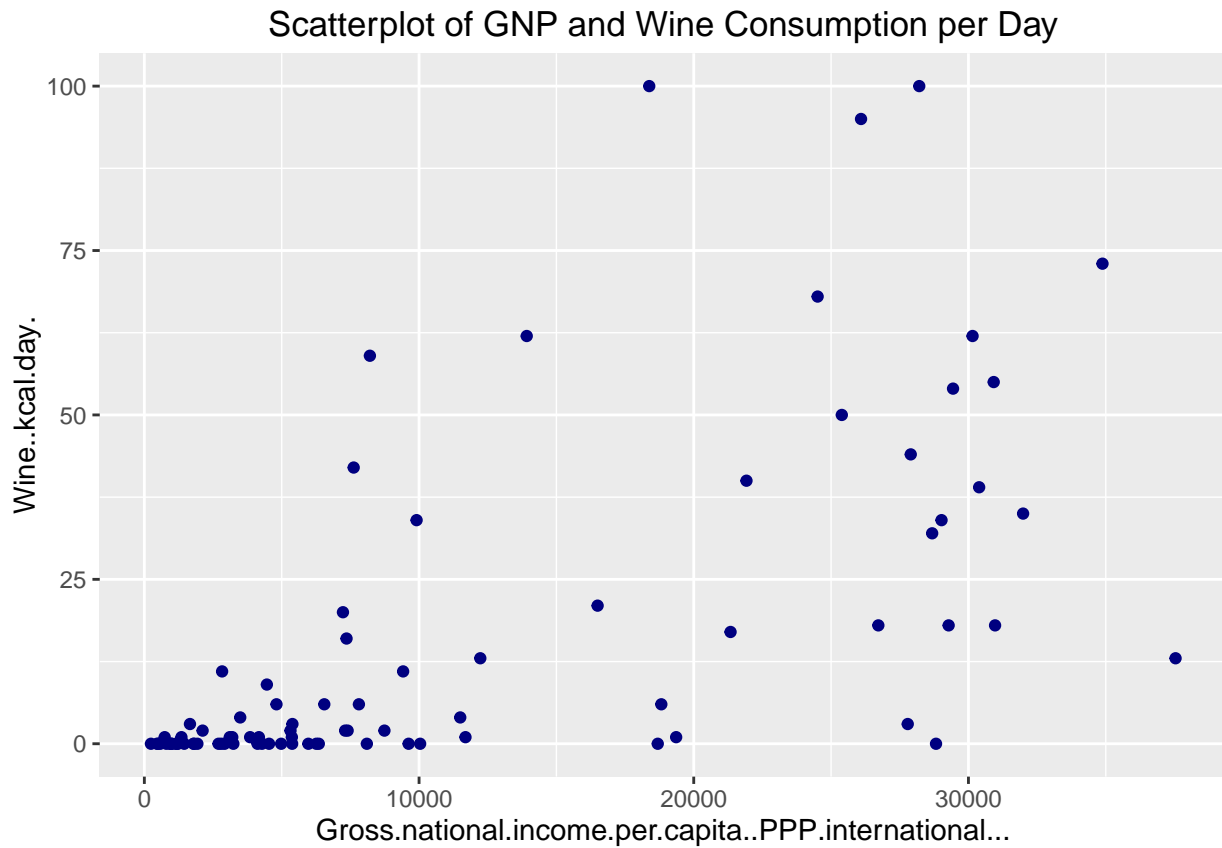
#look at correlation between wine / alcohol consumption and GNP to see if the above result appears to be
#There are very high correlations between wine and alcohol consumption with GNP, both being above 0.6.
cor(diet.data$Gross.national.income.per.capita..PPP.international..., diet.data$Wine..kcal.day.)

## [1] 0.6433328

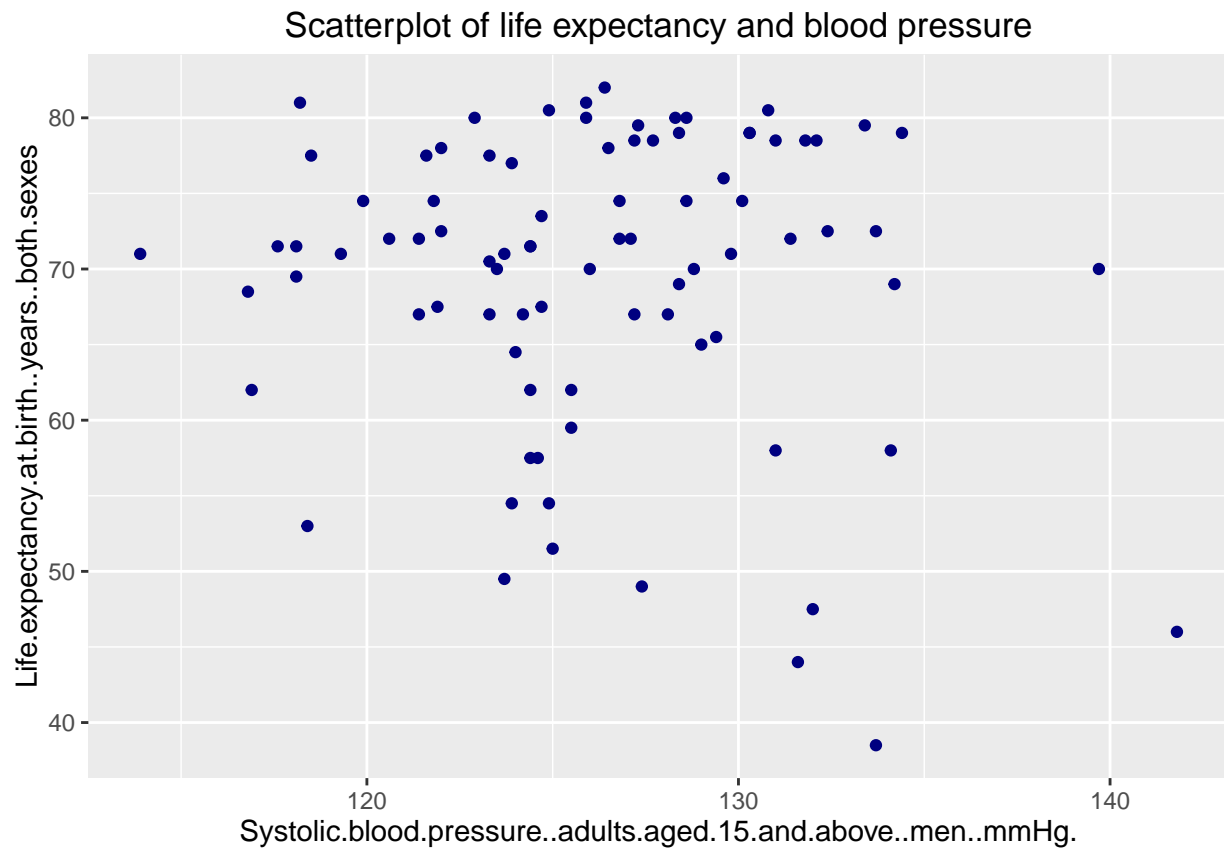
cor(diet.data$Gross.national.income.per.capita..PPP.international..., diet.data$Alcoholic.Beverages..kcal.day.)

## [1] 0.6576478
```

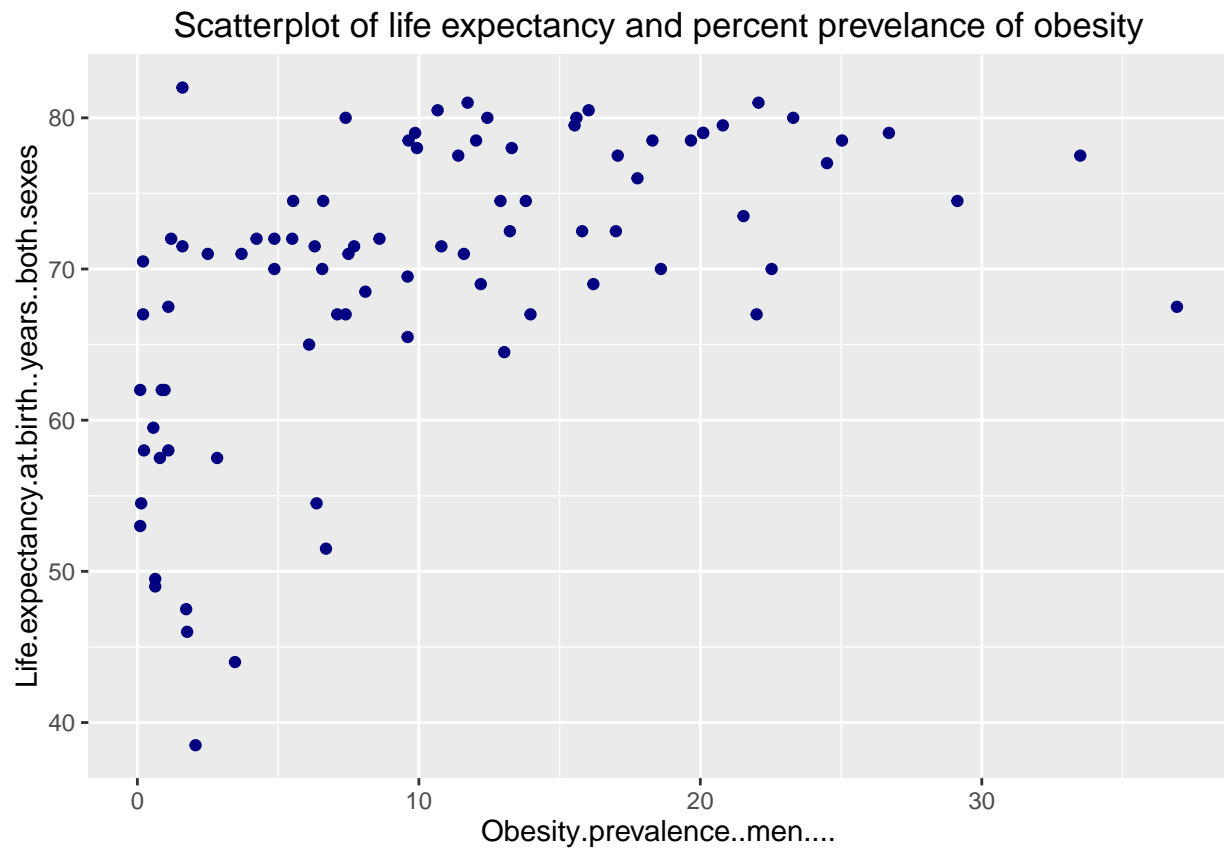
```
#diet.data$Alcoholic.Beverages..kcal.day. > 0
wine.gnp.scatter <- ggplot(data = diet.data, aes(x = Gross.national.income.per.capita..PPP.international...))
wine.gnp.scatter + geom_point(colour = "navy") + ggtitle("Scatterplot of GNP and Wine Consumption per Day")
```



```
# Scatterplot for Blood Pressure vs. Life Expectancy
bp.life.scatter <- ggplot(data = diet.data, aes(x = Systolic.blood.pressure..adults.aged.15.and.above..))
bp.life.scatter + geom_point(colour = "navy") + ggtitle("Scatterplot of life expectancy and blood pressure")
```



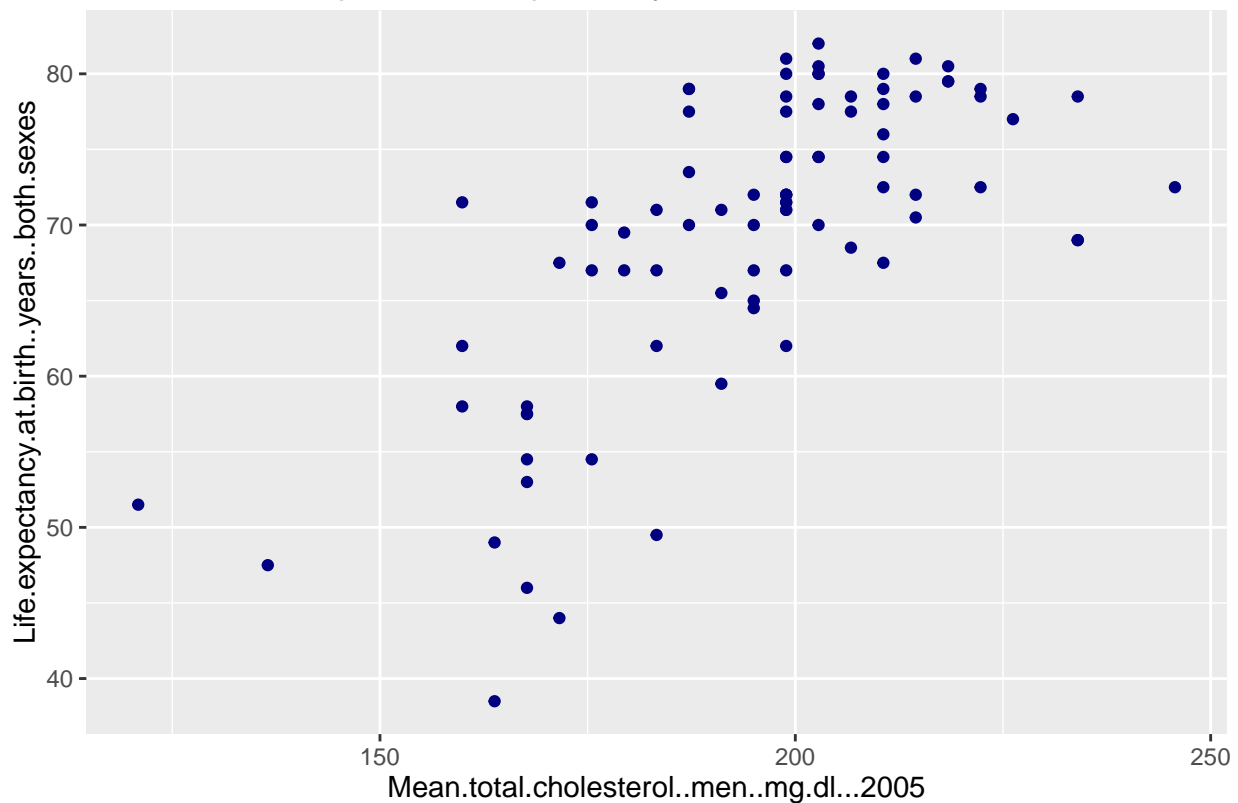
```
# Scatterplot for Obesity vs. Life Expectancy
Obesity.life.scatter <- ggplot(data = diet.data, aes(x = Obesity.prevalence..men...., y =Life.expectancy))
Obesity.life.scatter + geom_point(colour = "navy") + ggtitle("Scatterplot of life expectancy and percent")
```



```
# Scatterplot for Cholesterol vs. Life Expectancy
```

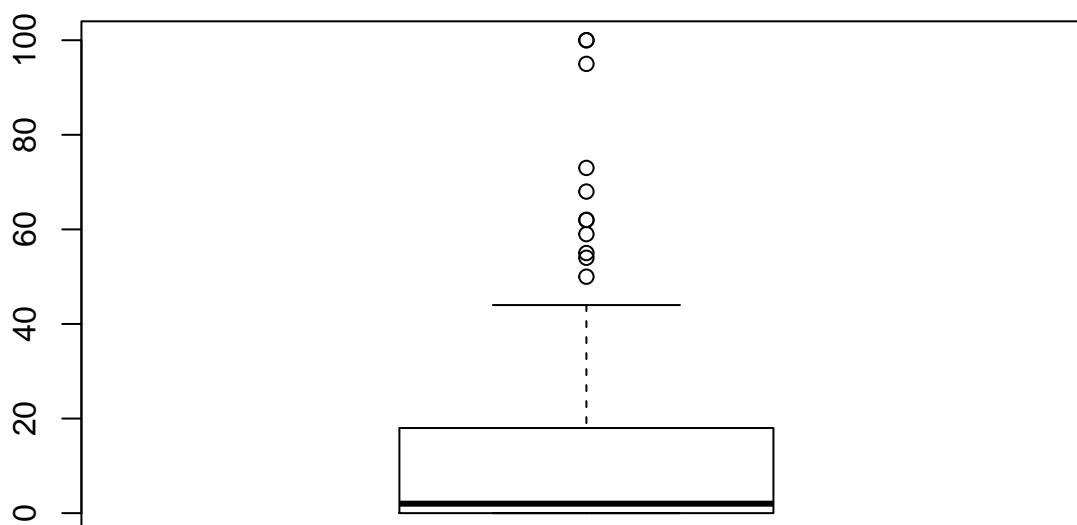
```
Cholesterol.life.scatter <- ggplot(data = diet.data, aes(x = Mean.total.cholesterol..men..mg.dl...2005,  
Cholesterol.life.scatter + geom_point(colour = "navy") + ggtitle("Scatterplot of life expectancy and me
```

Scatterplot of life expectancy and mean total cholesterol



```
#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 Consumption (kcal/day)")
```

Boxplot of Wine Consumption (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
```

From the scatterplots of our potential physical health indicators (predictors of life expectancy), we see potential relationships from all variables (blood pressure, cholesterol and obesity). Interestingly, only blood pressure appears to have a negative correlation with blood pressure, whereas both cholesterol and obesity have potentially positive correlations with life expectancy. Though we will posit this theory often, it seems likely that these variables may be good indicators of GNP (financial health) and through GNP they may be relating to life expectancy (physical health). While on a personal level this would seem strange (as being obese certainly shouldn't increase your life span), being obese indicates you can afford food and likely have easy access to things like clean water, which if you didn't have would decrease your life expectancy much more dramatically than obesity would.

Looking at the boxplot, we see countries with “outlier-level” wine consumption, and their life expectancy.

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 Europe, 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 indication that greater wine consumption could be a good predictor of life expectancy (physical health).

Data Validation

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

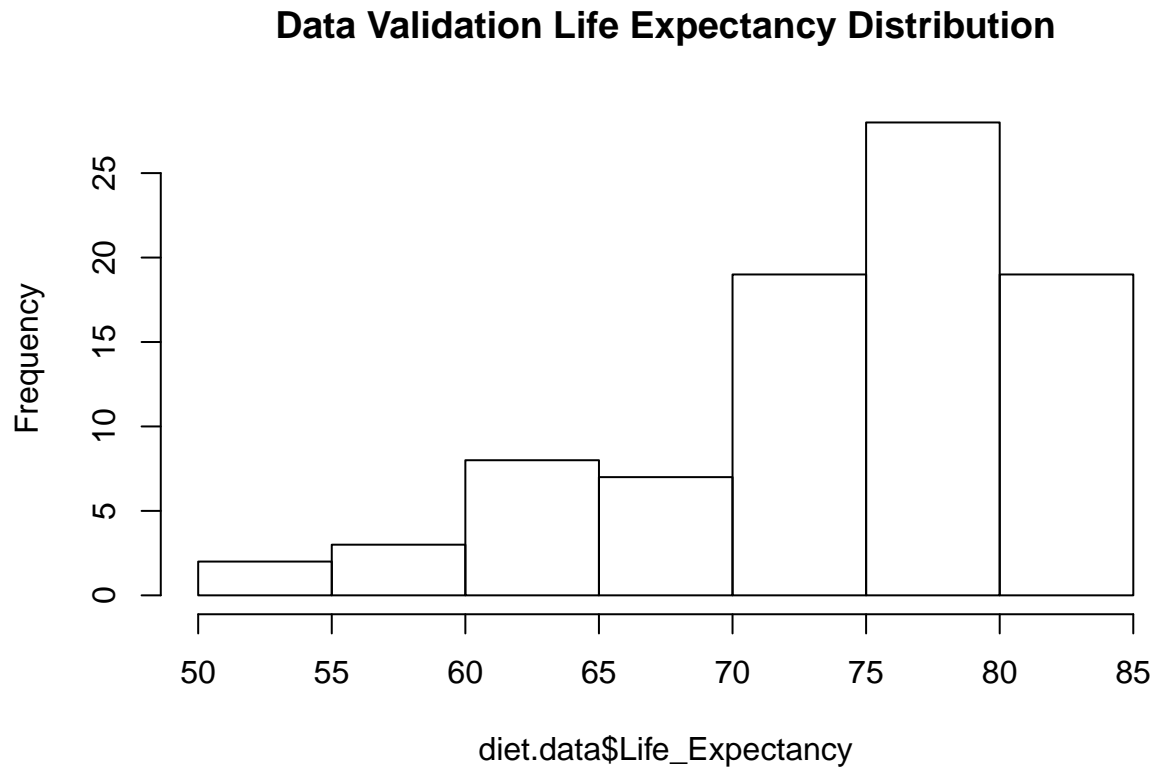
#Merge validating data into the main country dataset.
diet.data <- merge(diet.data, data.validation.growth_rate[,c("Country_main_dataset", "Growth_rate_2000"
diet.data <- merge(diet.data, data.validation.life_expect[,c("Country_main_dataset", "Life_Expectancy")

#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.dat

summary(diet.data$Life_Expectancy_pct_diff)
```

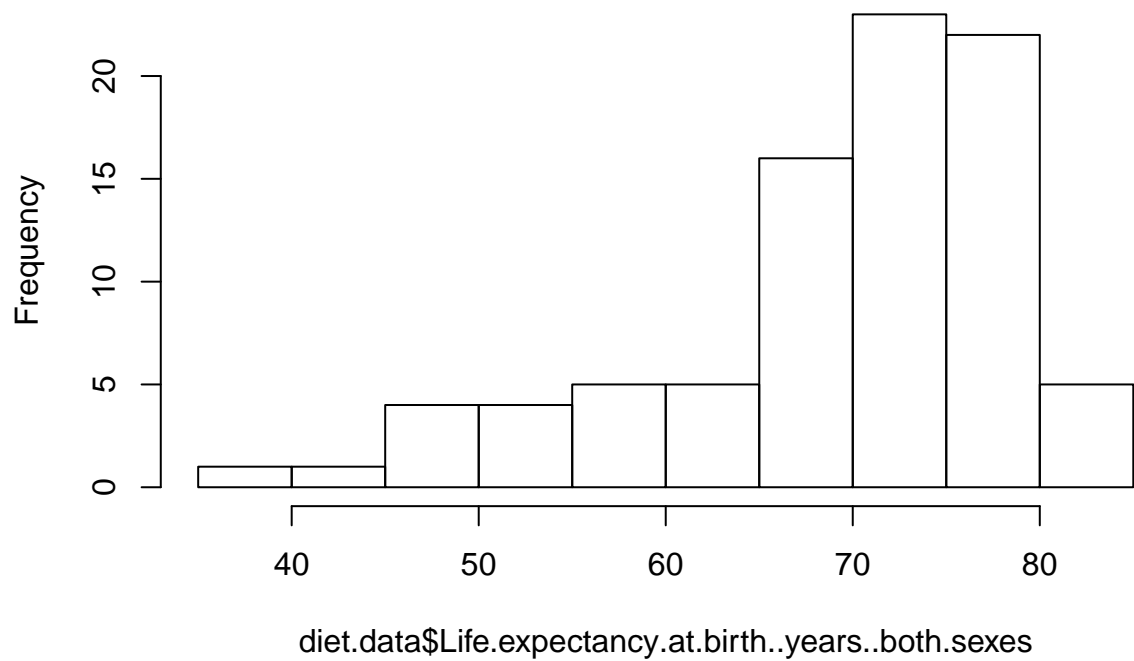
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.511700 -0.087700 -0.047750 -0.073340 -0.028300 -0.002548
```

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



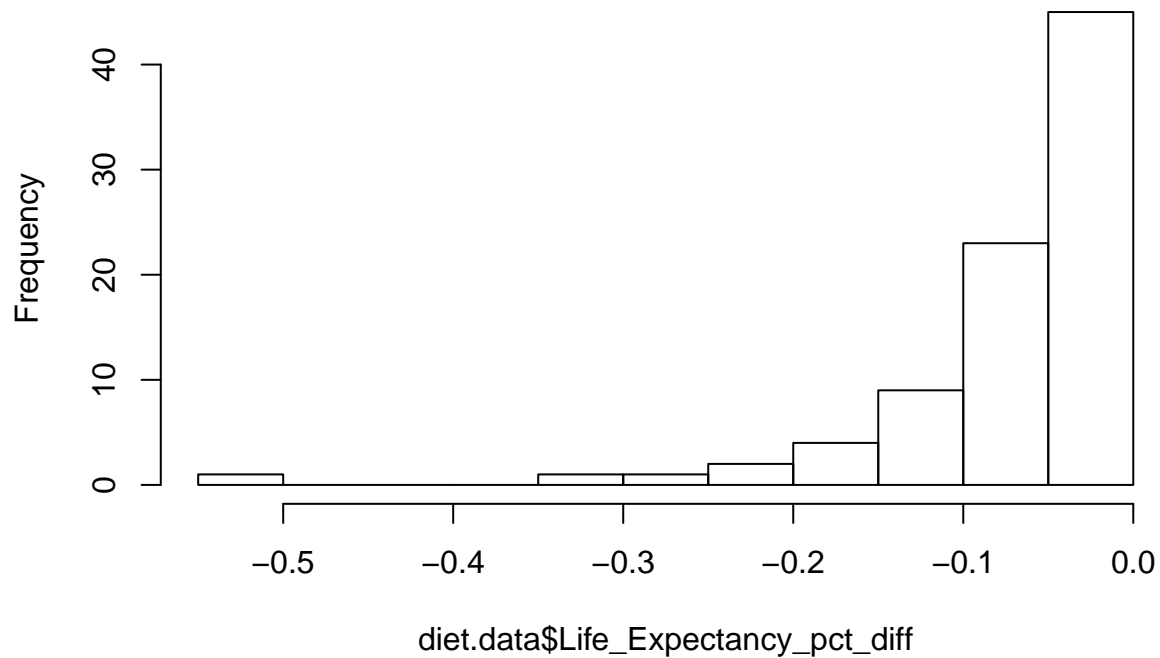
```
hist(diet.data$Life.expectancy.at.birth..years..both.sexes, main = "Data Validation Original Life Expectancy Distribution")
```


Data Validation Original Life Expectancy Distribution



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

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_rate_2010")]

Growth.rate.examination$Growth_rate_pct_diff_2000 <- (Growth.rate.examination$Population.annual.growth.rate_2000 - Growth.rate.examination$Population.annual.growth.rate_2005) / Growth.rate.examination$Population.annual.growth.rate_2005
Growth.rate.examination$Growth_rate_pct_diff_2005 <- (Growth.rate.examination$Population.annual.growth.rate_2005 - Growth.rate.examination$Population.annual.growth.rate_2010) / Growth.rate.examination$Population.annual.growth.rate_2005
Growth.rate.examination$Growth_rate_pct_diff_2010 <- (Growth.rate.examination$Population.annual.growth.rate_2010 - Growth.rate.examination$Population.annual.growth.rate_2005) / Growth.rate.examination$Population.annual.growth.rate_2005

#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)
```

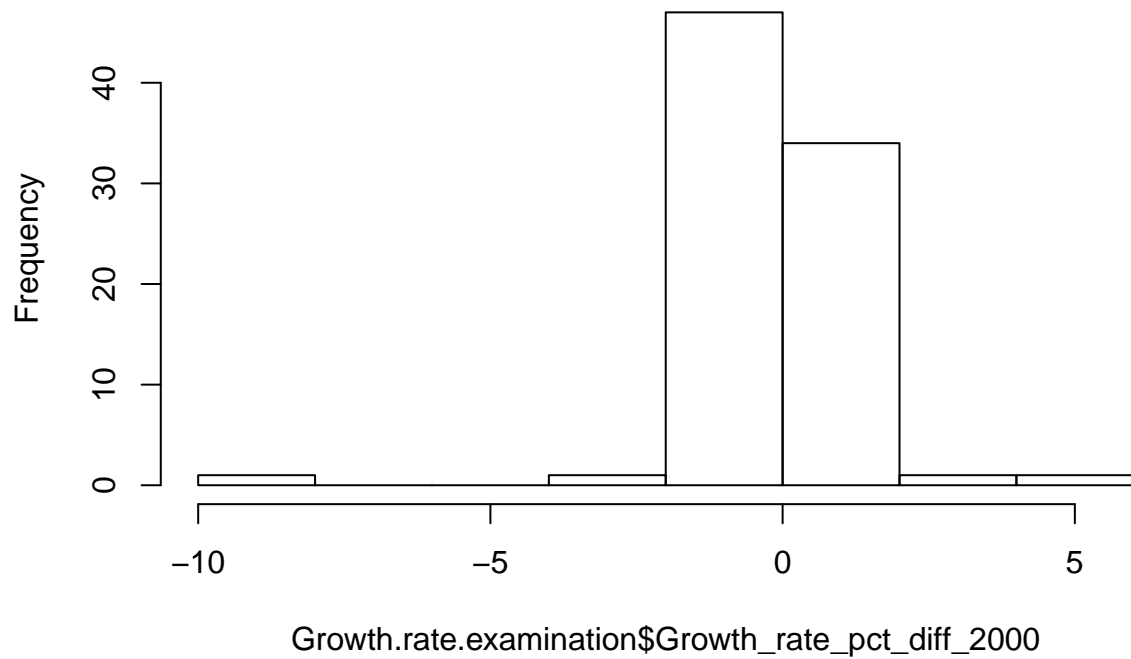
```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
## -1.3160  0.3746  1.0960  1.1720  1.6160  7.7870
```

```
#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
```

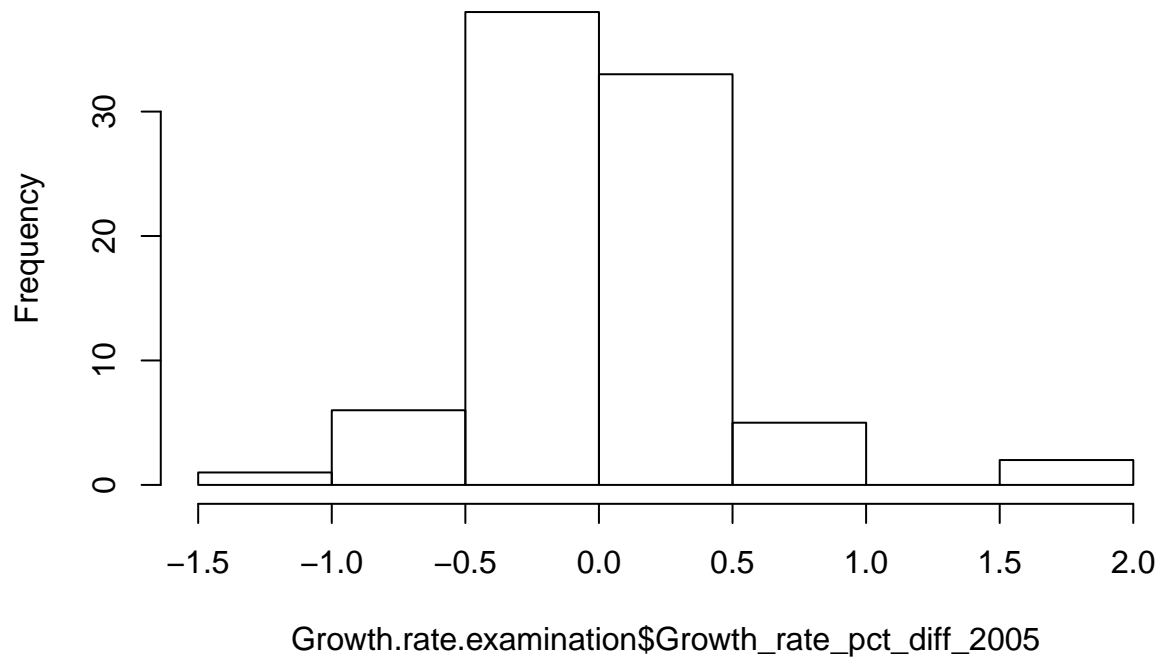
```
hist(Growth.rate.examination$Growth_rate_pct_diff_2000, main = "Histogram of Growth Rate % Diff with 2000")
```

Histogram of Growth Rate % Diff with 2000 Growth Rate



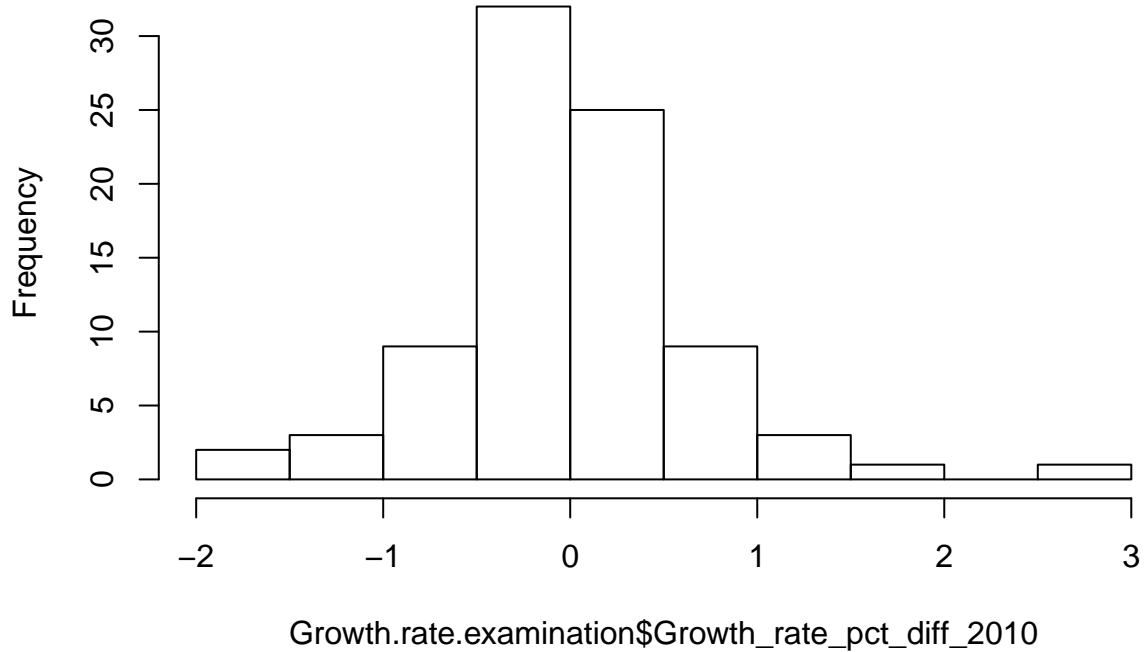
```
hist(Growth.rate.examination$Growth_rate_pct_diff_2005, main = "Histogram of Growth Rate % Diff with 2000 Growth Rate")
```

Histogram of Growth Rate % Diff with 2005 Growth Rate



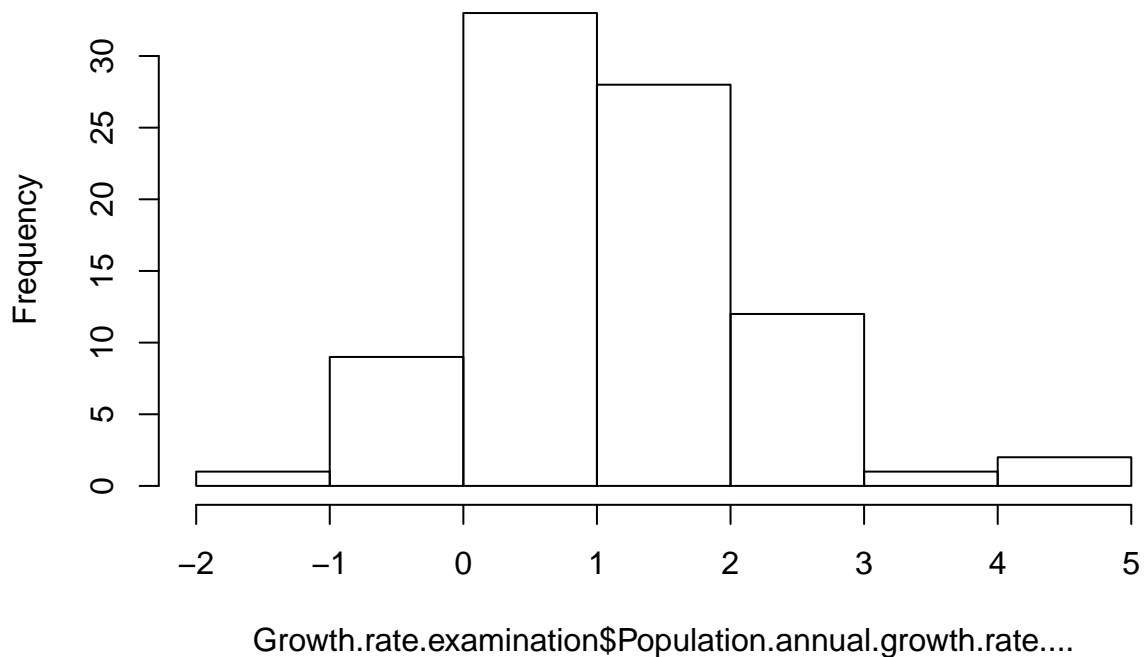
```
hist(Growth.rate.examination$Growth_rate_pct_diff_2010, main = "Histogram of Growth Rate % Diff with 2010 Growth Rate")
```

Histogram of Growth Rate % Diff with 2010 Growth Rate

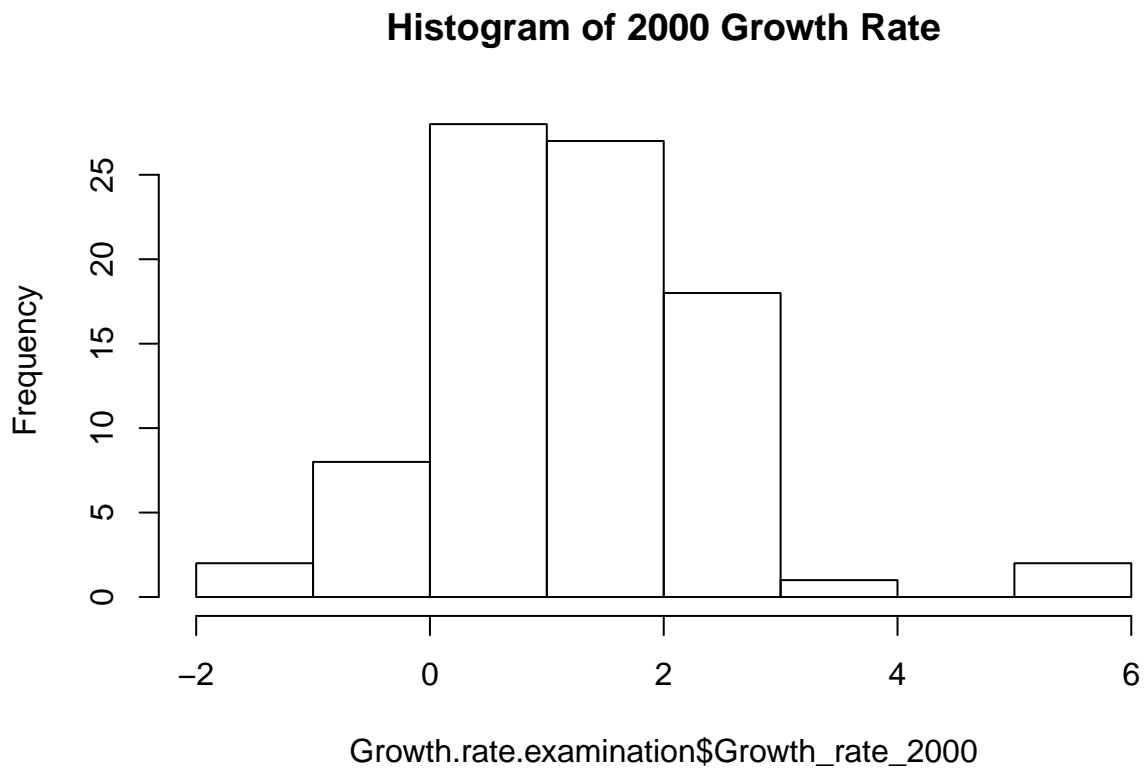


```
#Histograms of each growth rate  
hist(Growth.rate.examination$Population.annual.growth.rate..., main = "Histogram of Original Growth Rate")
```

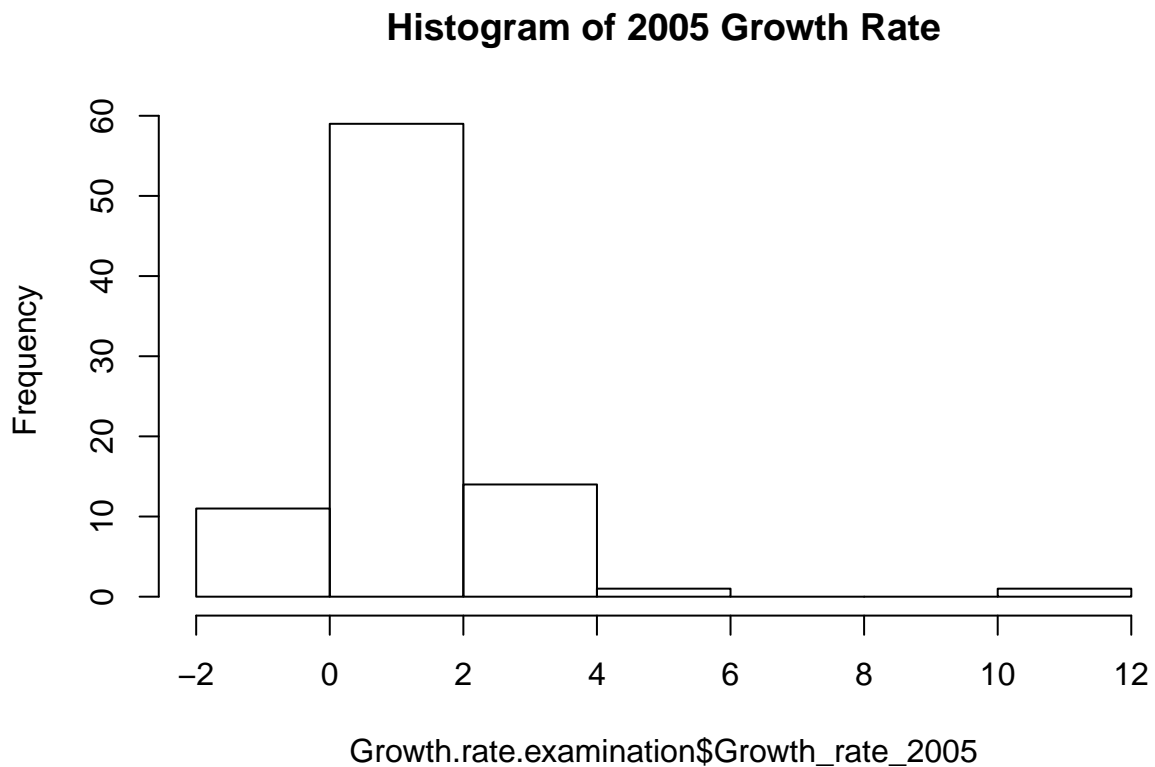
Histogram of Original Growth Rate



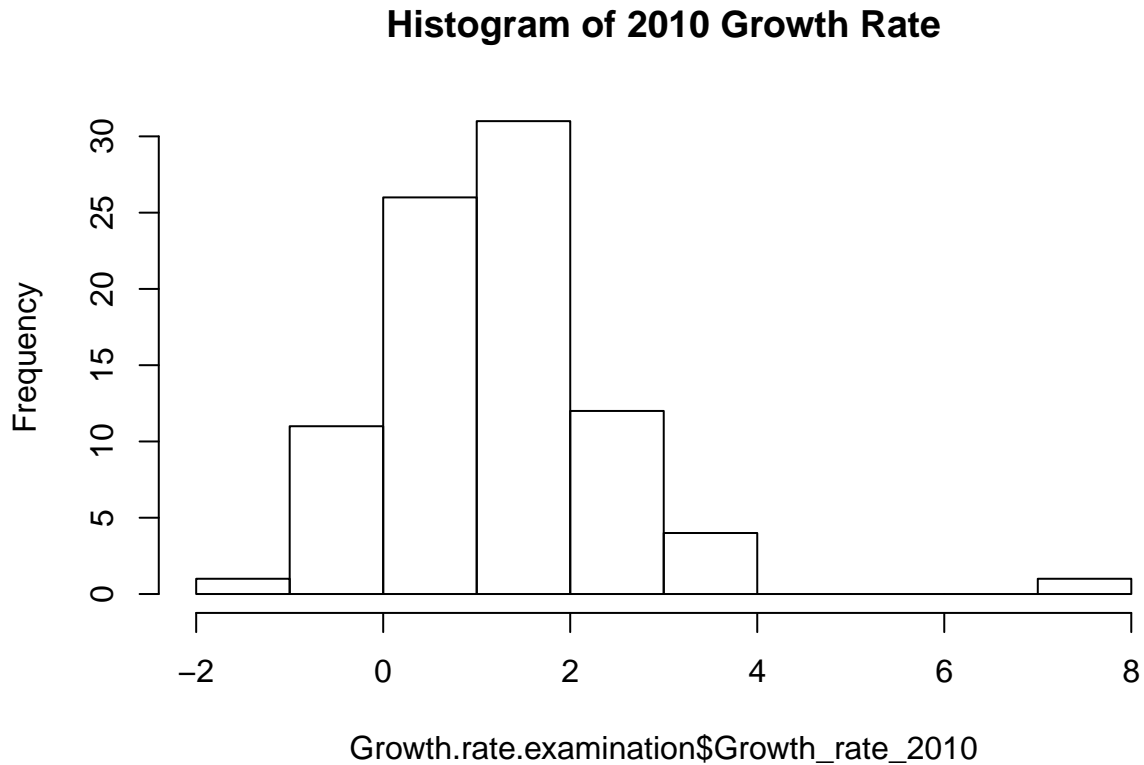
```
hist(Growth.rate.examination$Growth_rate_2000, main = "Histogram of 2000 Growth Rate")
```



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



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



#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

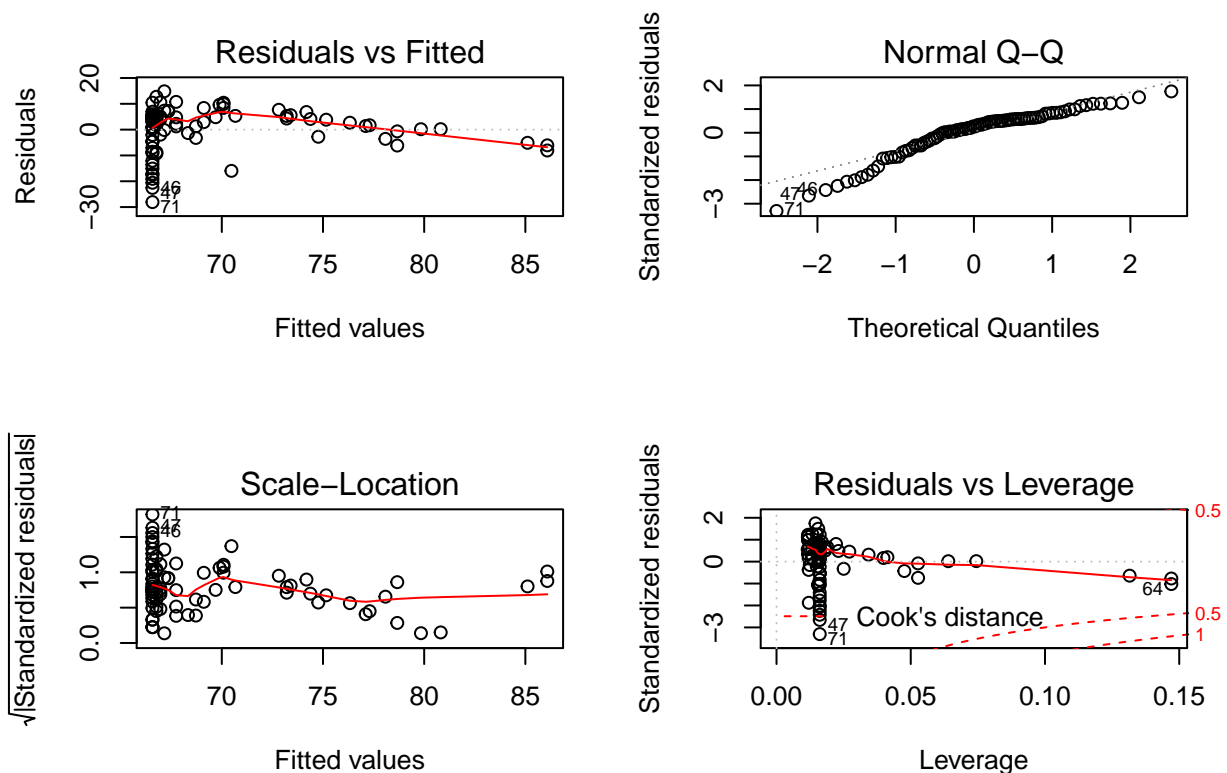
Null Hypotheses: Alcohol/wine consumption has no impact on GNP or life expectancy

Model 1 - parsimonious model - life expectancy ~ wine

```
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
```

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

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.2439961 1.511177 0.022
## Alternative hypothesis: rho != 0
```

```
#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 coefficient 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.birth..years..both.sexes)
```

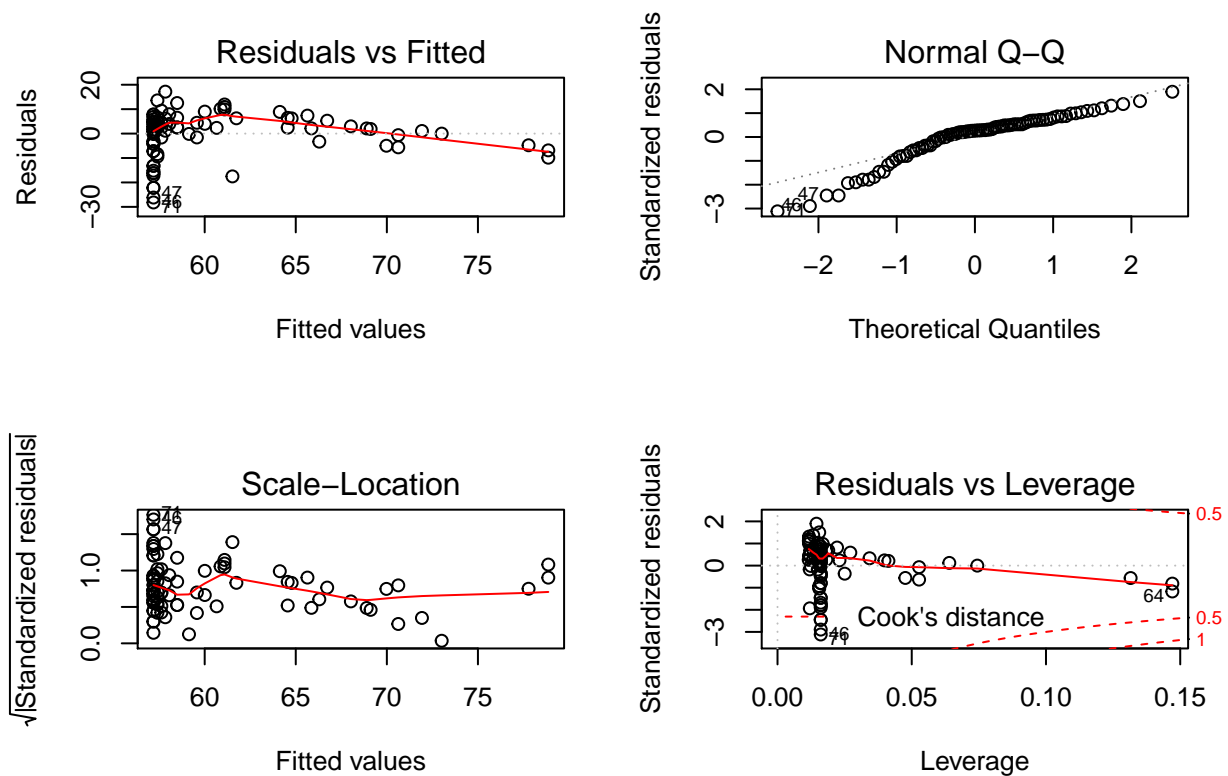
```
## [1] 0.9862559
```

#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.day, data = diet.data)
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. Healthy life expectancy and Life expectancy are consistent.

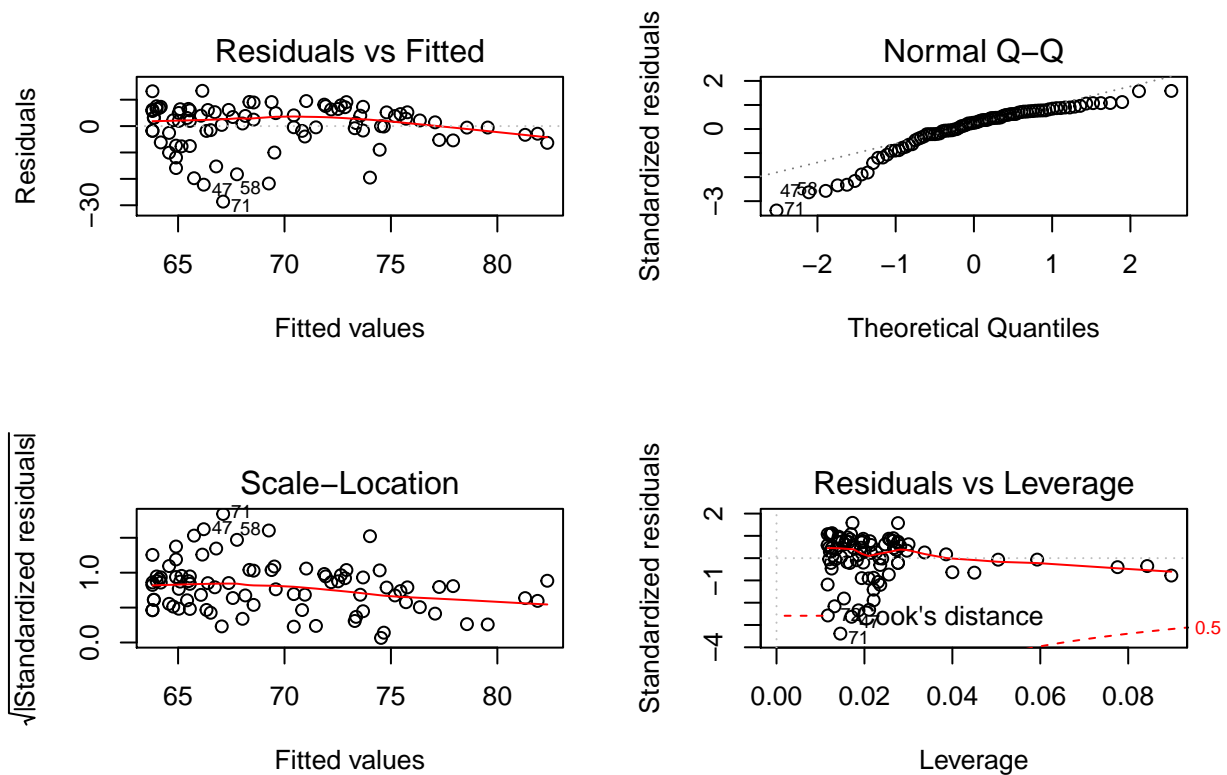
Model 2 - parsimonious model using alcohol consumption- life expectancy ~ alcohol

```
alc.model.1 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Alcoholic.Beverages..kcal.day., data = o
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 0.014
## 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.

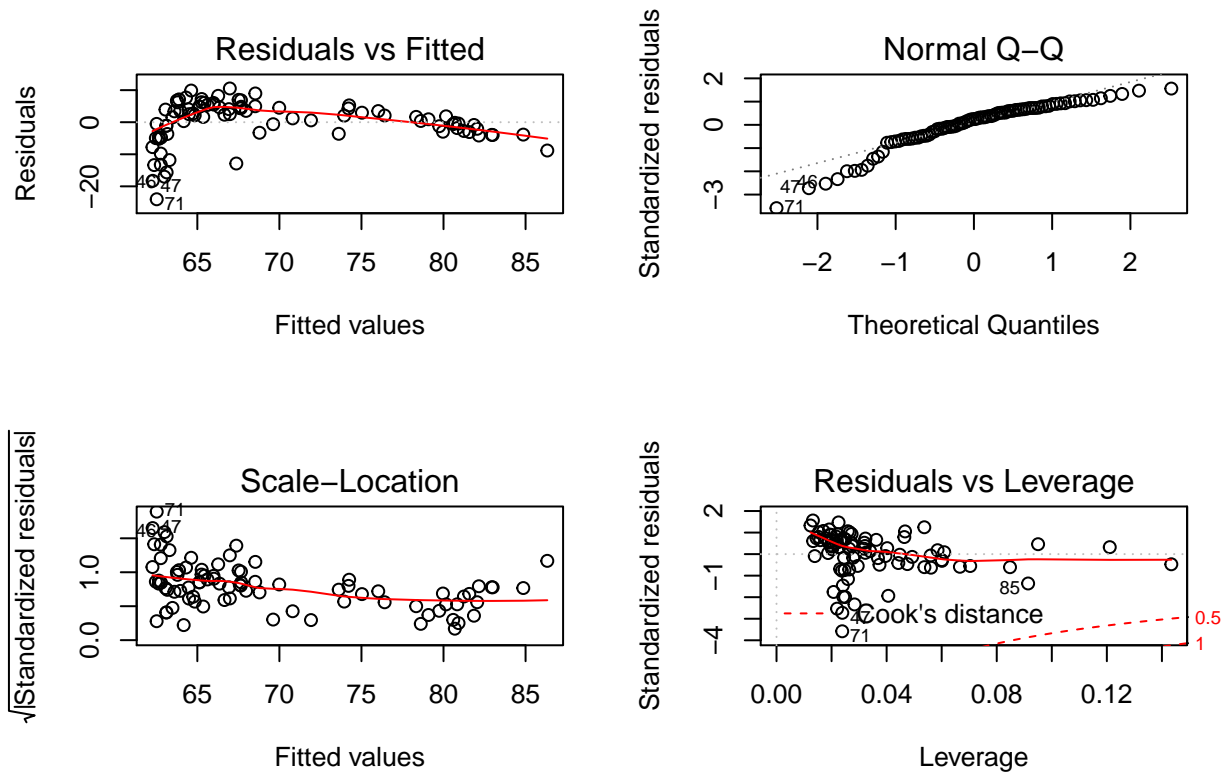
Model 3 - alcohol consumption with control for GNP - life expectancy ~ alcohol + GNP

```
alc.model.2 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Alcoholic.Beverages..kcal.day. + Gross.national.income.per.capita..PPP.international..., data = diet.data)
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
## 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 and of itself may not be as strongly related with life expectancy as previously suspected.

Model 4 - non-linear alcohol consumption with control for GNP - life expectancy ~ alcohol² + alcohol + GNP

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

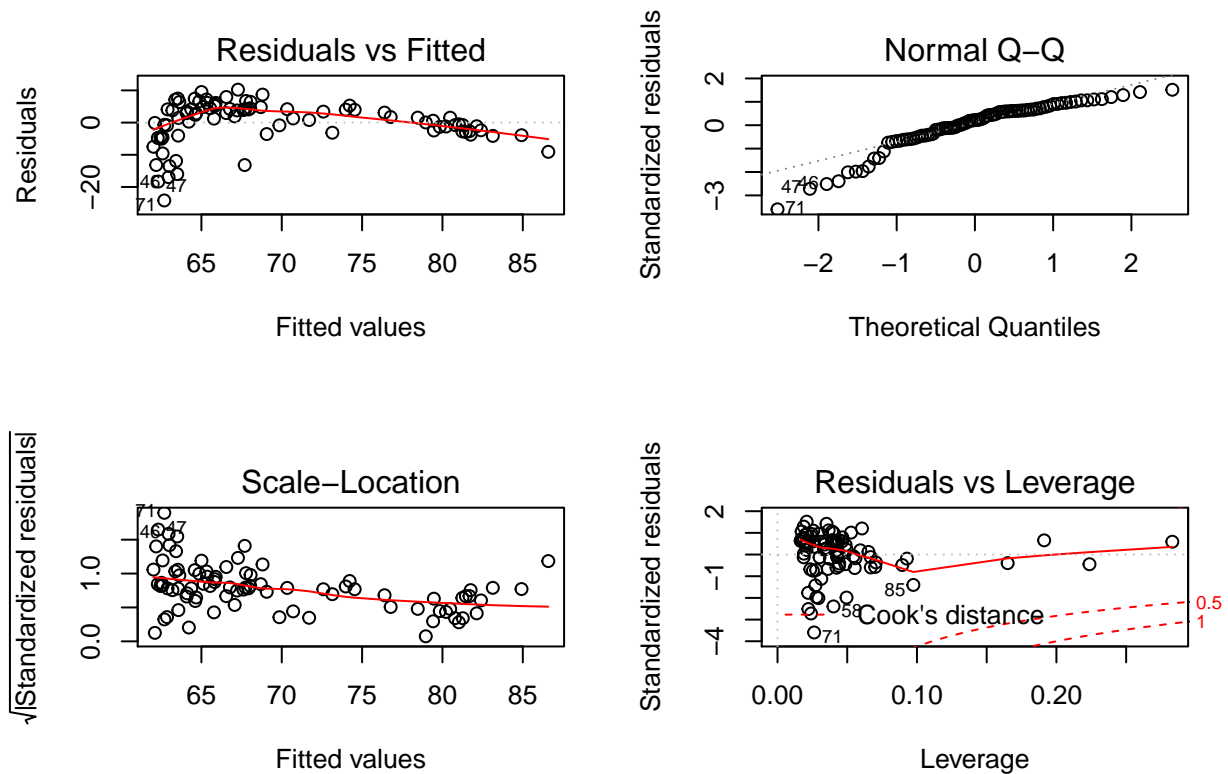
```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ I(Alcoholic.Beverages..kcal.day.^2) +
```

```

##      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
## 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

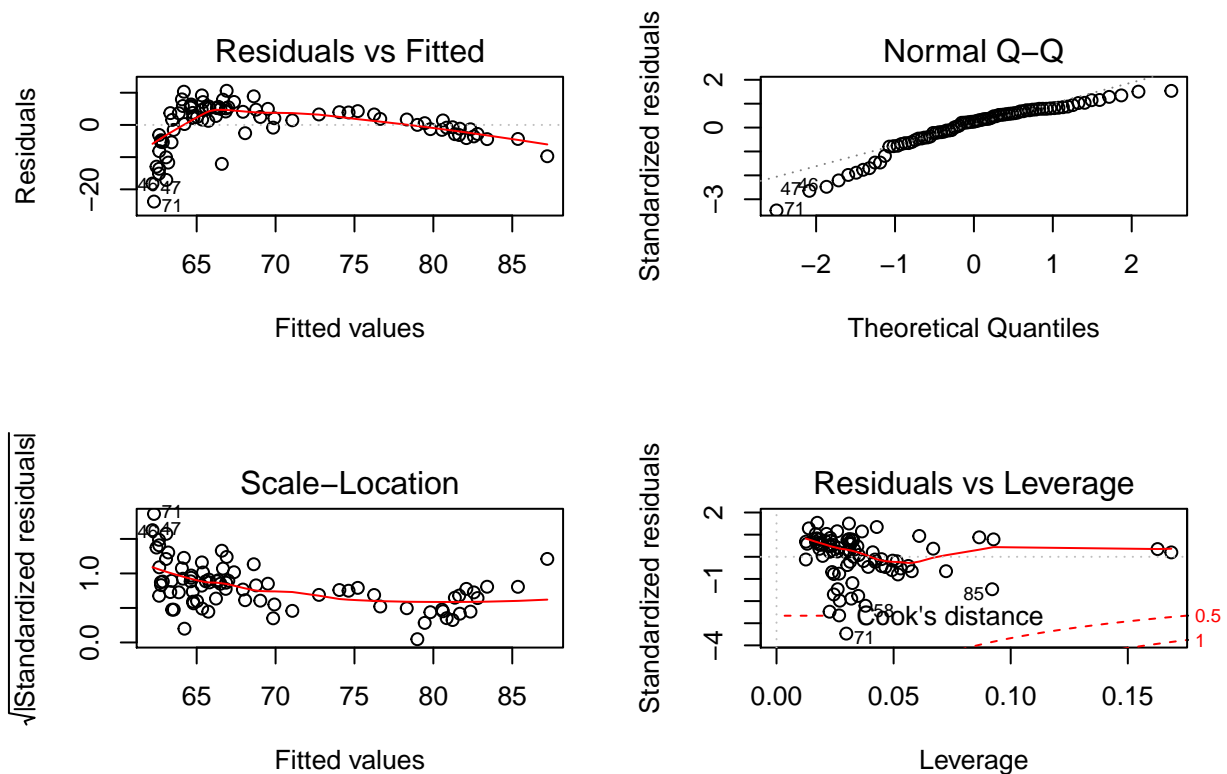
```
#First, remove observations of zero alcoholic beverage consumption so can implement a log transformation
diet.data.2 <- diet.data[diet.data$Alcoholic.Beverages..kcal.day. > 0, ]
```

```
#Estimate the model
alc.model.4 <- lm(Life.expectancy.at.birth..years..both.sexes ~ log(Alcoholic.Beverages..kcal.day.) + G
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 0.004
## 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.

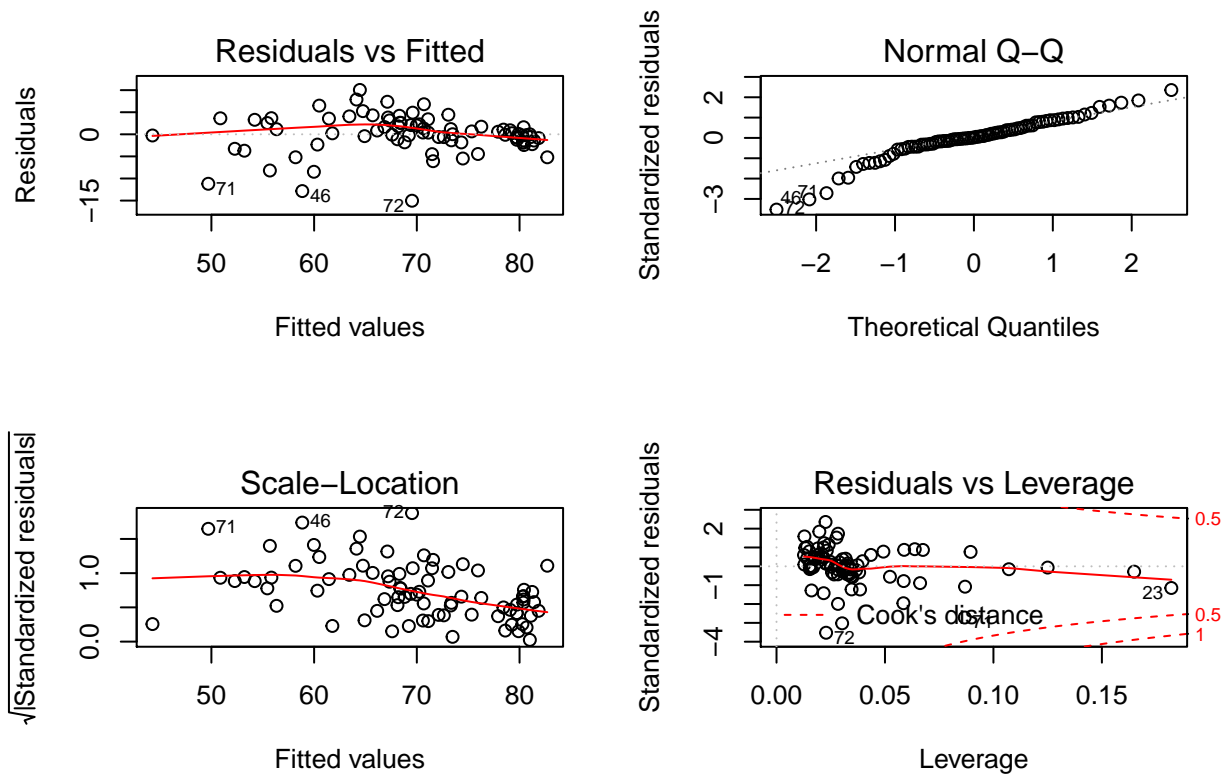
Model 6 - log transformation of alcohol consumption with control for log transformation of GNP - life expectancy \sim log(alcohol) + log(GNP)

#Estimate the model
`alc.model.5 <- lm(Life.expectancy.at.birth..years..both.sexes ~ log(Alcoholic.Beverages..kcal.day.) + log(Gross.national.income.per.capita..PPP.international...), data = diet.data.2)`
`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 0.166
## 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.

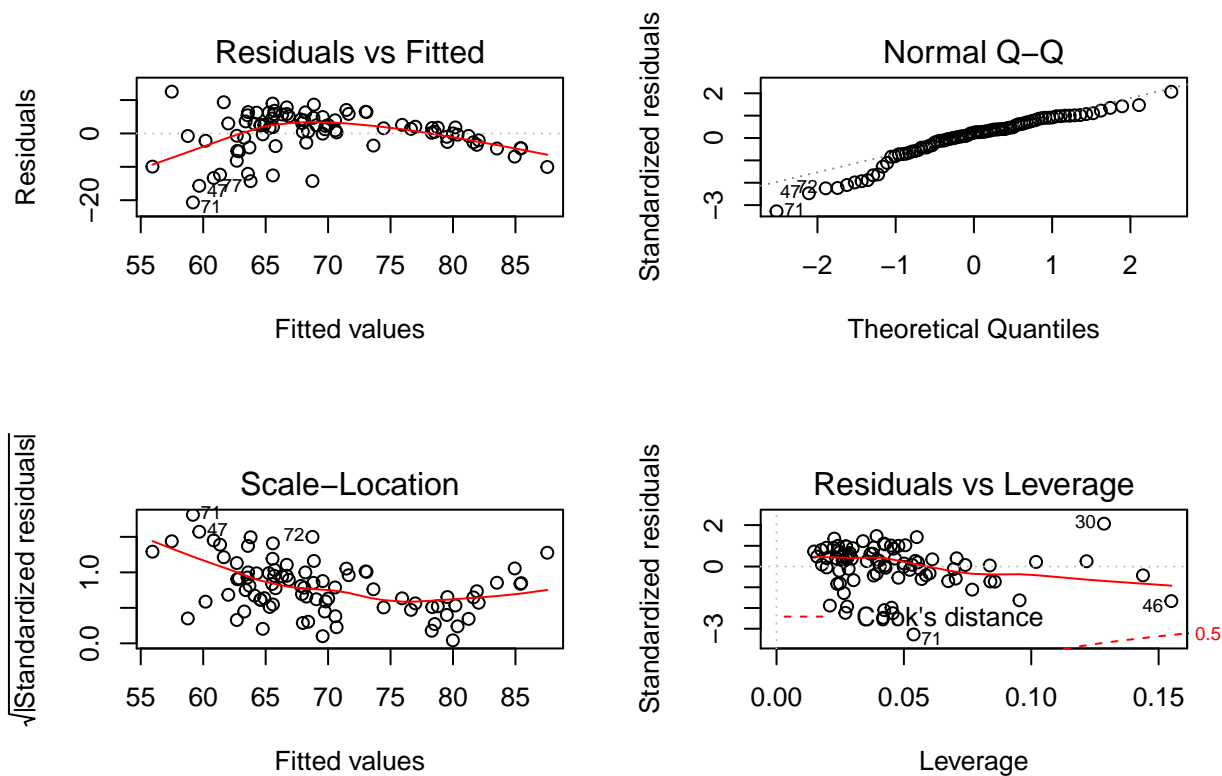
Model 7 - life expectancy ~ GNP + alcohol consumption + blood pressure

```
#Estimate the model
```

```
alc.model.7 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Gross.national.income.per.capita..PPP.int
summary(alc.model.7)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ Gross.national.income.per.capita..PPP.int
##      Alcoholic.Beverages..kcal.day. + Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.,
##      data = diet.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.676  -2.728   1.460   4.403  12.517
##
## Coefficients:
##                                     Estimate
## (Intercept)                        1.151e+02
## Gross.national.income.per.capita..PPP.international... 6.016e-04
## Alcoholic.Beverages..kcal.day.        1.663e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -4.272e-01
##                                     Std. Error
## (Intercept)                        1.788e+01
## Gross.national.income.per.capita..PPP.international... 8.593e-05
## Alcoholic.Beverages..kcal.day.        1.278e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 1.432e-01
##                                     t value
## (Intercept)                        6.440
## Gross.national.income.per.capita..PPP.international... 7.001
## Alcoholic.Beverages..kcal.day.        1.302
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -2.983
##                                     Pr(>|t|)
## (Intercept)                        7.64e-09 ***
## Gross.national.income.per.capita..PPP.international... 6.38e-10 ***
## Alcoholic.Beverages..kcal.day.        0.19669
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 0.00376 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.486 on 82 degrees of freedom
## Multiple R-squared:  0.5777, Adjusted R-squared:  0.5622
## F-statistic: 37.39 on 3 and 82 DF,  p-value: 2.505e-15
```

```
par(mfrow=c(2,2))
plot(alc.model.7)
```



```
bptest(alc.model.7)
```

```
##
## studentized Breusch-Pagan test
##
## data: alc.model.7
## BP = 13.906, df = 3, p-value = 0.003036
```

```
durbinWatsonTest(alc.model.7)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.2402217 1.505994 0.026
## Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.7, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##
## Estimate
## (Intercept) 1.1514e+02
## Gross.national.income.per.capita..PPP.international... 6.0163e-04
## Alcoholic.Beverages..kcal.day. 1.6632e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -4.2717e-01
## Std. Error
## (Intercept) 2.4185e+01
```

```
## Gross.national.income.per.capita..PPP.international... 8.3041e-05
## Alcoholic.Beverages..kcal.day. 1.1489e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 1.9869e-01
## t value
## (Intercept) 4.7608
## Gross.national.income.per.capita..PPP.international... 7.2449
## Alcoholic.Beverages..kcal.day. 1.4476
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -2.1499
## Pr(>|t|)
## (Intercept) 8.193e-06
## Gross.national.income.per.capita..PPP.international... 2.141e-10
## Alcoholic.Beverages..kcal.day. 0.15153
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 0.03451
##
## (Intercept) ***
## Gross.national.income.per.capita..PPP.international... ***
## Alcoholic.Beverages..kcal.day.
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We still see significant issues with diagnostic plots (one reason for using heteroskedastically robust standard errors).

Including a control variable for life expectancy (blood pressure), we see it is statistically significant (as might be expected; with heteroskedasticity robust standard errors, p-value = 0.03451) and with an estimate of -4.27 implies that there is a negative relationship between blood pressure and life expectancy as seen in the multivariate EDA (for every mmHg increase in blood pressure we see a decrease of 4.27 years on life expectancy).

We see alcohol consumption still statistically insignificant.

Model 8 - life expectancy ~ GNP + alcohol consumption + blood pressure + obesity

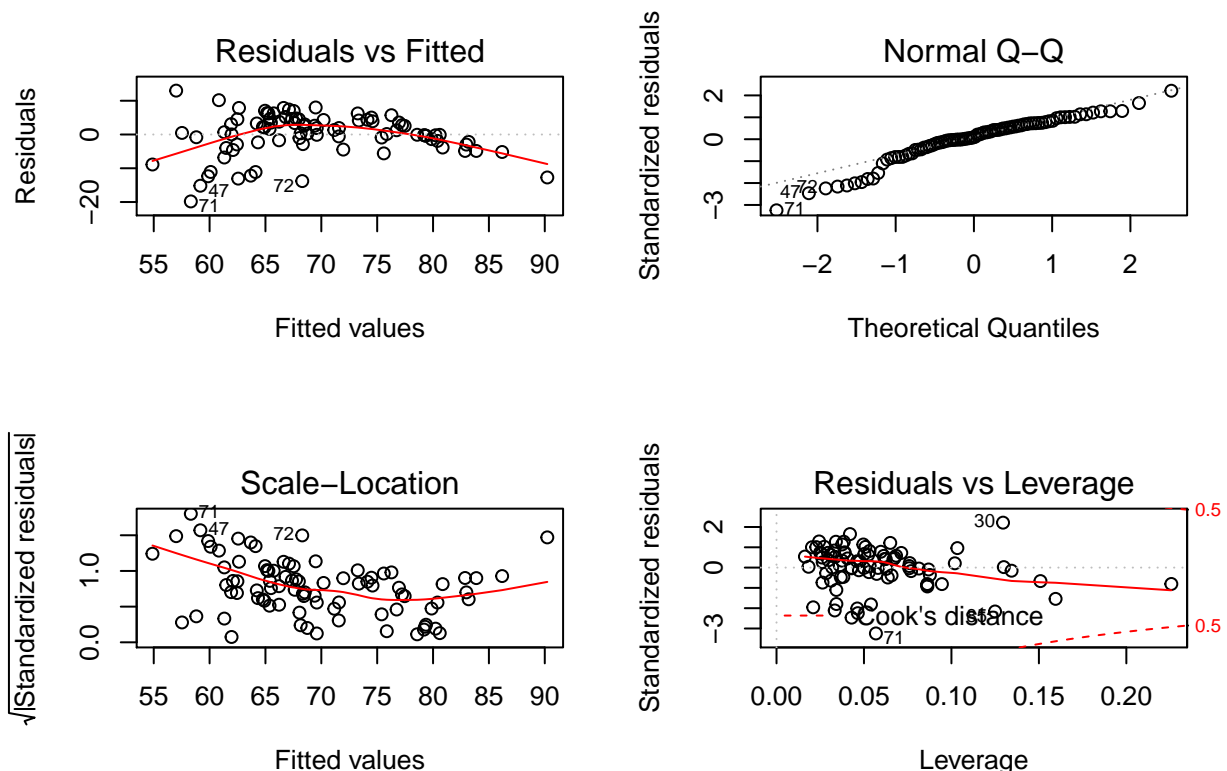
```
#Estimate the model
alc.model.8 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Gross.national.income.per.capita..PPP.international... +
Alcoholic.Beverages..kcal.day. + Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. +
Obesity.prevalence..men...., data = diet.data)
summary(alc.model.8)

##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ Gross.national.income.per.capita..PPP.international... +
##   Alcoholic.Beverages..kcal.day. + Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. +
##   Obesity.prevalence..men...., data = diet.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8210  -2.7219   0.5756   4.2596  12.9919
##
## Coefficients:
##              Estimate
## (Intercept)  1.136e+02
## Gross.national.income.per.capita..PPP.international... 4.899e-04
## Alcoholic.Beverages..kcal.day. 1.784e-02
```



```
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -4.261e-01
## Obesity.prevalence..men.... 2.405e-01
## Std. Error
## (Intercept) 1.736e+01
## Gross.national.income.per.capita..PPP.international... 9.479e-05
## Alcoholic.Beverages..kcal.day. 1.241e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 1.389e-01
## Obesity.prevalence..men.... 9.715e-02
## t value
## (Intercept) 6.548
## Gross.national.income.per.capita..PPP.international... 5.168
## Alcoholic.Beverages..kcal.day. 1.438
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -3.068
## Obesity.prevalence..men.... 2.476
## Pr(>|t|)
## (Intercept) 4.95e-09 ***
## Gross.national.income.per.capita..PPP.international... 1.66e-06 ***
## Alcoholic.Beverages..kcal.day. 0.15419
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 0.00293 **
## Obesity.prevalence..men.... 0.01538 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.292 on 81 degrees of freedom
## Multiple R-squared:  0.6074, Adjusted R-squared:  0.588
## F-statistic: 31.33 on 4 and 81 DF,  p-value: 9.202e-16
```

```
par(mfrow=c(2,2))
plot(alc.model.8)
```



```
bptest(alc.model.8)
```

```
##
## studentized Breusch-Pagan test
##
## data: alc.model.8
## BP = 12.089, df = 4, p-value = 0.0167
```

```
durbinWatsonTest(alc.model.8)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.2145628 1.564849 0.046
## Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.8, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##
## Estimate
## (Intercept) 1.1364e+02
## Gross.national.income.per.capita..PPP.international... 4.8992e-04
## Alcoholic.Beverages..kcal.day. 1.7844e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -4.2615e-01
## Obesity.prevalence..men.... 2.4051e-01
## Std. Error
## (Intercept) 2.3025e+01
## Gross.national.income.per.capita..PPP.international... 7.9915e-05
## Alcoholic.Beverages..kcal.day. 1.1568e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 1.8998e-01
## Obesity.prevalence..men.... 8.5790e-02
## t value
## (Intercept) 4.9357
## Gross.national.income.per.capita..PPP.international... 6.1305
## Alcoholic.Beverages..kcal.day. 1.5425
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -2.2431
## Obesity.prevalence..men.... 2.8034
## Pr(>|t|)
## (Intercept) 4.202e-06
## Gross.national.income.per.capita..PPP.international... 3.027e-08
## Alcoholic.Beverages..kcal.day. 0.126849
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 0.027623
## Obesity.prevalence..men.... 0.006325
##
## (Intercept) ***
## Gross.national.income.per.capita..PPP.international... ***
## Alcoholic.Beverages..kcal.day.
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. *
## Obesity.prevalence..men.... **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As in model 7, adding control variables has not brought significance to alcohol consumption as an estimator of life expectancy.

Obesity (newly added variable) has a p-value of 0.006325. We see that for every percentage increase in a country's prevalence of obesity, there is a 2.4 year increase in life expectancy. This positive relationship was seen in the EDA and it was noted that obesity is likely estimating financial health which in turn is estimating physical health.

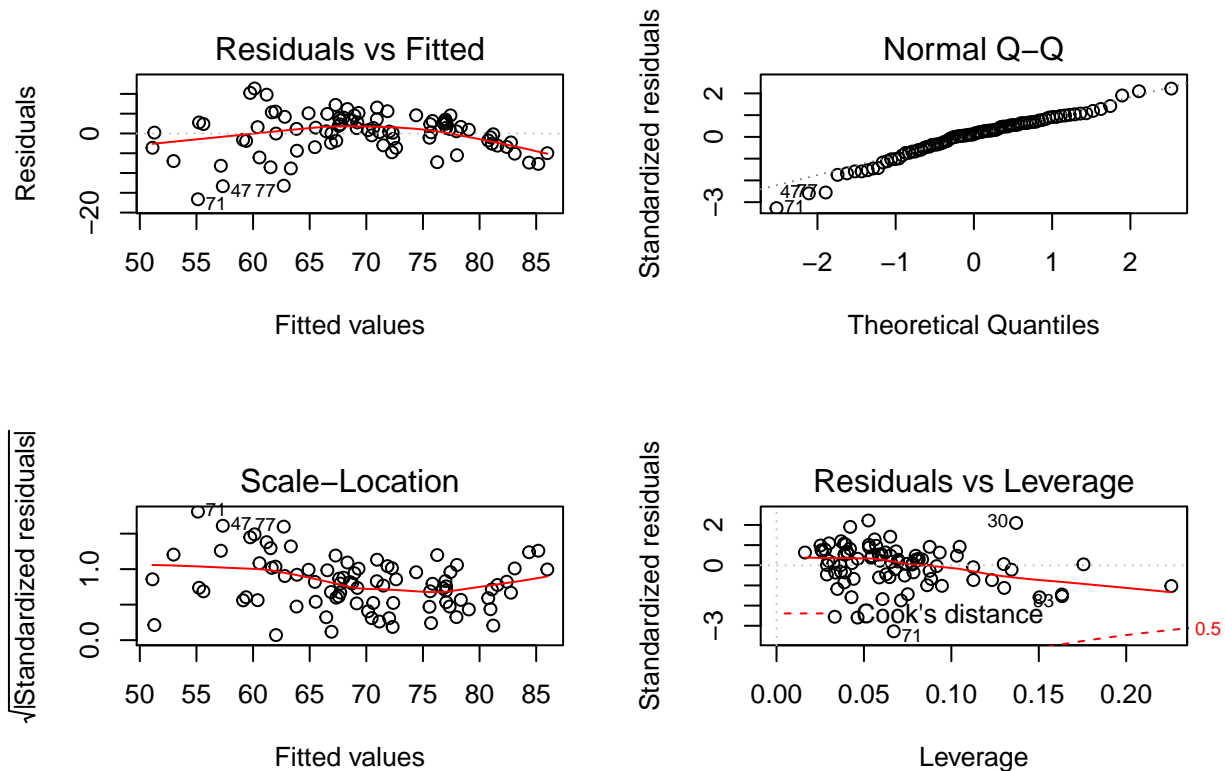
Model 9 - life expectancy ~ GNP + alcohol consumption + blood pressure + obesity + cholesterol

```
#Estimate the model
alc.model.9 <- lm(Life.expectancy.at.birth..years..both.sexes ~ Gross.national.income.per.capita..PPP.int
summary(alc.model.9)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..years..both.sexes ~ Gross.national.income.per.capita..PPP.int
##      Alcoholic.Beverages..kcal.day. + Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. +
##      Obesity.prevalence..men.... + Mean.total.cholesterol..men..mg.dl...2005,
##      data = diet.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6328  -2.8966   0.6919   3.2939  11.3671
##
## Coefficients:
##                                     Estimate
## (Intercept)                        7.431e+01
## Gross.national.income.per.capita..PPP.international... 4.195e-04
## Alcoholic.Beverages..kcal.day.      8.090e-03
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -3.817e-01
## Obesity.prevalence..men....         9.854e-02
## Mean.total.cholesterol..men..mg.dl...2005      1.894e-01
##                                     Std. Error
## (Intercept)                        1.592e+01
## Gross.national.income.per.capita..PPP.international... 8.008e-05
## Alcoholic.Beverages..kcal.day.      1.049e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 1.163e-01
## Obesity.prevalence..men....         8.456e-02
## Mean.total.cholesterol..men..mg.dl...2005      3.158e-02
##                                     t value
## (Intercept)                        4.669
## Gross.national.income.per.capita..PPP.international... 5.239
## Alcoholic.Beverages..kcal.day.      0.771
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -3.282
## Obesity.prevalence..men....         1.165
## Mean.total.cholesterol..men..mg.dl...2005      5.999
##                                     Pr(>|t|)
## (Intercept)                        1.20e-05 ***
## Gross.national.income.per.capita..PPP.international... 1.28e-06 ***
## Alcoholic.Beverages..kcal.day.      0.44302
```

```
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 0.00153 **
## Obesity.prevalence..men.... 0.24739
## Mean.total.cholesterol..men..mg.dl...2005 5.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.258 on 80 degrees of freedom
## Multiple R-squared:  0.7292, Adjusted R-squared:  0.7123
## F-statistic: 43.08 on 5 and 80 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(alc.model.9)
```



```
bptest(alc.model.9)
```

```
##
## studentized Breusch-Pagan test
##
## data: alc.model.9
## BP = 12.042, df = 5, p-value = 0.03422
```

```
durbinWatsonTest(alc.model.9)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.1167135 1.764049 0.256
## Alternative hypothesis: rho != 0
```

```
#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan
coeftest(alc.model.9, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                     Estimate
## (Intercept)                        7.4314e+01
## Gross.national.income.per.capita..PPP.international... 4.1951e-04
## Alcoholic.Beverages..kcal.day.      8.0903e-03
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -3.8173e-01
## Obesity.prevalence..men....          9.8536e-02
## Mean.total.cholesterol..men..mg.dl...2005          1.8944e-01
##                                     Std. Error
## (Intercept)                        1.9918e+01
## Gross.national.income.per.capita..PPP.international... 7.0795e-05
## Alcoholic.Beverages..kcal.day.      1.0649e-02
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 1.5635e-01
## Obesity.prevalence..men....          7.6468e-02
## Mean.total.cholesterol..men..mg.dl...2005          3.2314e-02
##                                     t value
## (Intercept)                        3.7309
## Gross.national.income.per.capita..PPP.international... 5.9258
## Alcoholic.Beverages..kcal.day.      0.7597
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. -2.4416
## Obesity.prevalence..men....          1.2886
## Mean.total.cholesterol..men..mg.dl...2005          5.8624
##                                     Pr(>|t|)
## (Intercept)                        0.0003552
## Gross.national.income.per.capita..PPP.international... 7.479e-08
## Alcoholic.Beverages..kcal.day.      0.4496639
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 0.0168286
## Obesity.prevalence..men....          0.2012554
## Mean.total.cholesterol..men..mg.dl...2005          9.771e-08
##
## (Intercept)                        ***
## Gross.national.income.per.capita..PPP.international... ***
## Alcoholic.Beverages..kcal.day.
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. *
## Obesity.prevalence..men....
## Mean.total.cholesterol..men..mg.dl...2005          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this final life expectancy model with an additional control variable (cholesterol) we see, interestingly, that obesity is no longer statistically significant, however cholesterol is. On an intuitive level we can appreciate the obvious connection between cholesterol and obesity.

In interpreting the cholesterol estimator, we would say that for every mg/dl increase in cholesterol there is a 0.19 year increase in life expectancy.

Also, note that the diagnostic plots look noticeably better (specifically the QQ and scale-location plots).

Model 10 - GNP ~ alcohol consumption (simple/parsimonious model)

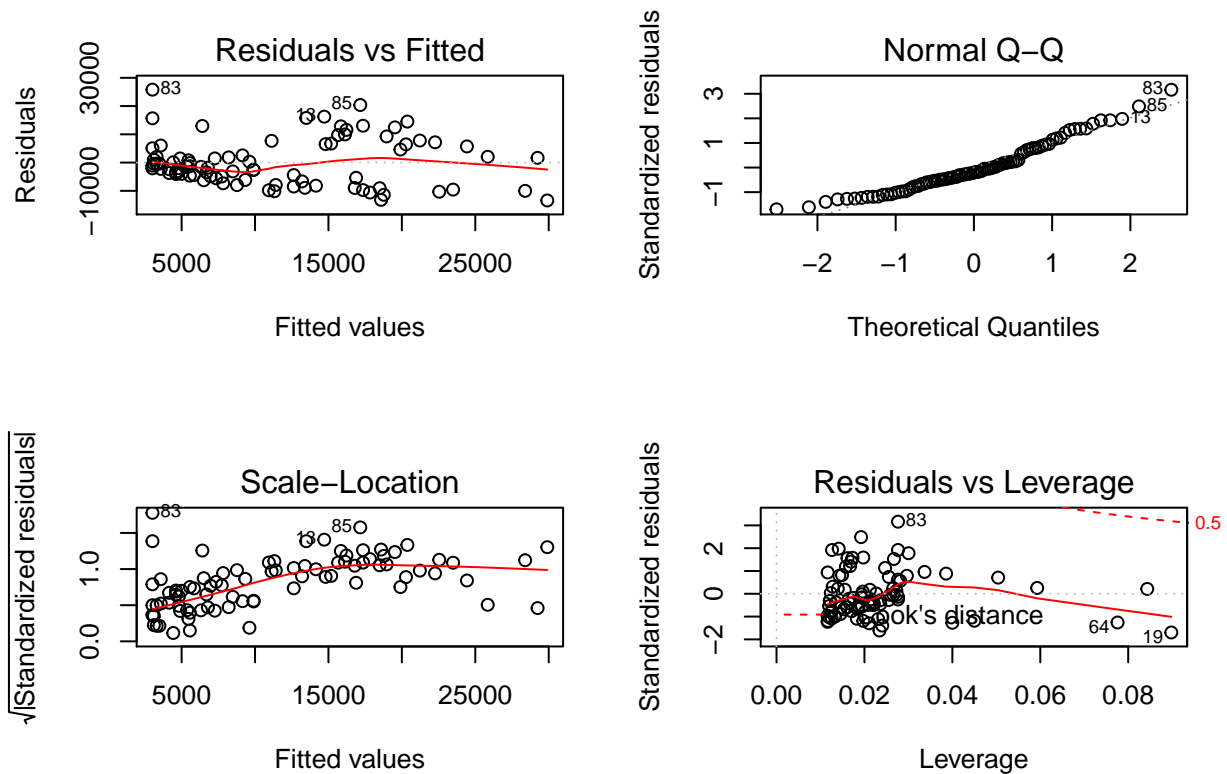
(note GNP is outcome variable here)

```
#Estimate the model
```

```
alc.model.10 <- lm(Gross.national.income.per.capita..PPP.international... ~ Alcoholic.Beverages..kcal.d  
summary(alc.model.10)
```

```
##  
## Call:  
## lm(formula = Gross.national.income.per.capita..PPP.international... ~  
##     Alcoholic.Beverages..kcal.day., data = diet.data)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -13404  -5520  -1667   5559  25804   
##  
## Coefficients:  
##                                Estimate Std. Error t value Pr(>|t|)      
## (Intercept)                   3016.03    1375.58    2.193   0.0311 *    
## Alcoholic.Beverages..kcal.day.    94.34      11.79    8.001 6.03e-12 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 8274 on 84 degrees of freedom  
## Multiple R-squared:  0.4325, Adjusted R-squared:  0.4257   
## F-statistic: 64.02 on 1 and 84 DF,  p-value: 6.035e-12
```

```
par(mfrow=c(2,2))  
plot(alc.model.10)
```



```
bptest(alc.model.10)
```

```
##
## studentized Breusch-Pagan test
##
## data: alc.model.10
## BP = 5.1111, df = 1, p-value = 0.02377
```

```
durbinWatsonTest(alc.model.10)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.09807197 1.801737 0.354
## Alternative hypothesis: rho != 0
```

#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan

```
coeftest(alc.model.10, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3016.029   1180.536   2.5548  0.01243 *
## Alcoholic.Beverages..kcal.day.  94.344    12.844   7.3451 1.219e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With alcohol consumption estimating GNP we see a statistically significant ($p\text{-value} = 1.2e-10$). For every daily calorie increase in alcoholic beverages we see a 94.3 PPP dollar (purchasing power parity) increase.

Robust standard errors were used to help account for concern seen in the diagnostic plots. Based on the scale-location plot and residuals vs. leverage, it may be worth taking log of GNP to see if this helps with heteroskedasticity.

This model illustrates that alcohol consumption is a good indicator of financial health (GNP) as would be expected from intuition and from the response we saw when GNP was added to Model 1 (simple model: life expectancy ~ alcohol consumption).

Model 11 - $\log(\text{GNP}) \sim \text{alcohol consumption} + \text{life expectancy} + \text{blood pressure} + \text{obesity} + \text{cholesterol}$

(note GNP is outcome variable here)

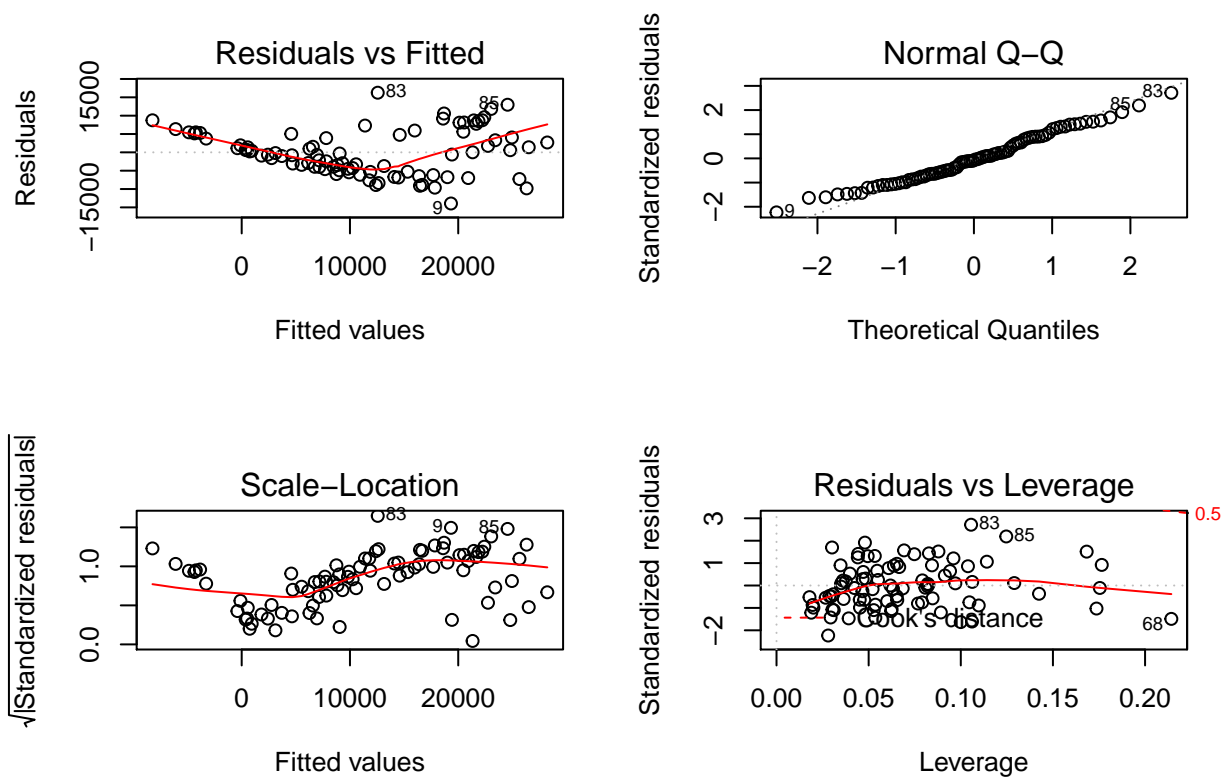
```
#Estimate the model
alc.model.11 <- lm(Gross.national.income.per.capita..PPP.international... ~ Alcoholic.Beverages..kcal.day.
summary(alc.model.11)
```

```
##
## Call:
## lm(formula = Gross.national.income.per.capita..PPP.international... ~
##      Alcoholic.Beverages..kcal.day. + Life.expectancy.at.birth..years..both.sexes +
##      Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. +
##      Obesity.prevalence..men.... + Mean.total.cholesterol..men..mg.dl...2005,
##      data = diet.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13941.1  -4559.0  -454.1   4973.2  16253.6
##
## Coefficients:
##                                     Estimate
## (Intercept)                       -43330.74
## Alcoholic.Beverages..kcal.day.         49.59
## Life.expectancy.at.birth..years..both.sexes    608.85
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.    152.73
## Obesity.prevalence..men....           263.18
## Mean.total.cholesterol..men..mg.dl...2005     -72.30
##                                     Std. Error
## (Intercept)                       21080.93
## Alcoholic.Beverages..kcal.day.        11.41
## Life.expectancy.at.birth..years..both.sexes    116.22
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.    148.29
## Obesity.prevalence..men....           98.43
## Mean.total.cholesterol..men..mg.dl...2005     45.09
##                                     t value
## (Intercept)                       -2.055
## Alcoholic.Beverages..kcal.day.         4.344
## Life.expectancy.at.birth..years..both.sexes    5.239
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.    1.030
## Obesity.prevalence..men....           2.674
## Mean.total.cholesterol..men..mg.dl...2005     -1.604
##                                     Pr(>|t|)
## (Intercept)                       0.04310 *
```



```
## Alcoholic.Beverages..kcal.day. 4.06e-05 ***
## Life.expectancy.at.birth..years..both.sexes 1.28e-06 ***
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 0.30615
## Obesity.prevalence..men.... 0.00909 **
## Mean.total.cholesterol..men..mg.dl...2005 0.11275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6335 on 80 degrees of freedom
## Multiple R-squared:  0.6832, Adjusted R-squared:  0.6634
## F-statistic: 34.51 on 5 and 80 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(alc.model.11)
```



```
bptest(alc.model.11)
```

```
##
## studentized Breusch-Pagan test
##
## data:  alc.model.11
## BP = 15.773, df = 5, p-value = 0.007523
```

```
durbinWatsonTest(alc.model.11)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.102805 1.779824 0.268
## Alternative hypothesis: rho != 0
```

#Look at coefficient estimates with heteroskedasticity robust standard errors because the Breusch-Pagan

```
coeftest(alc.model.11, vcov = vcovHC) # ignore estimate values
```

```
##
## t test of coefficients:
##
##                                     Estimate
## (Intercept)                        -43330.741
## Alcoholic.Beverages..kcal.day.      49.587
## Life.expectancy.at.birth..years..both.sexes  608.849
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.  152.731
## Obesity.prevalence..men....  263.175
## Mean.total.cholesterol..men..mg.dl...2005  -72.305
##                                     Std. Error
## (Intercept)                        20407.796
## Alcoholic.Beverages..kcal.day.      13.312
## Life.expectancy.at.birth..years..both.sexes  132.064
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.  141.851
## Obesity.prevalence..men....  140.182
## Mean.total.cholesterol..men..mg.dl...2005  39.662
##                                     t value
## (Intercept)                        -2.1232
## Alcoholic.Beverages..kcal.day.      3.7250
## Life.expectancy.at.birth..years..both.sexes  4.6102
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.  1.0767
## Obesity.prevalence..men....  1.8774
## Mean.total.cholesterol..men..mg.dl...2005  -1.8230
##                                     Pr(>|t|)
## (Intercept)                        0.0368242
## Alcoholic.Beverages..kcal.day.      0.0003623
## Life.expectancy.at.birth..years..both.sexes  1.503e-05
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.  0.2848489
## Obesity.prevalence..men....  0.0641102
## Mean.total.cholesterol..men..mg.dl...2005  0.0720374
##
## (Intercept)                        *
## Alcoholic.Beverages..kcal.day.      ***
## Life.expectancy.at.birth..years..both.sexes  ***
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg.
## Obesity.prevalence..men.... .
## Mean.total.cholesterol..men..mg.dl...2005 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coeftest(alc.model.11, vcov = vcovHC))
```

```
##
## t test of coefficients:
##
##                                     Estimate
## (Intercept)                        0.0000e+00
```

```

## Alcoholic.Beverages..kcal.day. 3.4314e+21
## Life.expectancy.at.birth..years..both.sexes 2.6276e+264
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 2.1394e+66
## Obesity.prevalence..men.... 1.9748e+114
## Mean.total.cholesterol..men..mg.dl...2005 3.9665e-32
## Std. Error
## (Intercept) Inf
## Alcoholic.Beverages..kcal.day. 6.0434e+05
## Life.expectancy.at.birth..years..both.sexes 2.2637e+57
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 4.0269e+61
## Obesity.prevalence..men.... 7.5885e+60
## Mean.total.cholesterol..men..mg.dl...2005 1.6793e+17
## t value
## (Intercept) 0.1196
## Alcoholic.Beverages..kcal.day. 41.4726
## Life.expectancy.at.birth..years..both.sexes 100.5085
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 2.9350
## Obesity.prevalence..men.... 6.5364
## Mean.total.cholesterol..men..mg.dl...2005 0.1615
## Pr(>|t|)
## (Intercept) 1.038
## Alcoholic.Beverages..kcal.day. 1.000
## Life.expectancy.at.birth..years..both.sexes 1.000
## Systolic.blood.pressure..adults.aged.15.and.above..men..mmHg. 1.330
## Obesity.prevalence..men.... 1.066
## Mean.total.cholesterol..men..mg.dl...2005 1.075

```

Finally, we looked at GNP as an outcome variable with the variables from Model 9. We see that life expectancy is highly statistically significant ($p\text{-value} = 1.5e-05$) as is alcoholic beverages ($p\text{-value} = 0.00036$).

Note, no other variable is statistically significant.

Also, the estimate for alcoholic beverages has decreased. In Model 10 a 1 calorie daily increase from alcoholic beverages produced a 94.3 PPP increase in GNP while in this current model a 1 calorie daily increase only produces a 49.6 PPP increase in GNP. Still a practically significant large effect.

Concluding remarks

In analyzing any of this data, it is important to remember that little to nothing is known about the origin of this data. Therefore, many large issues could exist with the way the data was gathered that would make the analysis of this data completely invalid for real world application. In a highly likely but less severe problem, it could be that the data for different variables was collected at very different time points (several years apart). Although something like a country's mean blood pressure is probably highly correlated from year to year it still represents a likely significant factor that keeps any conclusions from being causal.

Moving forward, when we look at the models we see that a simple model of life expectancy estimated by alcohol consumption is statistically significant, however any model that includes GNP as an estimator causes alcohol consumption to lose significance. Alternatively, any model we created to estimate GNP has alcohol consumption as a statistically significant estimator. It seems that while the alcohol consumption of a country may be a good indication of its financial health, it is not a good indication of its physical health. However because GNP is a good indicator of life expectancy, alcohol consumption may serve as a rough indirect estimate of a country's physical health. The reason alcohol consumption is predicting GNP so well can only

be speculated at this point. However an obvious reason is that alcohol is not a necessity for life and thus money spent on alcohol may be viewed as disposable income. Any country with a high disposable income seems likely to have a high GNP. In the future it may be interesting to see how other uses of disposable income predict GNP. For instance, countries with large entertainment industries.