STAT 291 – Final Project Report

Danesh Patel

We will import the necessary packages

```
library(moderndive)
library(dplyr)
library(ggplot2)
library(readxl)
library(gapminder)
```

The data is imported below.

```
dat = read_excel("FinalProjectDat.xlsx")
summary(dat)
```

```
##        X            country            population           lifeexp
##  Min.   :  1.0   Length:115         Min.   :   180000   Min.   :32.50
##  1st Qu.: 29.5   Class :character   1st Qu.:  4565000   1st Qu.:61.15
##  Median : 58.0   Mode  :character   Median : 10900000   Median :73.20
##  Mean   : 58.0                      Mean   : 28945287   Mean   :69.87
##  3rd Qu.: 86.5                      3rd Qu.: 32350000   3rd Qu.:78.00
##  Max.   :115.0                      Max.   :309000000   Max.   :82.80
##    childmort          income        gdpcapita      chdperwoman
##  Min.   :  2.620   Min.   :  846   Min.   :  234   Min.   :1.260
##  1st Qu.:  6.705   1st Qu.: 3015   1st Qu.: 1145   1st Qu.:1.875
##  Median : 26.000   Median :10300   Median : 4630   Median :2.620
##  Mean   : 45.547   Mean   :17089   Mean   :12835   Mean   :3.231
##  3rd Qu.: 77.250   3rd Qu.:24450   3rd Qu.:13550   3rd Qu.:4.900
##  Max.   :209.000   Max.   :78200   Max.   :87700   Max.   :7.490
##   healthspend          co2               water           popdensity
##  Min.   :  11.9   Min.   : 0.0304   Min.   : 67.30   Min.   :   1.75
##  1st Qu.:  51.2   1st Qu.: 0.3720   1st Qu.: 85.85   1st Qu.:  25.95
##  Median : 247.0   Median : 2.1000   Median : 96.30   Median :  72.70
##  Mean   :1143.5   Mean   : 3.8027   Mean   : 91.64   Mean   : 168.37
##  3rd Qu.: 968.5   3rd Qu.: 6.1200   3rd Qu.: 99.55   3rd Qu.: 131.50
##  Max.   :8360.0   Max.   :18.5000   Max.   :100.00   Max.   :7330.00
##      murder          continent              baby2
##  Min.   :    2.43   Length:115         Min.   :0.0000
##  1st Qu.:  128.00   Class :character   1st Qu.:0.0000
##  Median :  382.00   Mode  :character   Median :0.0000
##  Mean   : 1975.78                      Mean   :0.3478
##  3rd Qu.:  994.00                      3rd Qu.:1.0000
##  Max.   :57500.00                      Max.   :1.0000
```

Choosing A Model

We are building a model for predicting the life expectancy in 2009 (denoted by the variable `lifeexp` in the dataset) using other numerical and/or categorical predictors. The variables that we have selected for the model are

- `water`: percentage of individuals that use basic water services
- `healthspend`: health spending by government
- `chdperwoman`: average number of children birthed by each woman

We chose these variables by a mixture of their adequacy in a statistical regression model and also by using common sense and prior knowledge to see which variables would logically impact the life expectancy.

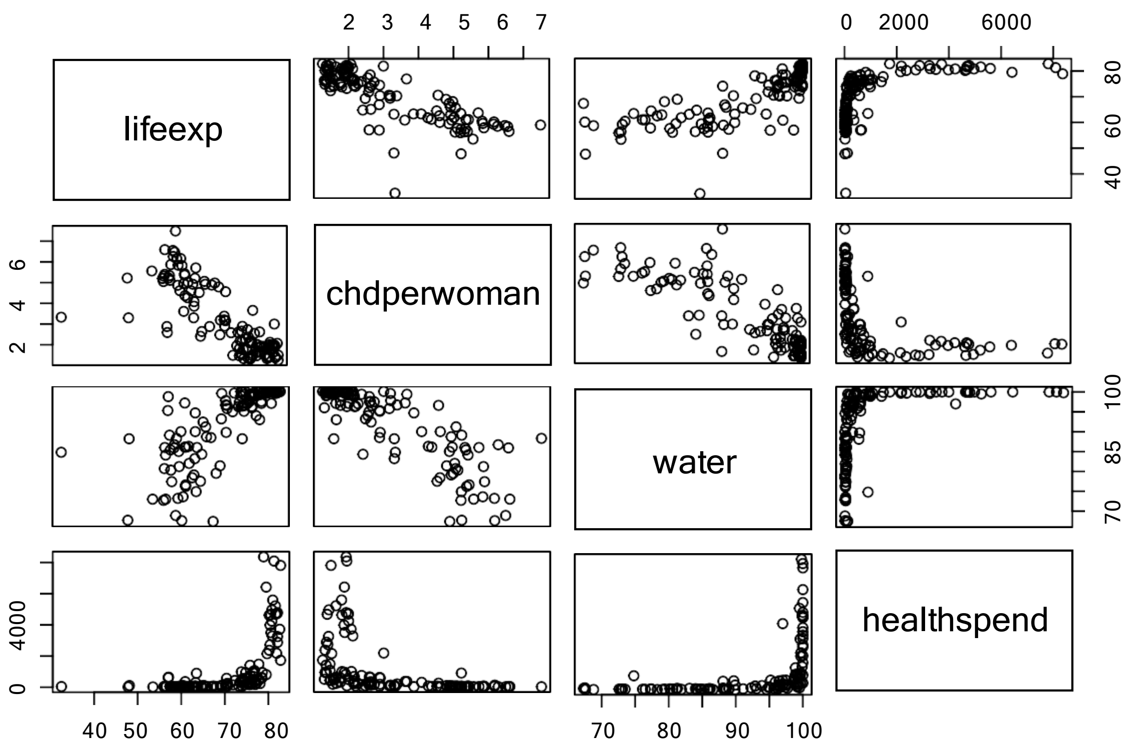
Initial Considerations In Building A Model

At first, we explored the possibility that the only categorical variable of use `continent` would provide any useful information about the life expectancy. However, models fitted with the continent variable were often poor and did not reveal much about the life expectancy. Because of the low significancy of the models, we opted to not use a categorical predictor in our model.

We also explored the effects of interaction terms in the model in search of a better fit. However, higher order interaction terms were deemed to be insignificant with very high p-values. Many two-way interaction terms were not significant either in finding a good model. While some interaction terms were marginally significant, we opted to not include these in the model due to parsimony since no-interaction models provided just as good fits, if not better ones. In this regard, we did not include any interaction terms in the model.

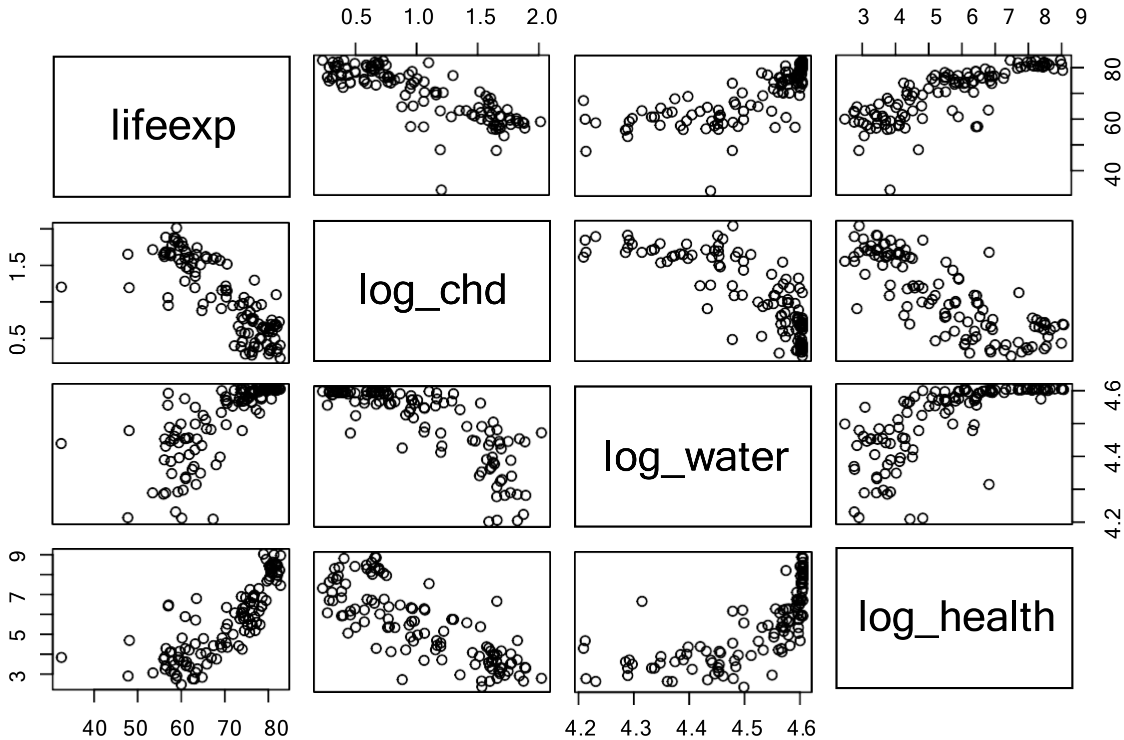
An initial look at the scatterplots reveal to be a nonlinear trend between the variables and the life expectancy. Consider the scatterplot matrix below.

```r
# Data frame that only contains the predictors we want
smallerDat = dat %>%
  select(lifeexp, chdperwoman, water, healthspend)
plot(smallerDat)
```



The output of the plot reveals many non-linear trends in the some of the variables. This also caused problems later on in the analysis when trying to address the distribution of residuals. To remedy this, we applied a logarithmic transformation to the variables. A new scatterplot matrix, using the transformed variables is given below.

```
# Transformed variables
transformedDat = dat %>%
  transmute(lifeexp = lifeexp,
            log_chd = log(chdperwoman),
            log_water = log(water),
            log_health = log(healthspend))
plot(transformedDat)
```



Transforming the data improved the linear relationship between many of the variables with the response. Hence, we decided to keep the logarithmic transformation and use the transformed variables in the model.

Another important consideration in the model was the issue of multicollinearity. This dataset, by its design, features a lot of multicollinearity. The multicollinearity can visually be noticed by looking at the scatterplot matrix and by analyzing the correlation matrix generated below.

```
cor(smallerDat)
```

```
##               lifeexp chdperwoman      water healthspend
## lifeexp     1.0000000  -0.7851377  0.7618874   0.5916694
## chdperwoman-0.7851377   1.0000000 -0.8387799  -0.4882149
## water       0.7618874  -0.8387799  1.0000000   0.4730726
## healthspend 0.5916694  -0.4882149  0.4730726   1.0000000
```

Most of the correlations are moderate to high in this model. While the transformed variables provide a better fit, they do not fix the multicollinearity problem. The correlation matrix for the transformed data is given below.

```
cor(transformedDat)
```

```
##            lifeexp    log_chd  log_water log_health
## lifeexp    1.0000000 -0.7984198  0.7446555  0.8028051
## log_chd   -0.7984198  1.0000000 -0.8022205 -0.7976727
## log_water  0.7446555 -0.8022205  1.0000000  0.7138868
## log_health 0.8028051 -0.7976727  0.7138868  1.0000000
```

Many of the correlations between the transformed variables are high. Unfortunately, this is an intrinsic problem in the dataset as a whole. To see this, we select all numerical data columns and analyze the correlation matrix.

```r
# Get only numerical columns
numericVarsDat = dat %>%
  select(where(is.numeric))

# Correlation matrix
cor(numericVarsDat)
```

```
##                      X  population     lifeexp   childmort      income
## X           1.00000000  0.14155421 0.091253698 -0.11919328  0.08940520
## population  0.14155421  1.00000000 0.053906847 -0.03972533  0.03107100
## lifeexp     0.09125370  0.05390685 1.000000000 -0.91453757  0.72365496
## childmort  -0.11919328 -0.03972533 -0.914537574  1.00000000 -0.63592645
## income      0.08940520  0.03107100 0.723654958 -0.63592645  1.00000000
## gdpcapita   0.07561731  0.02629010 0.638135656 -0.53499870  0.94793487
## chdperwoman-0.09127879 -0.06915419 -0.785137729  0.85963009 -0.64204576
## healthspend 0.06706235  0.13853817 0.591669418 -0.49318906  0.88351574
## co2         0.03163469  0.11828576 0.637048345 -0.61317552  0.85053575
## water       0.08570850  0.07894043 0.761887354 -0.83503964  0.59978903
## popdensity  0.12784809 -0.04040448 0.138157067 -0.10132407  0.32987088
## murder     -0.04150399  0.57280530 0.005740065 -0.05576277 -0.05052643
## baby2       0.01704945  0.07531160 0.650108053-0.61667452  0.68032400
##              gdpcapita chdperwoman healthspend         co2       water
## X           0.07561731 -0.09127879 0.067062349  0.03163469  0.08570850
## population  0.02629010 -0.06915419 0.138538169  0.11828576  0.07894043
## lifeexp     0.63813566 -0.78513773 0.591669418  0.63704834  0.76188735
## childmort  -0.53499870  0.85963009 -0.493189059 -0.61317552 -0.83503964
## income      0.94793487 -0.64204576 0.883515739  0.85053575  0.59978903
## gdpcapita   1.00000000 -0.52771650 0.963353290  0.77828832  0.50740140
## chdperwoman-0.52771650  1.00000000 -0.488214885 -0.61683502 -0.83877994
## healthspend 0.96335329 -0.48821488 1.000000000  0.74974787  0.47307257
## co2         0.77828832 -0.61683502 0.749747872  1.00000000  0.56923505
## water       0.50740140 -0.83877994 0.473072570  0.56923505  1.00000000
## popdensity  0.18425943 -0.13456995 0.040203801  0.16174012  0.10816744
## murder     -0.04893390 -0.08154242 0.002388096  0.00145267  0.09780670
## baby2       0.62806892 -0.70377512 0.616114763  0.61699999  0.55861920
##             popdensity       murder       baby2
## X           0.12784809 -0.041503986  0.01704945
## population -0.04040448  0.572805304  0.07531160
## lifeexp     0.13815707  0.005740065  0.65010805
## childmort  -0.10132407 -0.055762770 -0.61667452
## income      0.32987088 -0.050526434  0.68032400
## gdpcapita   0.18425943 -0.048933897  0.62806892
## chdperwoman-0.13456995 -0.081542424 -0.70377512
## healthspend 0.04020380  0.002388096  0.61611476
## co2         0.16174012  0.001452670  0.61699999
```

```
## water        0.10816744  0.097806698  0.55861920
## popdensity   1.00000000 -0.038595421  0.15070813
## murder      -0.03859542  1.000000000  0.04038717
## baby2        0.15070813  0.040387169  1.00000000
```

Many of the variables have very high correlations and the ones that have low correlations did not provide a good fit for the model and were insignificant in regression models. This made it challenging to completely remove multicollinearity. We opted to go for models and variables that provided better fits, even if it brought extra multicollinearity to the model.

Picking A Model

With the challenges mentioned above, we decided on using a linear model with no interaction terms that utilized the logarithmic transformation of the variables. The final model of choice was fitted below.

```r
model = lm(lifeexp ~ log(water) + log(healthspend) + log(chdperwoman), data = dat)
summary(model)
```

```
##
## Call:
## lm(formula = lifeexp ~ log(water) + log(healthspend) + log(chdperwoman),
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.3100  -1.6450   0.6316   2.6186   8.7289
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -19.3647    34.9280  -0.554  0.58041
## log(water)        18.4072     7.5671   2.433  0.01659 *
## log(healthspend)   2.1872     0.4431   4.936 2.82e-06 ***
## log(chdperwoman)  -5.7608     1.8741  -3.074  0.00266 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.186 on 111 degrees of freedom
## Multiple R-squared:  0.7277,  Adjusted R-squared:  0.7203
## F-statistic: 98.87 on 3 and 111 DF,  p-value: < 2.2e-16
```

From the summary output, we see that all the predictors chosen are significant as is the overall model. We also see that the adjusted R-squared is 0.7203, which means that the model explains 72% of the variability in life expectancy. Of the models we tested, this one had a higher R-squared and adjusted R-squared than other models. This was also a factor when we chose the model.

Interpreting The Model

We will interpret the model coefficients below

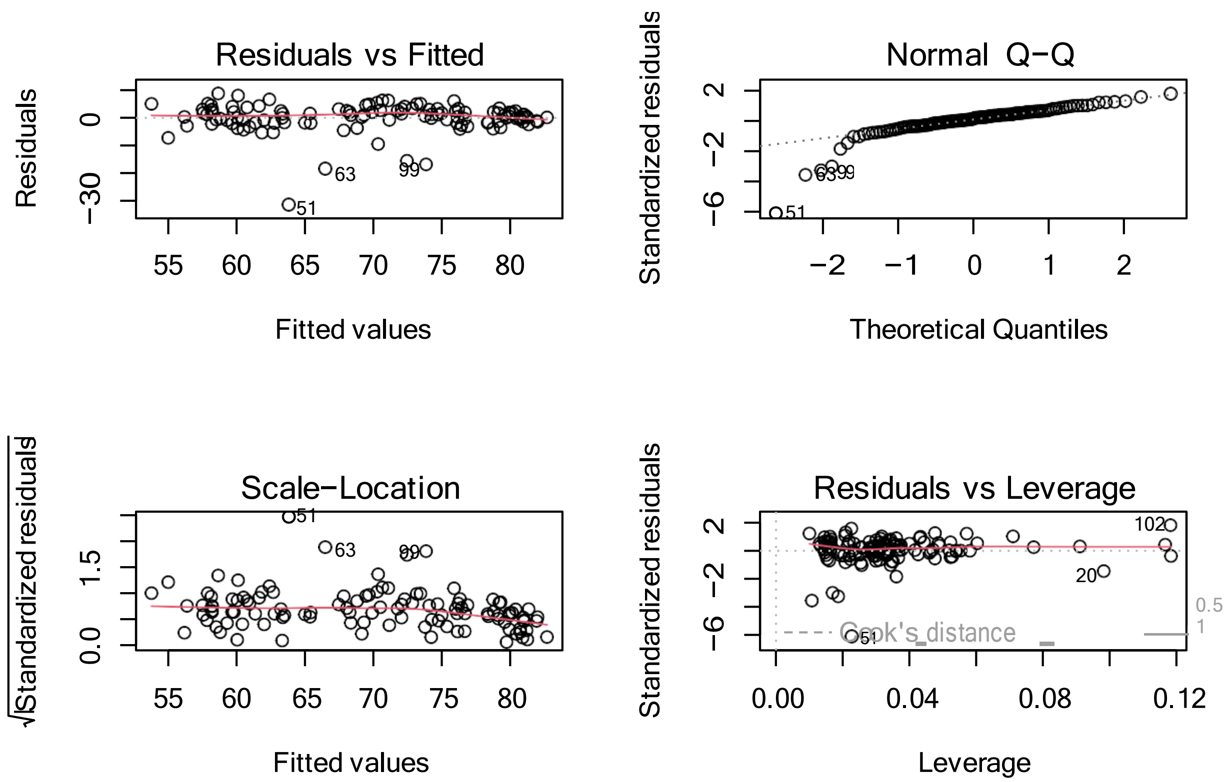
- Slope of `log(water)`: for a one unit increase in `log(water)`, the predicted life expectancy increases by about 18.407 years, if other predicts are held constant. If `log(water)` increases by 1, then the transformed `water` variable would roughly increase by 2.7 units.
- Slope of `log(healthspend)`: for a one unit increase in `log(healthspend)` (or about a 2.7 units increase in `healthspend`), the expected life expectancy increases by about 2.19 years, if other predictors are held constant.

- Slope of `log(chdperwoman)`: for a one unit increase in `log(chdperwoman)` (or about a 2.7 units increase in `chdperwoman`), tje expected life expectancy decreases by 5.76 years, if other predictors are held constant.

Residual Analysis

In order to further assess the usefulness of the model, we will generate residual plots using the diagnostic plots produced automatically by R and using the ggplot package. The default diagnostic plots produced by R are given below (these plots explictly highlight outliers, making it easier to track them down).

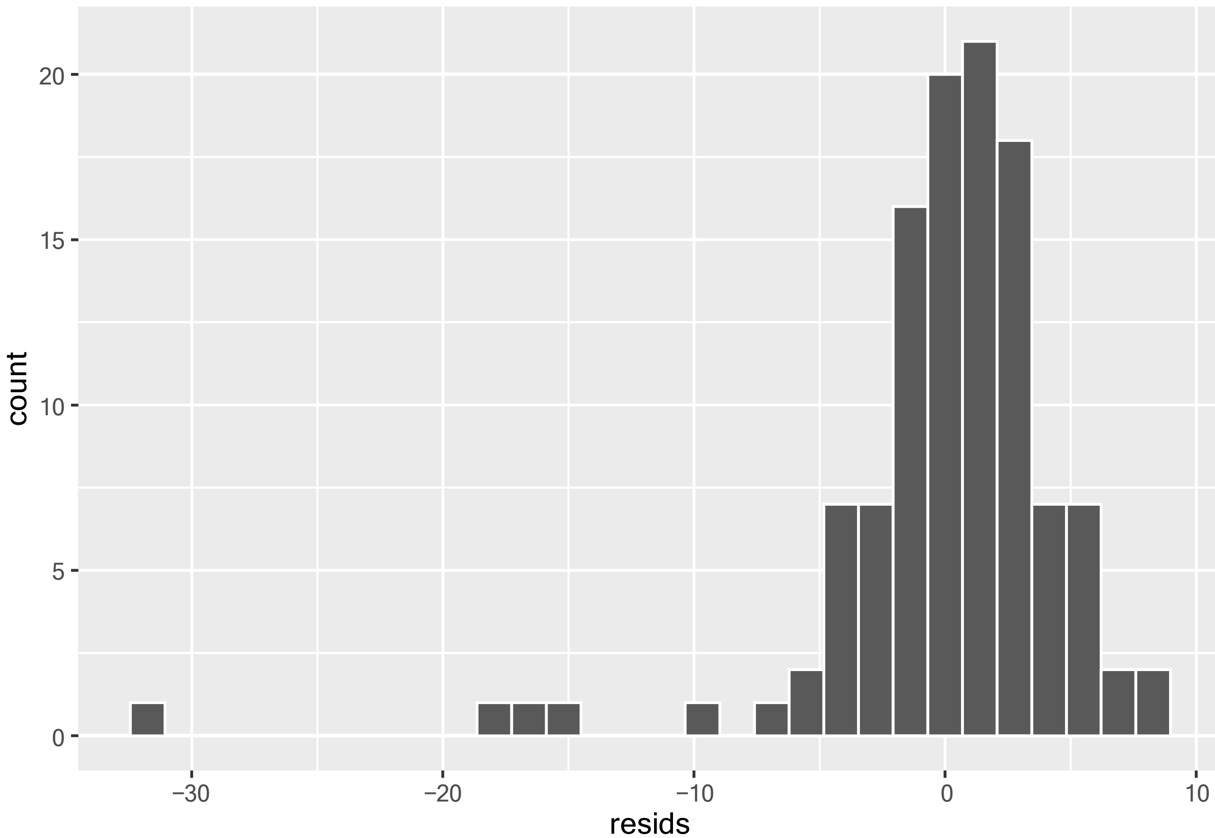
```
# Create a 2 x 2 grid to display all plots in one screen
par(mfrow = c(2, 2))
plot(model)
```



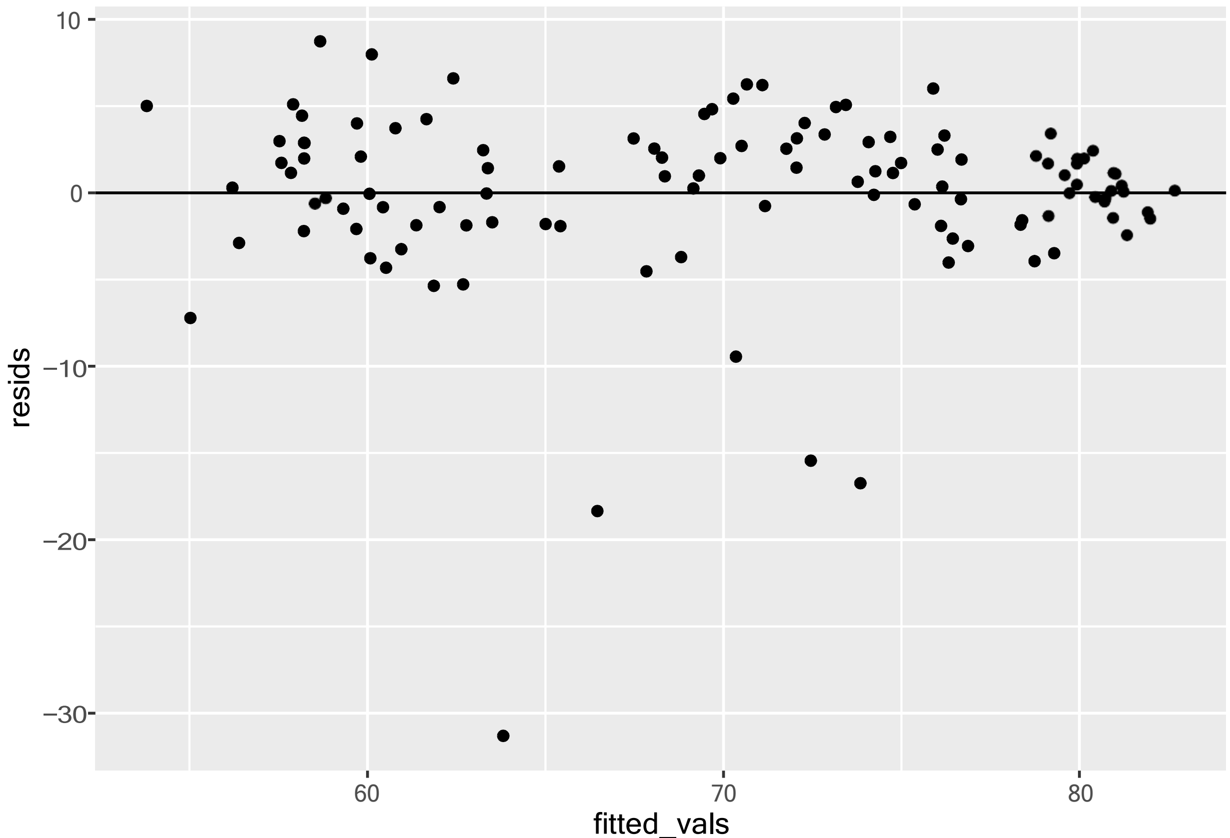
The diagnostic plots produced by R highlight observations with ID 51, 63, 99, 20, and 102 as influential observations/outliers. We will discuss these observations in the next section.

We use the ggplot package to generate more detailed plots, include a residual vs. fitted plot, and histogram and Q-Q plot of residuals.

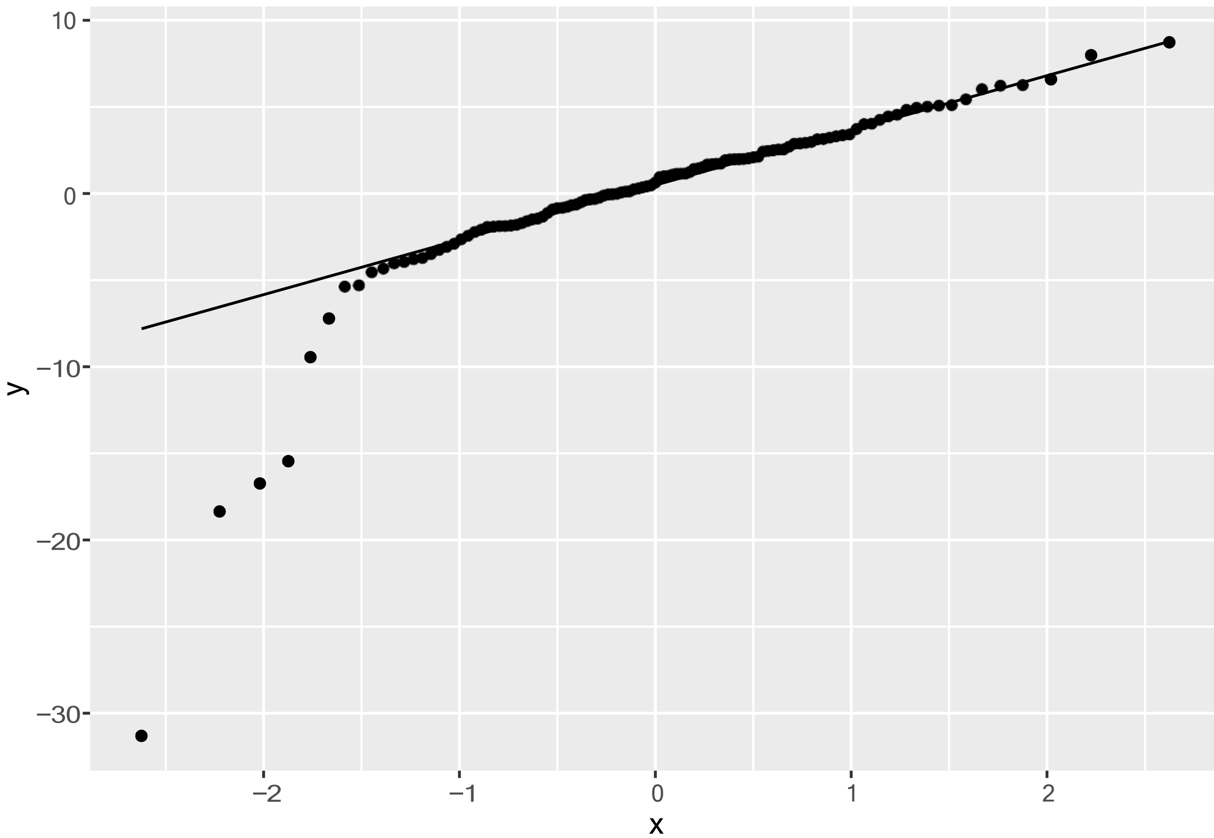
```
# Data frame of residuals and fitted values
modelpoints = data.frame(resids = model$residuals, fitted_vals = model$fitted.values)

# Histogram of residuals
ggplot(data=modelpoints, aes(x=resids)) + geom_histogram(bins=30, color = "white")
```




```r
# Residuals vs. fitted values
ggplot(data=modelpoints, aes(x= fitted_vals, y=resids)) +
  geom_point() +
  geom_hline(yintercept = 0)
```




```
# Q-Q plot of residuals
ggplot(data=modelpoints, aes(sample=resids)) +
  stat_qq() +
  stat_qq_line()
```



From the residual , we see that while most observations on the line, there are some outliers that are signifcantly skewing the distribution. This is supported by the histogram that suggests the residuals are left-skewed. To verify the assumptions, we see if the LINE property is satisfied.

- Linearity: from the residuals vs. fitted plot, it seems that every interval has mean roughly 0, excluding some of the intervals with outliers. Overall, we can conclude that there is likely a linear relationship between the variables.
- Independence: the residual vs. fitted plot indicates no discernible pattern. We can conclude that the observations are independent.
- Normality: this is the only assumption that is not satisfied. The logarithmic transformation helped with this slightly as the model with the untransformed variables were more strongly skewed. Even so, this assumption is the only one that isn't satisfied.
- Equal Variances: from the residual vs. fitted plot, this assumption is satisfied as there is no strong "cone shape" present (excluding outliers).

Outliers

In the previous section, we identified observations 51, 63, 99, 20, and 102 as the outliers/potentially influential observations. We subset these observations into its own data frame.

```
influentialObs = dat[c(20, 51, 63, 99, 102),]
influentialObs
```

```
## # A tibble: 5 x 15
##       X country    popul~1 lifeexp child~2 income gdpca~3 chdpe~4 healt~5co2
##    <dbl> <chr>       <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl> <dbl>
## 1    20 Central A~ 4.39e6    47.8    150     1200     488    5.22    18.2 0.0602
## 2    51 Haiti       9.95e6    32.5    209     2740    1170    3.33    46.4 0.211
```

```
## 3    63 Lesotho      2    e6     48.1   100      2270    1120    3.3     109   1.14
## 4    99 South Afr~  5.12e7   57.1    52.5  12500    7330    2.59    649   9.16
## 5   102 Sudan       3.45e7   67.4    75.9   3220    1490    4.88    83.9 0.417
## # ... with 5 more variables: water <dbl>, popdensity <dbl>, murder <dbl>,
## #   continent <chr>, baby2 <dbl>, and abbreviated variable names 1: population,
## #   2: childmort, 3: gdpcapita, 4: chdperwoman, 5: healthspend
```

We do some external research to try to explain the reason for the outliers in the data. Here is a brief summary of the results.

- Central African Republic: the year 2009 was unusually active for criminal gangs and a lot of political instability, leading to a lot of violence.
- Haiti: also very politically unstable which was made worse by an extremely deadly hurricane season.
- Lesotho: no obvious reason found.
- South Africa: high poverty and crime rates with generally high violence.
- Sudan: war and famine in the region during the time period.

All the information was obtained from Human Rights Watch (HRW) reports on each country for 2009. Most of the situations seem quite extraordinary and would create outliers in life expectancy. Due to these exceptional events, we can generally ignore these outliers as special and extraordinary cases.