Final Project

Lawrence Ho Paul Jung

2024-03-06

```
library(moderndive)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(epiDisplay)
```

```
## Loading required package: foreign
```

```
## Loading required package: survival
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: nnet
```

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
## The following object is masked from 'package:epiDisplay':
##
##     alpha
```

```
library(rockchalk)
```

```
##
## Attaching package: 'rockchalk'
```

```
## The following object is masked from 'package:MASS':
##
##     mvrnorm
```

```
## The following object is masked from 'package:dplyr':
##
##     summarize
```

```
Countries<- read.csv("Countries (1).csv")
glimpse(Countries)
```

```
## Rows: 171
## Columns: 14
## $ X                 <int> 1, 2, 3, 6, 8, 9, 10, 12, 13, 14, 15, 16, 18, 19, 20…
## $ country           <chr> "Afghanistan", "Albania", "Algeria", "Angola", "Anti…
## $ childmortality    <dbl> 88.00, 13.30, 27.40, 120.00, 9.59, 14.40, 18.50, 4.7…
## $ co2capita         <dbl> 0.290, 2.260, 3.280, 1.240, 5.960, 4.170, 1.780, 17.…
## $ fertility         <dbl> 5.82, 1.65, 2.89, 6.16, 2.13, 2.37, 1.55, 1.93, 1.44…
## $ gdpcapita         <chr> "569", "3580", "3930", "2990", "14.4k", "13.6k", "28…
## $ healthspending    <dbl> 37.7, 241.0, 178.0, 123.0, 690.0, 742.0, 134.0, 4780…
## $ income            <dbl> 4.50, 9.77, 9.08, 6.29, 18.40, 23.10, 7.12, 61.10, 6…
## $ lifexpectancy     <dbl> 60.5, 78.1, 74.5, 60.2, 75.9, 75.9, 73.9, 82.1, 80.8…
## $ murder            <chr> "4130", "65.9", "530", "824", "5.05", "2450", "154",…
## $ population        <chr> "28.2M", "2.91M", "35.9M", "23.4M", "85.7k", "41.1M"…
## $ populationdensity <dbl> 43.40, 106.00, 15.10, 18.70, 195.00, 14.70, 104.00, …
## $ wateraccess       <dbl> 48.8, 91.4, 92.3, 50.4, 98.4, 98.4, 98.1, 99.9, 100.…
## $ continent         <chr> "Asia", "Europe", "Africa", "Africa", "Americas", "A…
```

Exploratory data analysis

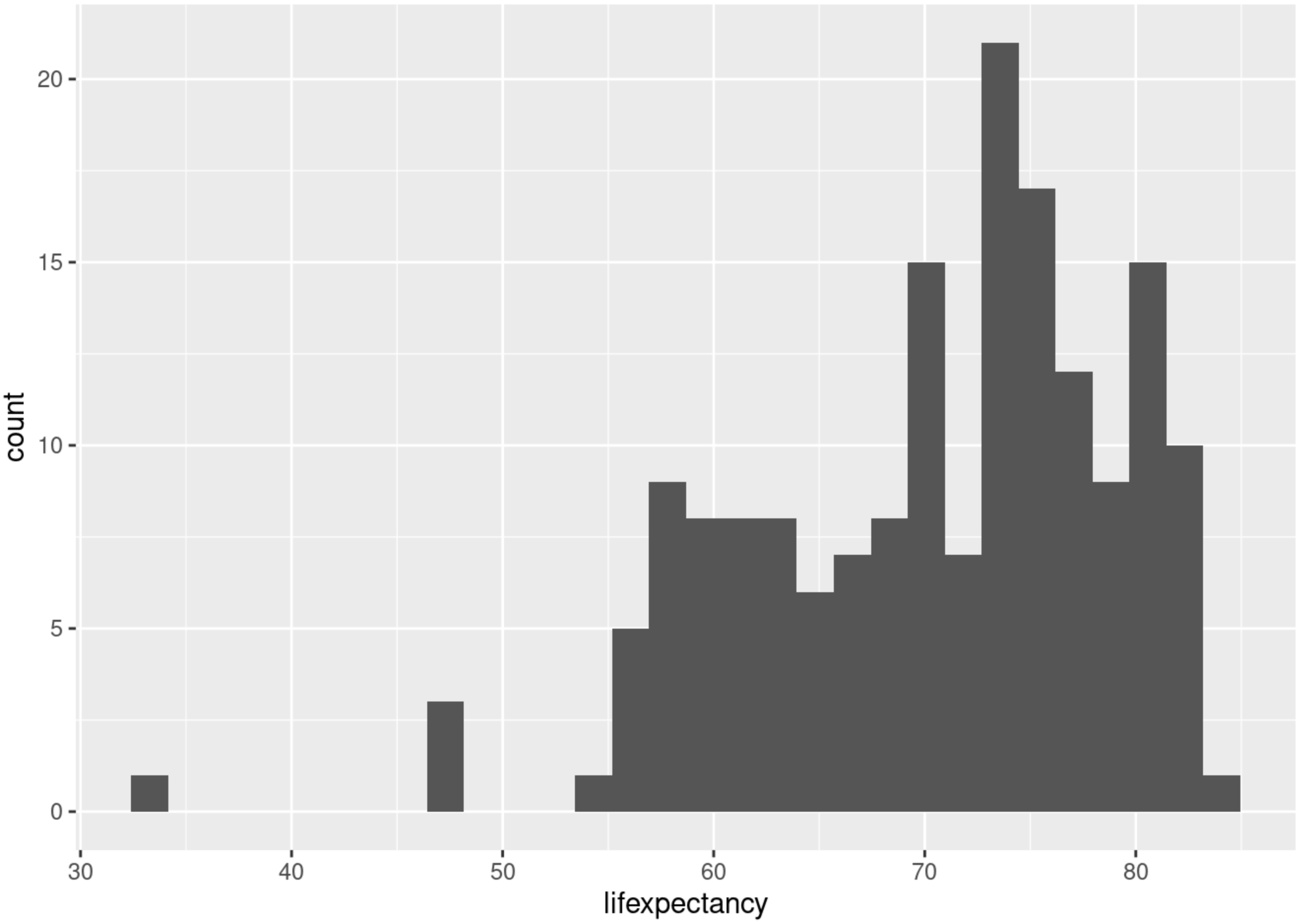
Figure 1 shows that the distribution of life expectancy is right skewed and there are outliers with extremely low life expactancys. NOTE: look into the countries with expectancy lower than 40 years

Figure 2 shows a positive maybe linear relationship but more likely a log relationship.

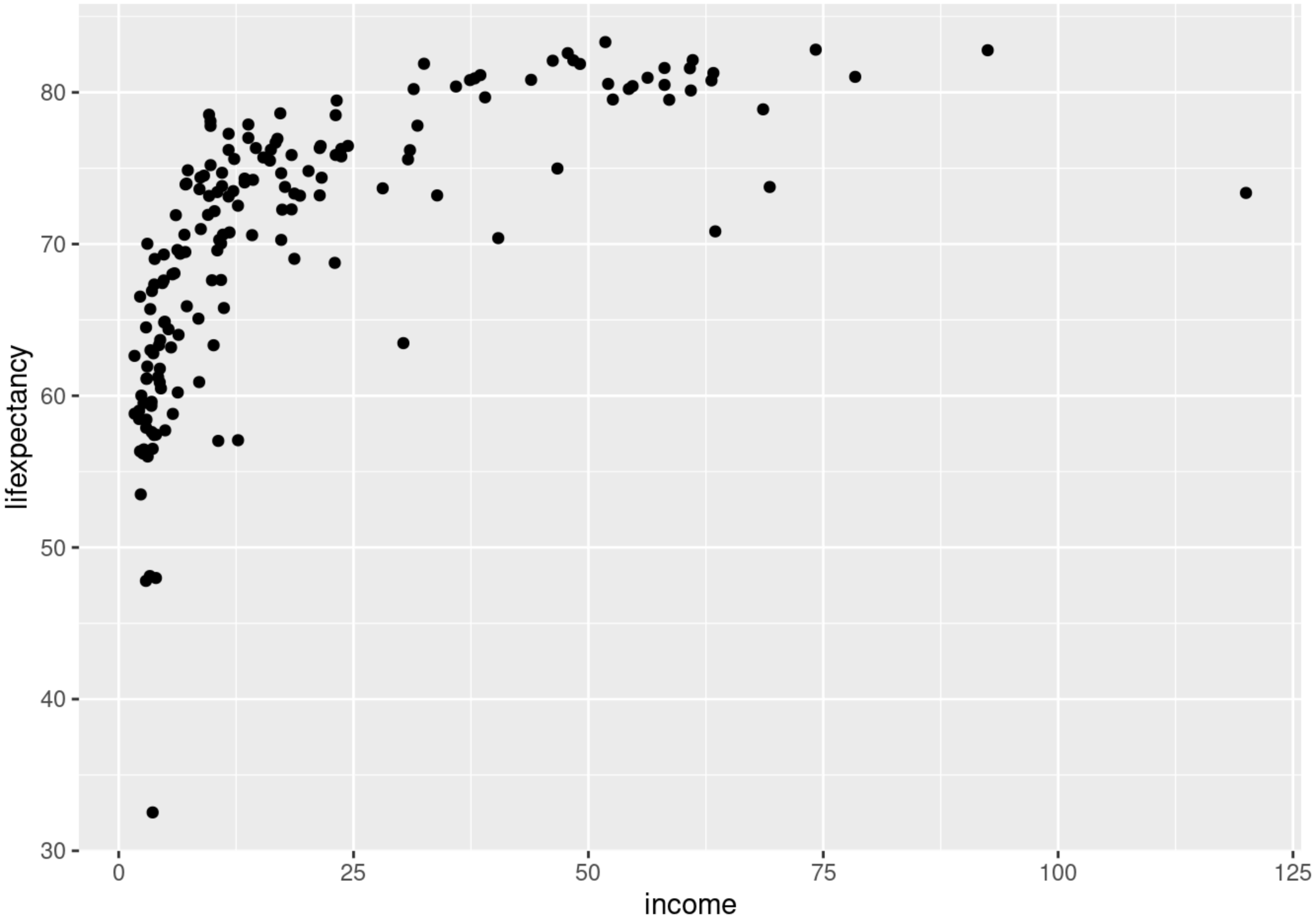
Figure 3 shows another positive relationship that resembles a log curve

```
ggplot(Countries, mapping = aes( x=lifexpectancy, col=))+geom_histogram()
```

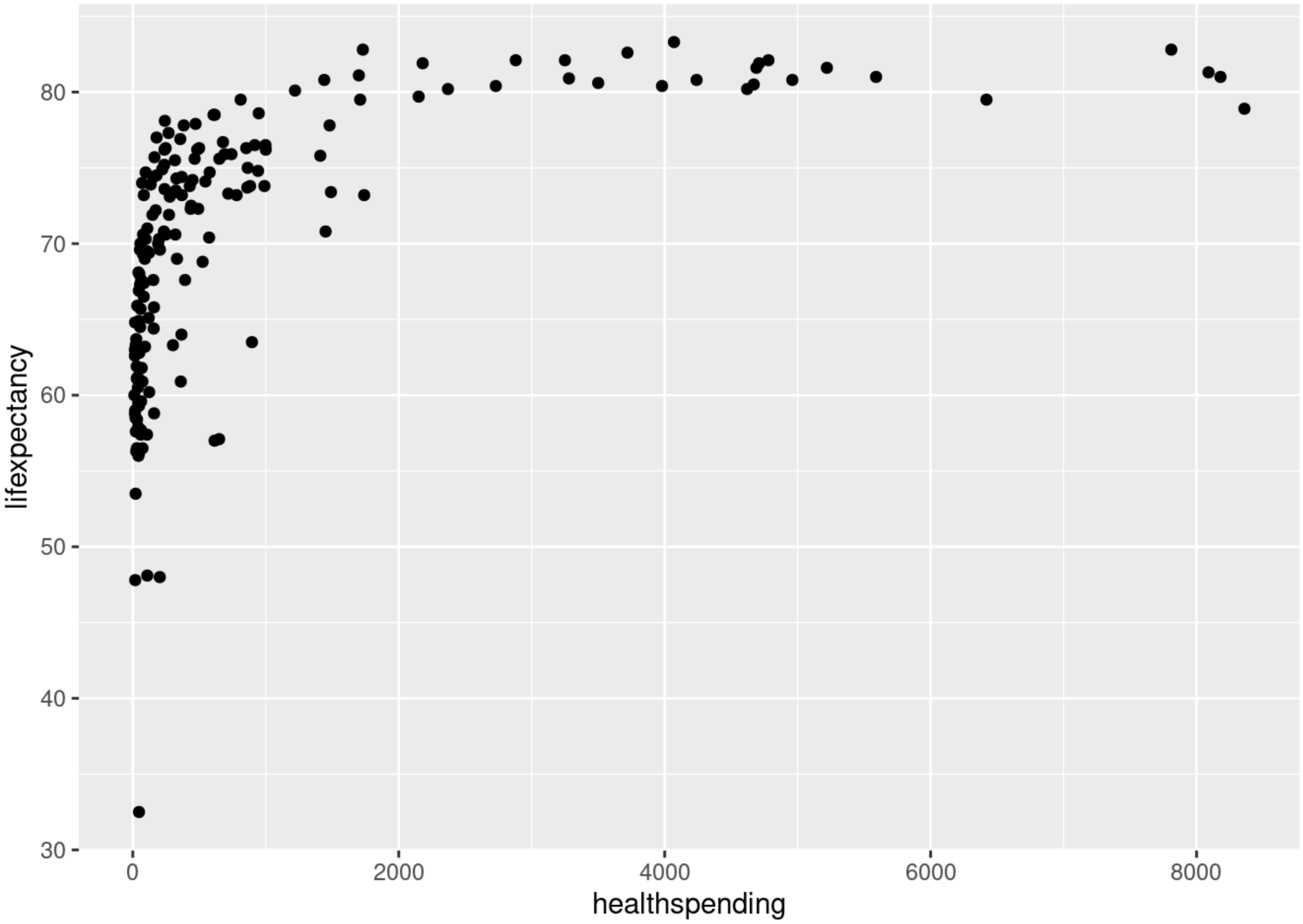
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(Countries, mapping = aes(x=income, y=lifexpectancy))+geom_jitter()
```



```
ggplot(Countries, mapping = aes(x = healthspending, y=lifexpectancy))+geom_point()
```



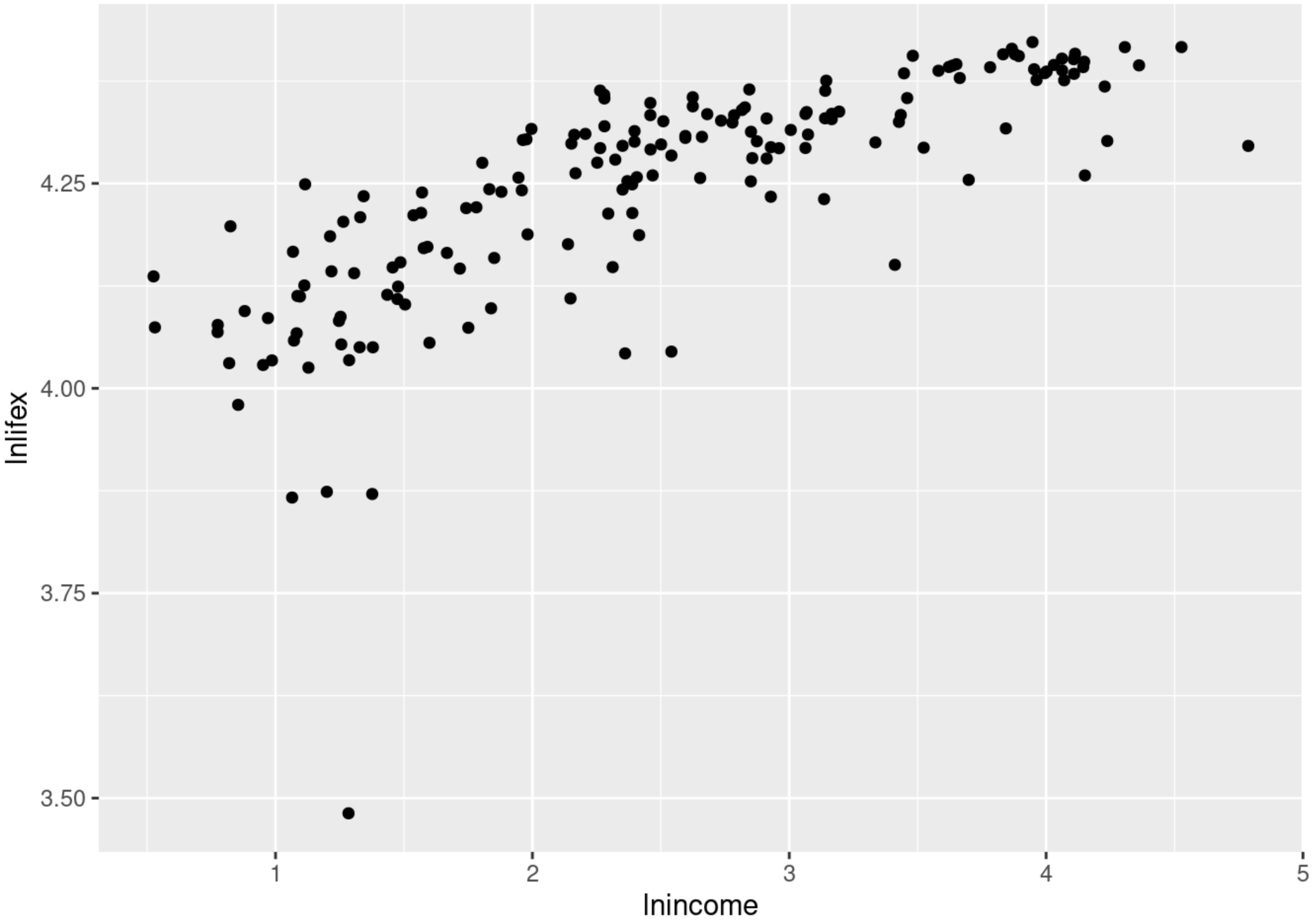
Log Transformation

```
logcountriesmoney<- Countries %>%
  mutate(lnlifex=log(Countries$lifexpectancy)) %>%
  mutate(lnincome=log(Countries$income)) %>%
  mutate(lnhealthspending=log(Countries$healthspending))
```

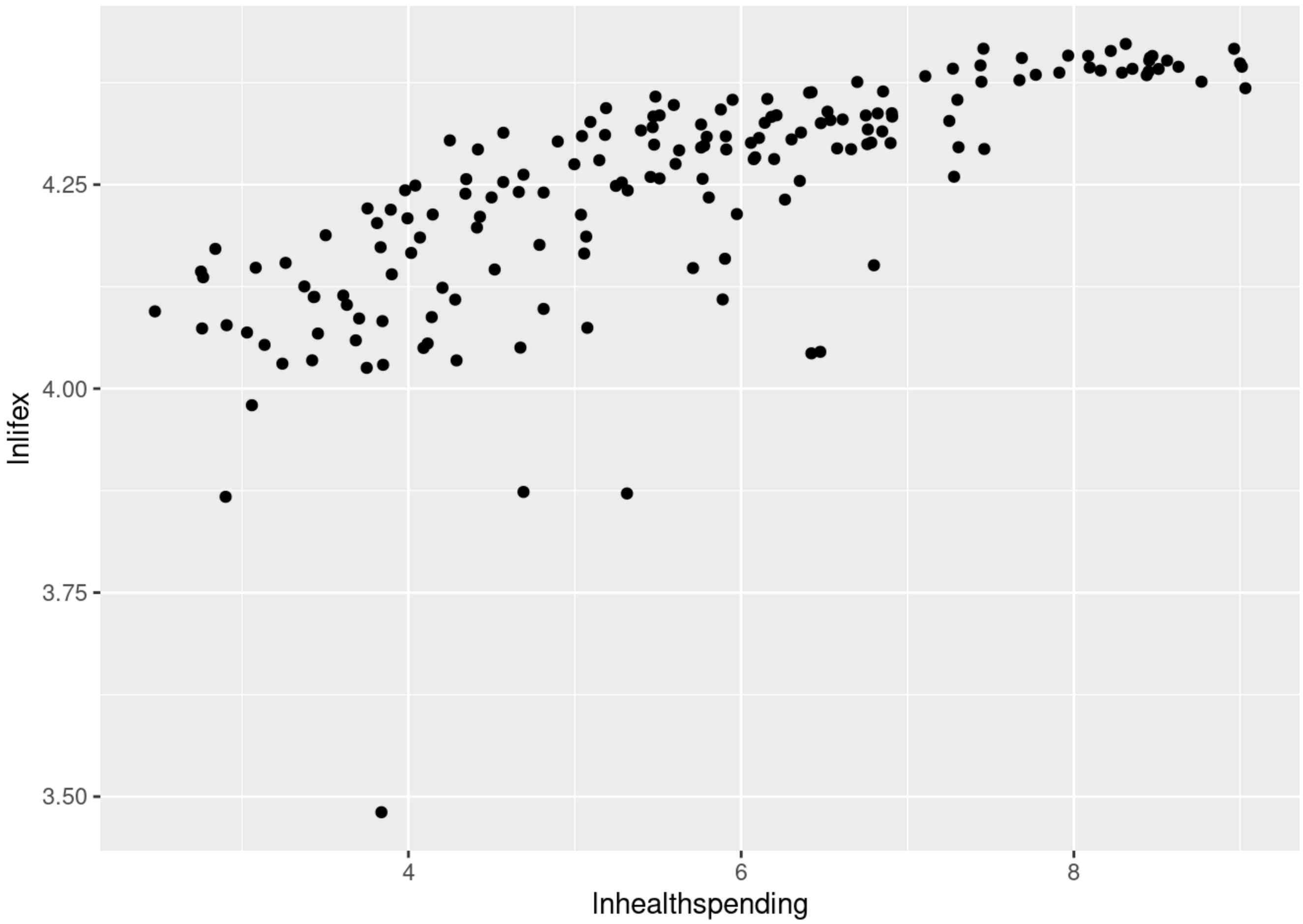
Log scatterplot

After “In”, the relationship shows a positive linear sloping trend. This means income and health spending are have a logarithmic relationship to life expectancy

```
ggplot(data = logcountriesmoney, mapping = aes(x=lnincome, y=lnlifex))+geom_jitter()
```



```
ggplot(data=logcountriesmoney, mapping = aes(x=lnhealthspending, y=lnlifex))+geom_jitter()
```



Regression for income and health spending (continuous variables)

Seperately, all regressors are significant. The log models have a higher r^2 so, I will create a multiple regression model with the two log models.

```
model1a<- lm(data = Countries, formula = lifexpectancy~income)
get_regression_table(model1a)
```

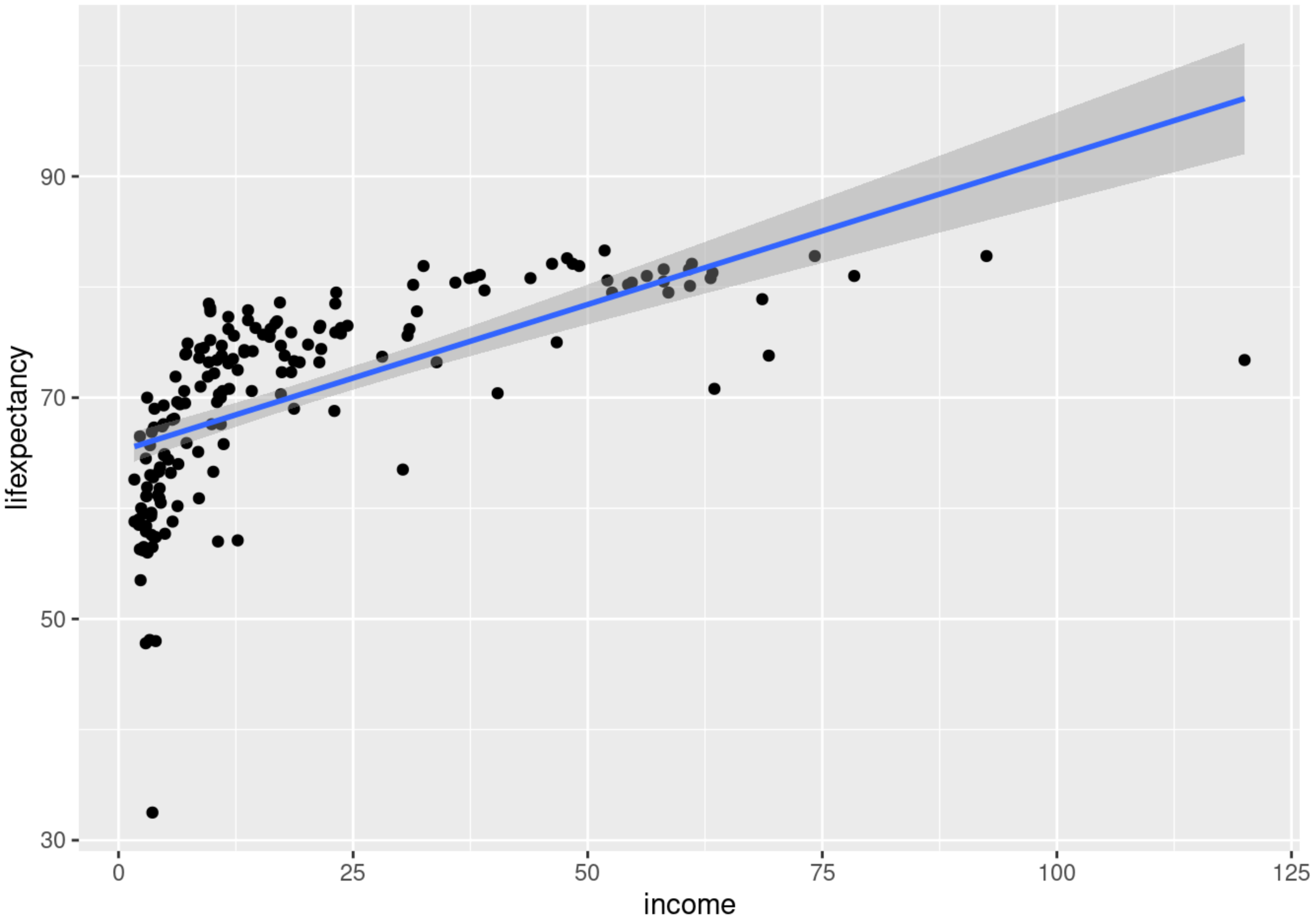
```
## # A tibble: 2 × 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   65.1       0.714      91.2       0   63.7     66.5
## 2 income       0.266     0.025      10.7       0    0.217    0.315
```

```
get_regression_summaries(model1a)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.404         0.401  45.7  6.76  6.80      115.       0     1   171
```

```
ggplot(data = Countries, mapping = aes(x=income, y=lifexpectancy))+geom_point()+geom_smooth(meth
od = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
model1b<- lm(data = Countries, formula = lifexpectancy~healthspending)
get_regression_table(model1b)
```

```
## # A tibble: 2 × 7
##   term           estimate std_error statistic p_value lower_ci upper_ci
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept        67.6       0.647    104.         0   66.3     68.8
## 2 healthspending    0.003     0           8.70      0    0.002    0.003
```

```
get_regression_summaries(model1b)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.309         0.305  53.0  7.28  7.32      75.7       0     1   171
```

```
model1c<- lm(data = logcountriesmoney, formula = lnlifex~lnincome)
get_regression_table(model1c)
```

```
## # A tibble: 2 × 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    4.00      0.017     228.        0    3.96     4.03 
## 2 lnincome     0.101     0.007      15.4       0    0.088    0.113
```

```
get_regression_summaries(model1c)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared     mse   rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>   <dbl>  <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.583          0.58 0.00771 0.0878 0.088      236.       0     1   171
```

```
model1d<- lm(data = logcountriesmoney, formula = lnlifex~lnhealthspending)
get_regression_table(model1d)
```

```
## # A tibble: 2 × 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept            3.91     0.025     154.        0    3.86     3.96 
## 2 lnhealthspending     0.06     0.004      13.8       0    0.051    0.068
```

```
get_regression_summaries(model1d)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared     mse   rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>   <dbl>  <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.531         0.528 0.00867 0.0931 0.094      192.       0     1   171
```

Multiple regression model

```
model1e<- lm(data = logcountriesmoney, formula = lnlifex~lnincome+lnhealthspending)
get_regression_table(model1e)
```

```
## # A tibble: 3 × 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           3.98      0.028   140.      0        3.92     4.04 
## 2 lnincome            0.086     0.019     4.64    0        0.049    0.123
## 3 lnhealthspending    0.01      0.012     0.833   0.406   -0.013    0.032
```

```
model1f<- lm(data = logcountriesmoney, formula = lnlifex~lnincome*lnhealthspending)
get_regression_table(model1f)
```

```
## # A tibble: 4 × 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                  3.84      0.061     63.1    0         3.72     3.96 
## 2 lnincome                   0.145     0.029      5.01   0         0.088    0.202
## 3 lnhealthspending           0.038     0.016      2.43   0.016     0.007    0.069
## 4 lnincome:lnhealthspend…   -0.011     0.004     -2.62   0.01     -0.019   -0.003
```

```
get_regression_summaries(model1f)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared     mse   rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>   <dbl>  <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.601         0.594 0.00738 0.0859 0.087      83.8       0     3   171
```

Residuals

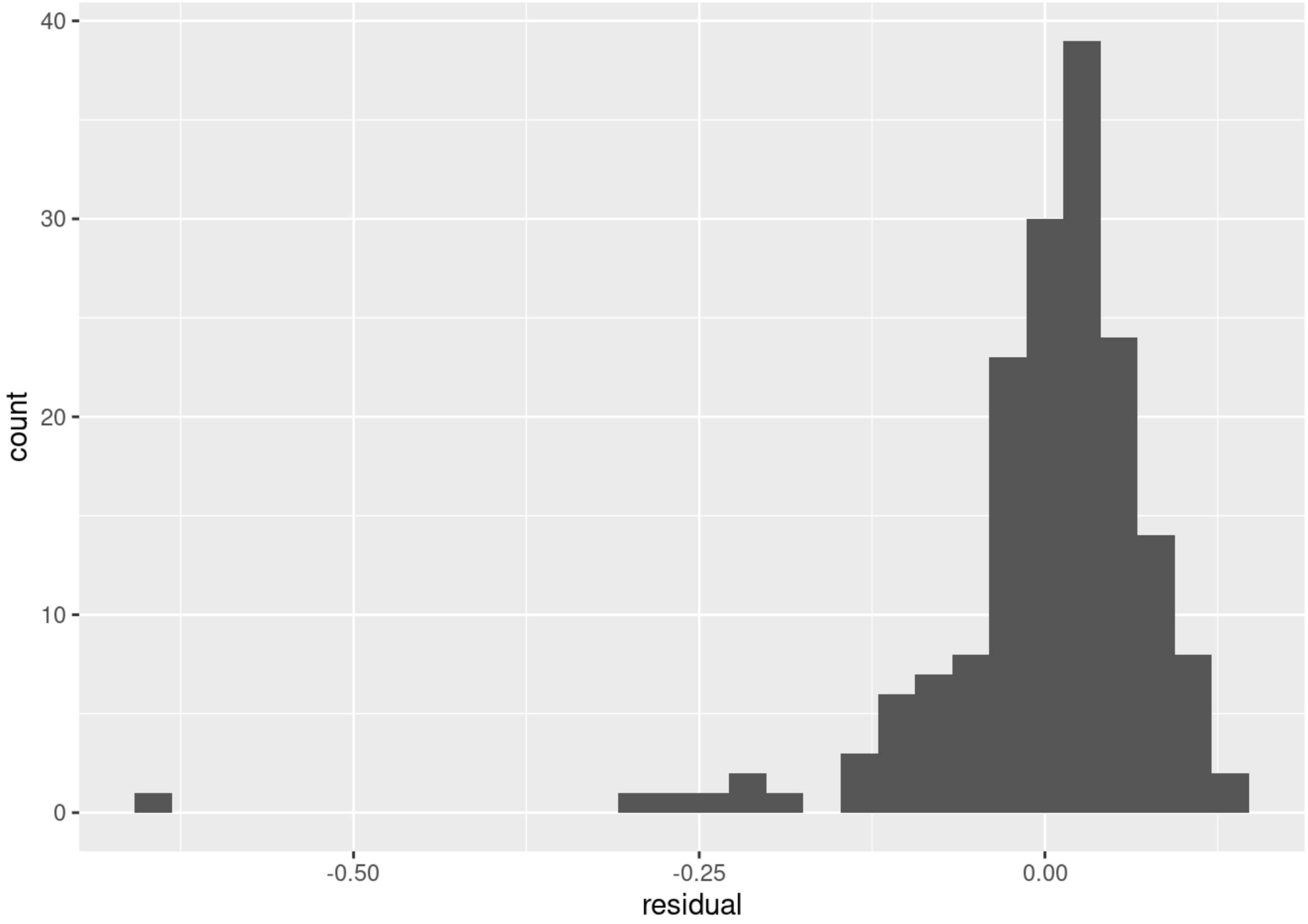
Residuals don’t follow bell shaped curve and residual vs expected life expectancy contains an outlier with some cone shape tendencies.

```
money_regPoints<- get_regression_points(model1f)
glimpse(money_regPoints)
```

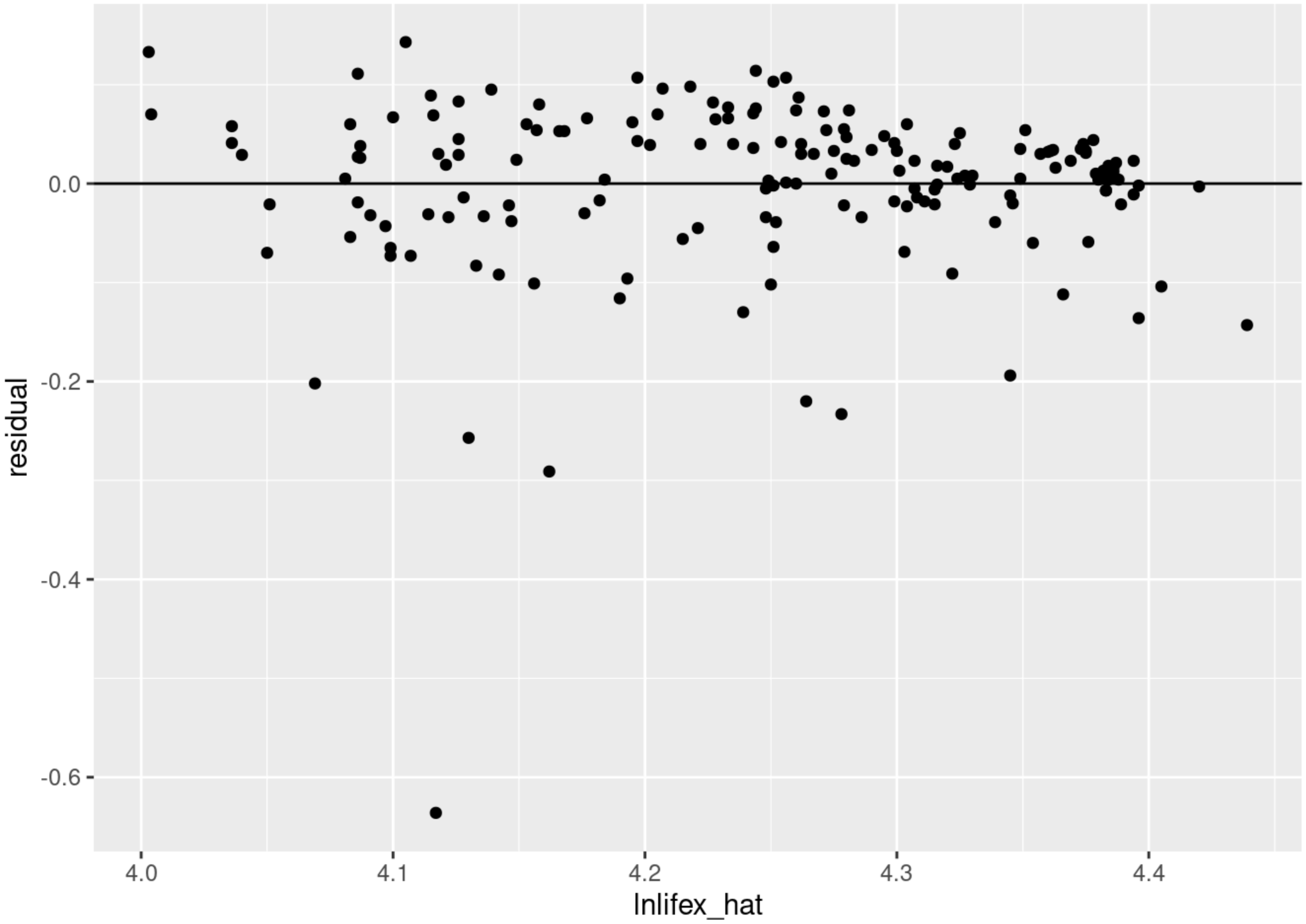
```
## Rows: 171
## Columns: 6
## $ ID               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16…
## $ lnlifex          <dbl> 4.103, 4.358, 4.311, 4.098, 4.329, 4.329, 4.303, 4.40…
## $ lnincome         <dbl> 1.504, 2.279, 2.206, 1.839, 2.912, 3.140, 1.963, 4.11…
## $ lnhealthspending <dbl> 3.630, 5.485, 5.182, 4.812, 6.537, 6.609, 4.898, 8.47…
## $ lnlifex_hat      <dbl> 4.136, 4.244, 4.233, 4.193, 4.307, 4.324, 4.207, 4.38…
## $ residual         <dbl> -0.033, 0.114, 0.077, -0.096, 0.023, 0.005, 0.096, 0.…
```

```
ggplot(data = money_regPoints, mapping = aes(x=residual))+ geom_histogram()
```

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



```
ggplot(data=money_regPoints, mapping = aes(x=lnlifex_hat, y=residual))+geom_point()+geom_hline(y
intercept = 0)
```



We can see that the R^2 value is fairly high with the interactions between income, healthspent, babies per woman, childmortality, and continent.

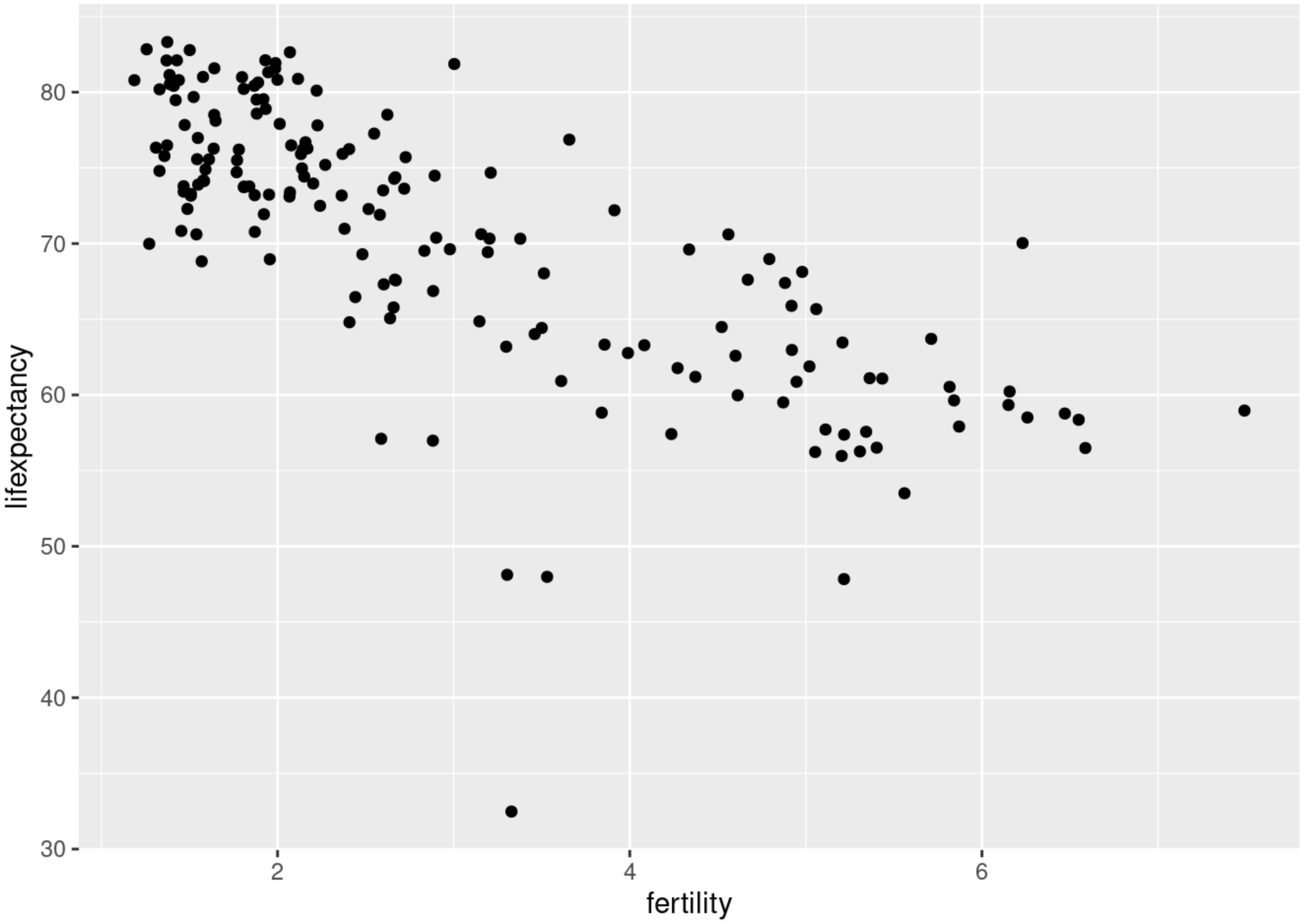
Exploratory Analysis

The correlation between life expectancy and babies per woman has a negative relationship, where as life expectancy rise, the babies per woman decreases. We can also conclude that Africa and Oceana has the lower side of life expectancy. The data is widely spread at Oceana and Africa consists of an outlier. The relationship between life expectancy and child mortality is negative. As life expectancy increases, child mortality decreases.

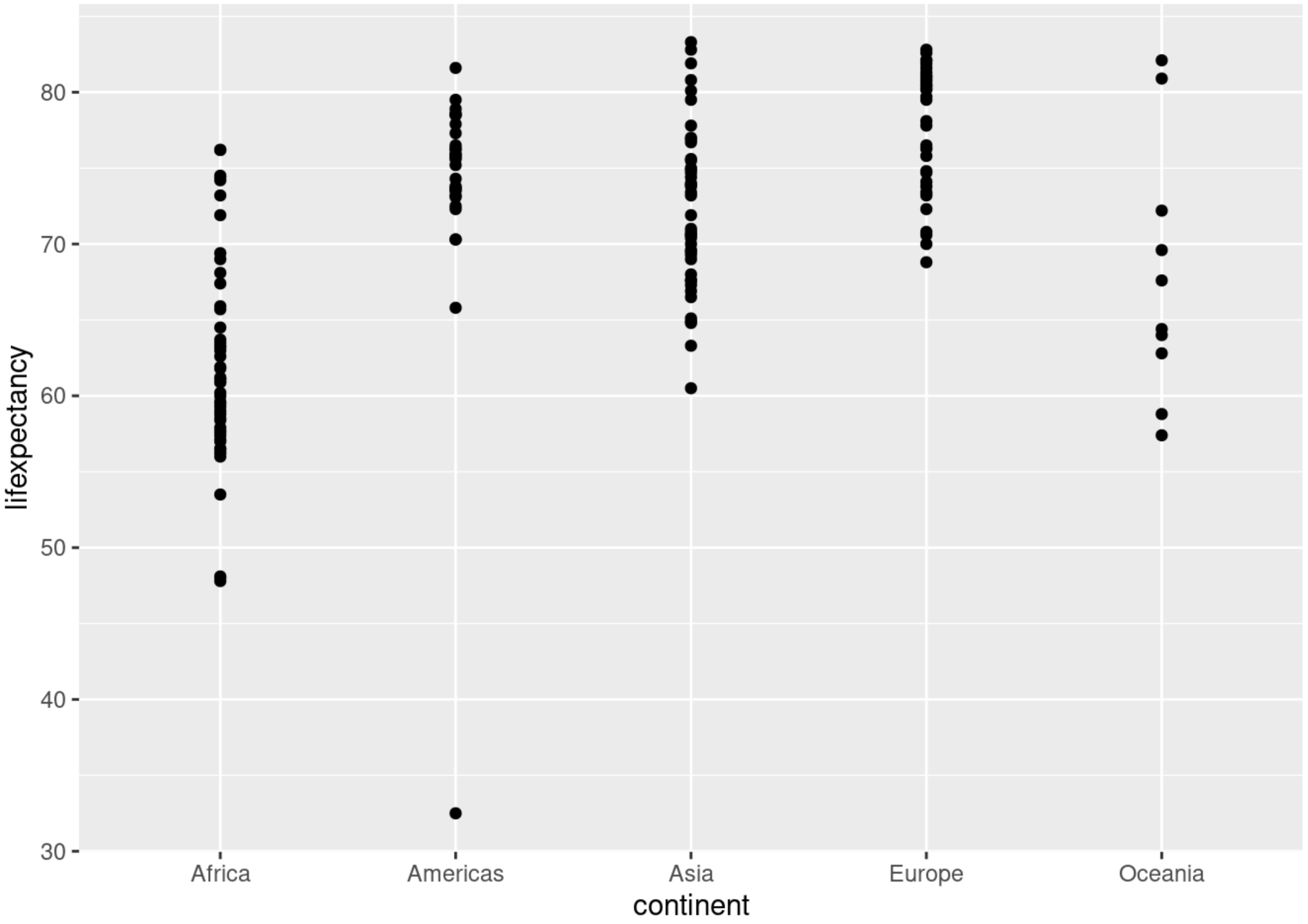
```
Countries$continent <- as.factor(Countries$continent)
model_full5 <- lm(lifexpectancy ~ income +
                 healthspending +
                 fertility + childmortality +
                 continent,
                 data = Countries)
summary(model_full5)
```

```
##
## Call:
## lm(formula = lifexpectancy ~ income + healthspending + fertility +
##     childmortality + continent, data = Countries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5514  -1.6770   0.1282   2.5100   8.9155
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       72.2554989  1.5482938  46.668  < 2e-16 ***
## income             0.0409501  0.0245254   1.670  0.09691 .
## healthspending     0.0008221  0.0002950   2.787  0.00596 **
## fertility          0.8228372  0.3962459   2.077  0.03942 *
## childmortality    -0.1796813  0.0136993 -13.116  < 2e-16 ***
## continentAmericas  2.5340041  1.0982264   2.307  0.02230 *
## continentAsia      1.5907023  0.9765662   1.629  0.10528
## continentEurope    1.5236443  1.2520957   1.217  0.22542
## continentOceania  -3.2956634  1.3644856  -2.415  0.01683 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.631 on 162 degrees of freedom
## Multiple R-squared:  0.8372, Adjusted R-squared:  0.8291
## F-statistic: 104.1 on 8 and 162 DF,  p-value: < 2.2e-16
```

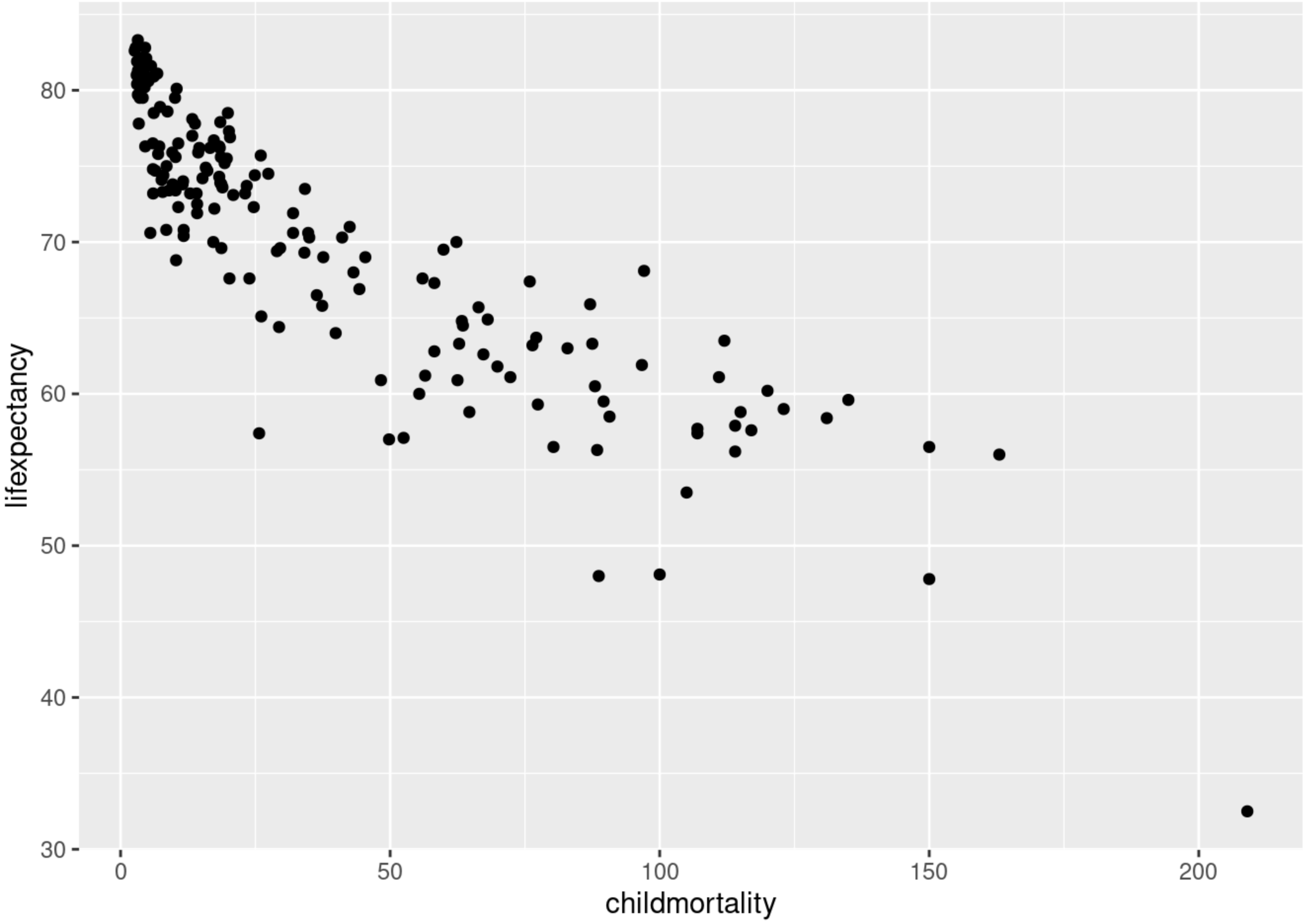
```
ggplot(Countries, mapping = aes(x=fertility, y=lifexpectancy))+geom_jitter()
```



```
ggplot(Countries, mapping = aes(x = continent, y=lifexpectancy))+geom_point()
```



```
ggplot(Countries, mapping = aes(x = childmortality, y=lifeexpectancy))+geom_point()
```



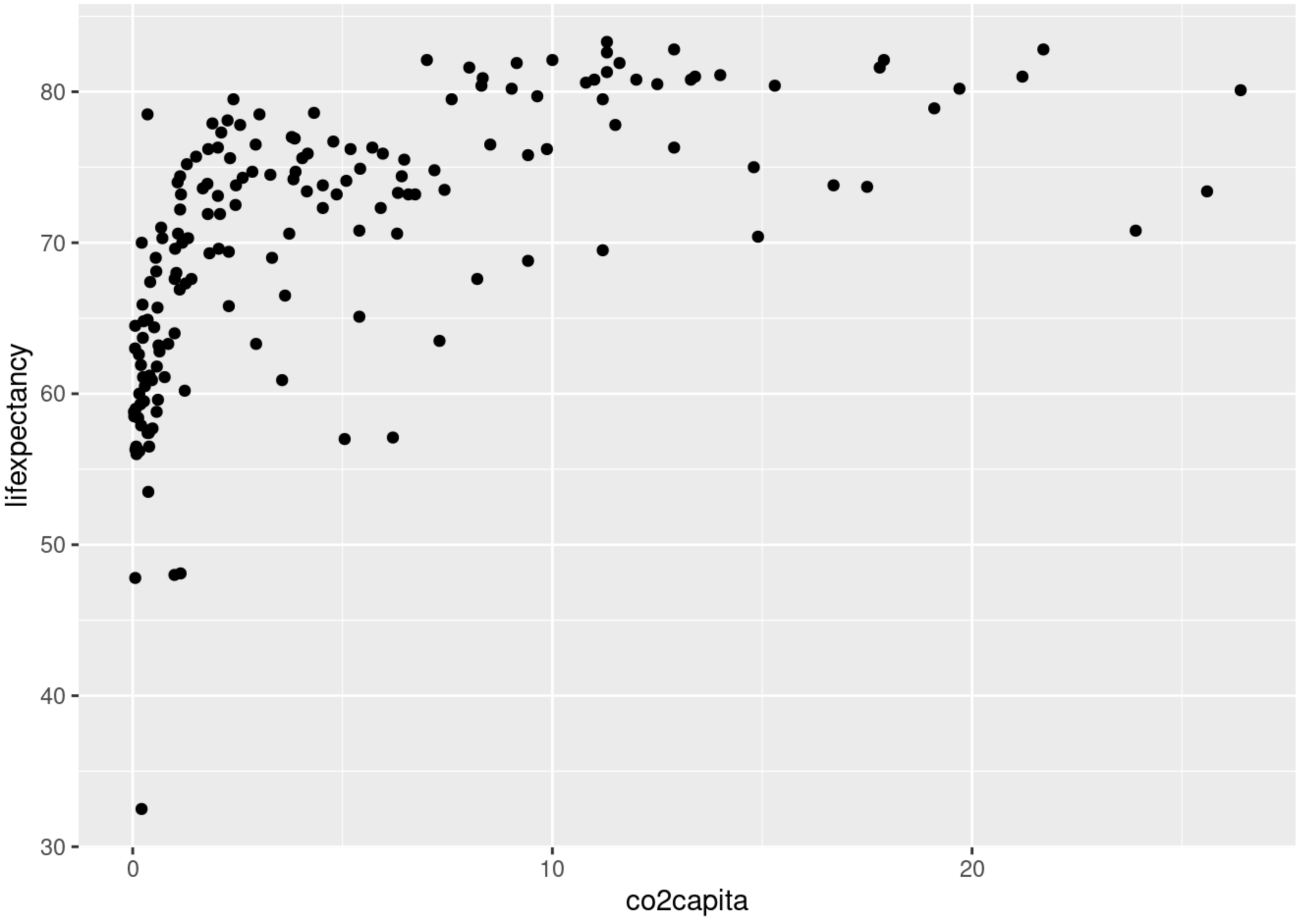
##

Lets also check the relationship with CO2 emissions and basic water with life expectancy. ## We can visually conclude that the relationship between co2 emission and life expectancy is negative, however there exists a lot of outliers. CO2 emissions will not be used. The correlation between basic water and life expectancy is positive, where as life expectancy increases, so does the percentage of basic water.

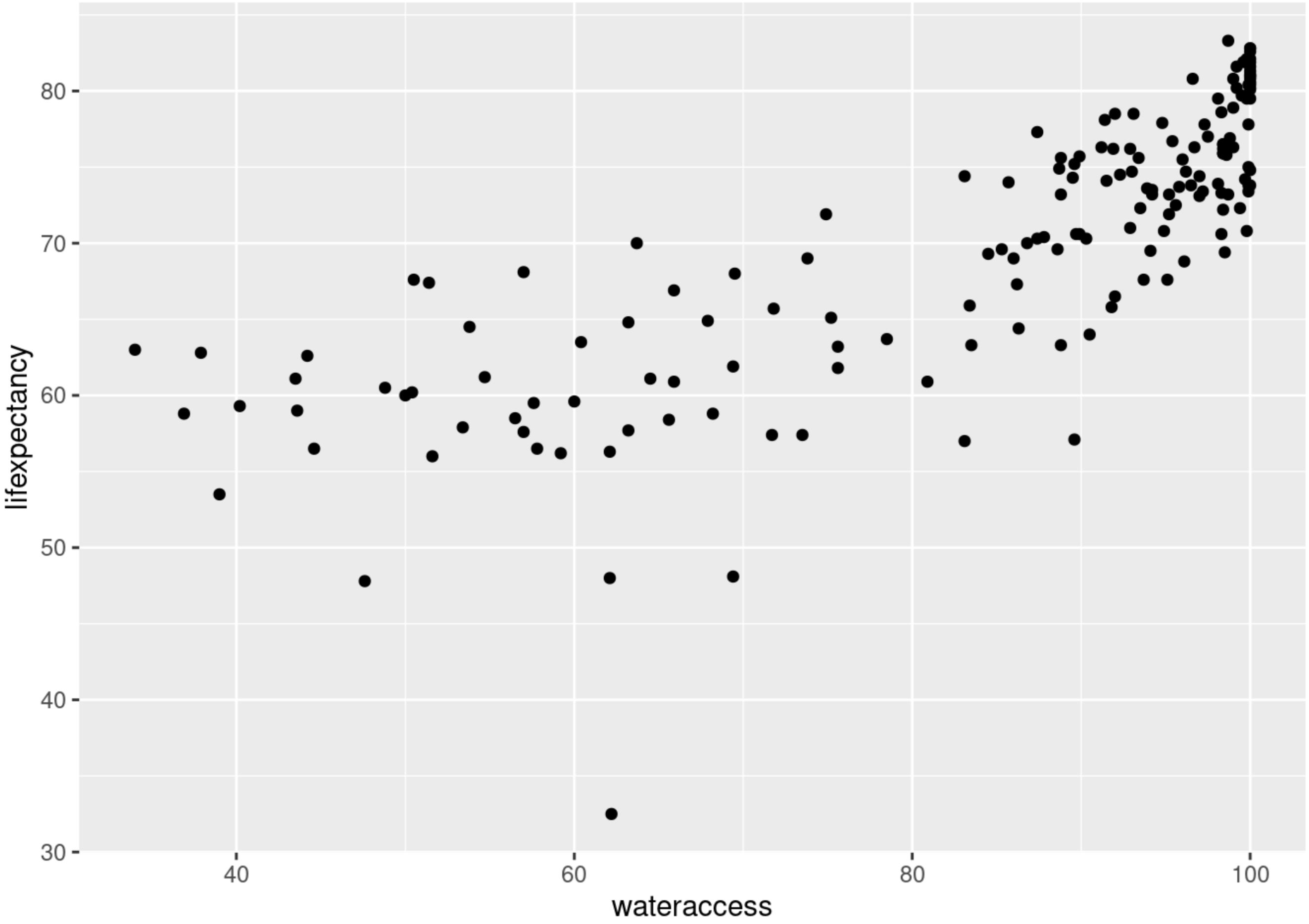
```
Countries$continent <- as.factor(Countries$continent)
model_full6 <- lm(lifexpectancy ~ healthspending +
                 co2capita+
                 wateraccess,
                 data = Countries)
summary(model_full6)
```

```
##
## Call:
## lm(formula = lifexpectancy ~ healthspending + co2capita + wateraccess,
##     data = Countries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.4309  -2.2669   0.2364   3.1367  10.0703
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.211e+01  1.994e+00  21.123  < 2e-16 ***
## healthspending 1.236e-03  2.997e-04   4.123 5.87e-05 ***
## co2capita      6.286e-02  9.832e-02   0.639    0.523
## wateraccess    3.175e-01  2.572e-02  12.343  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5 on 167 degrees of freedom
## Multiple R-squared:  0.6818, Adjusted R-squared:  0.6761
## F-statistic: 119.3 on 3 and 167 DF,  p-value: < 2.2e-16
```

```
ggplot(Countries, mapping = aes(x = co2capita, y=lifexpectancy))+geom_point()
```



```
ggplot(Countries, mapping = aes(x = wateraccess, y=lifeexpectancy))+geom_point()
```



Second-order terms and Residual Analysis

Residual Analysis

The Residual v fitted plot shows that the residuals are randomly dispered, suggesting normality.

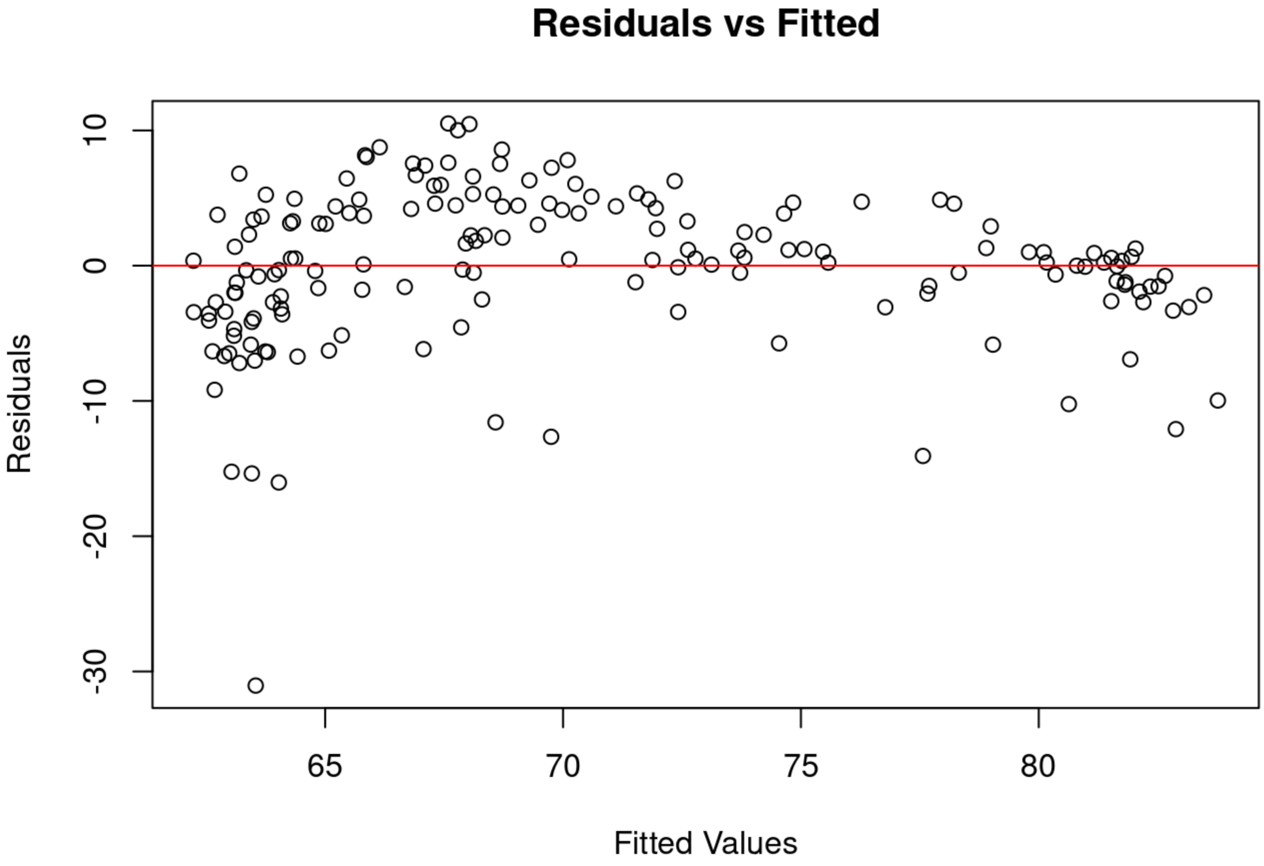
The qq pot shows slight deviation from the line suggesting possible outliers or non linearity

The scale plot shows a relatively uniform spread and suggests that homoscedasticity might be reasonable

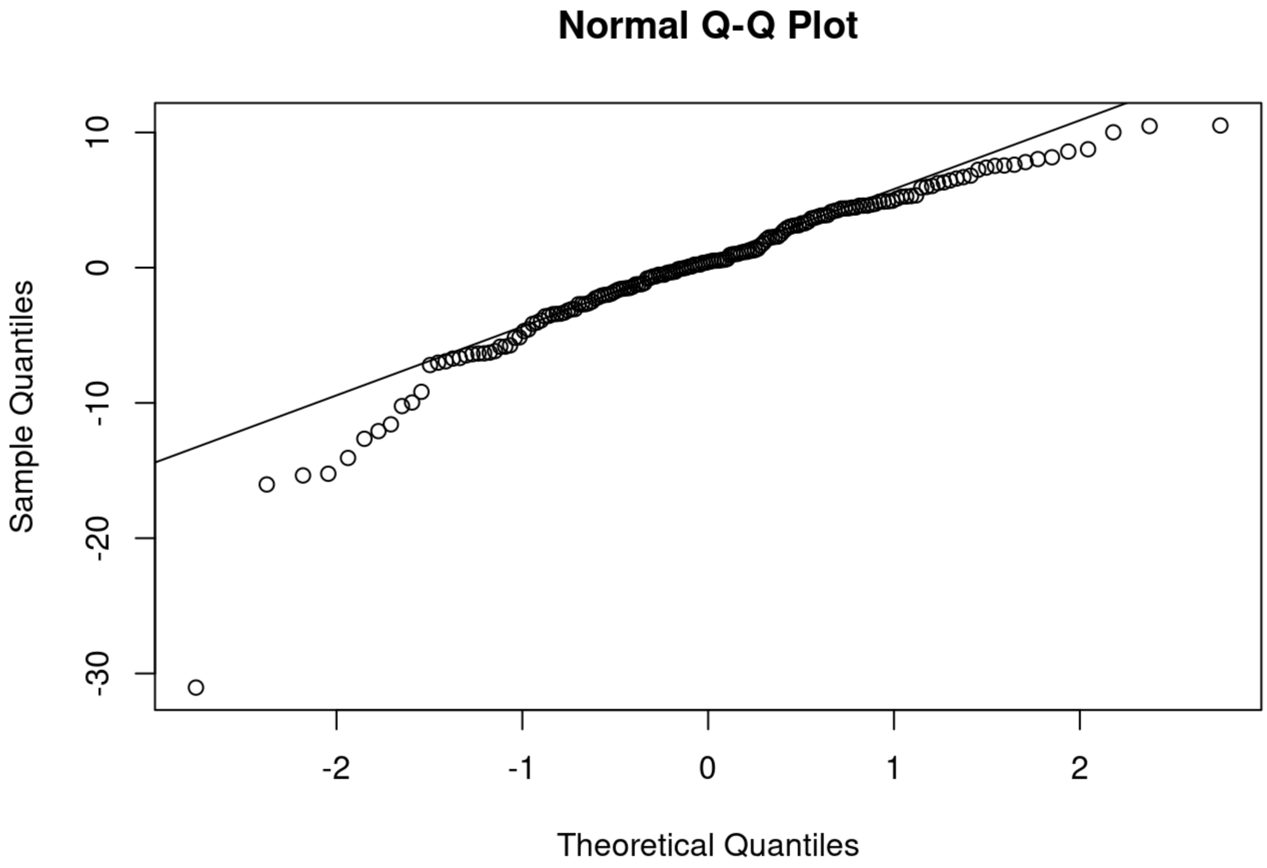
Model Fit

Around 57.7% of the variability in life expectancy is explained by the model. The low p value in the F statistic test suggests that the model is statistically significant, and has a strong influence on life expectancy

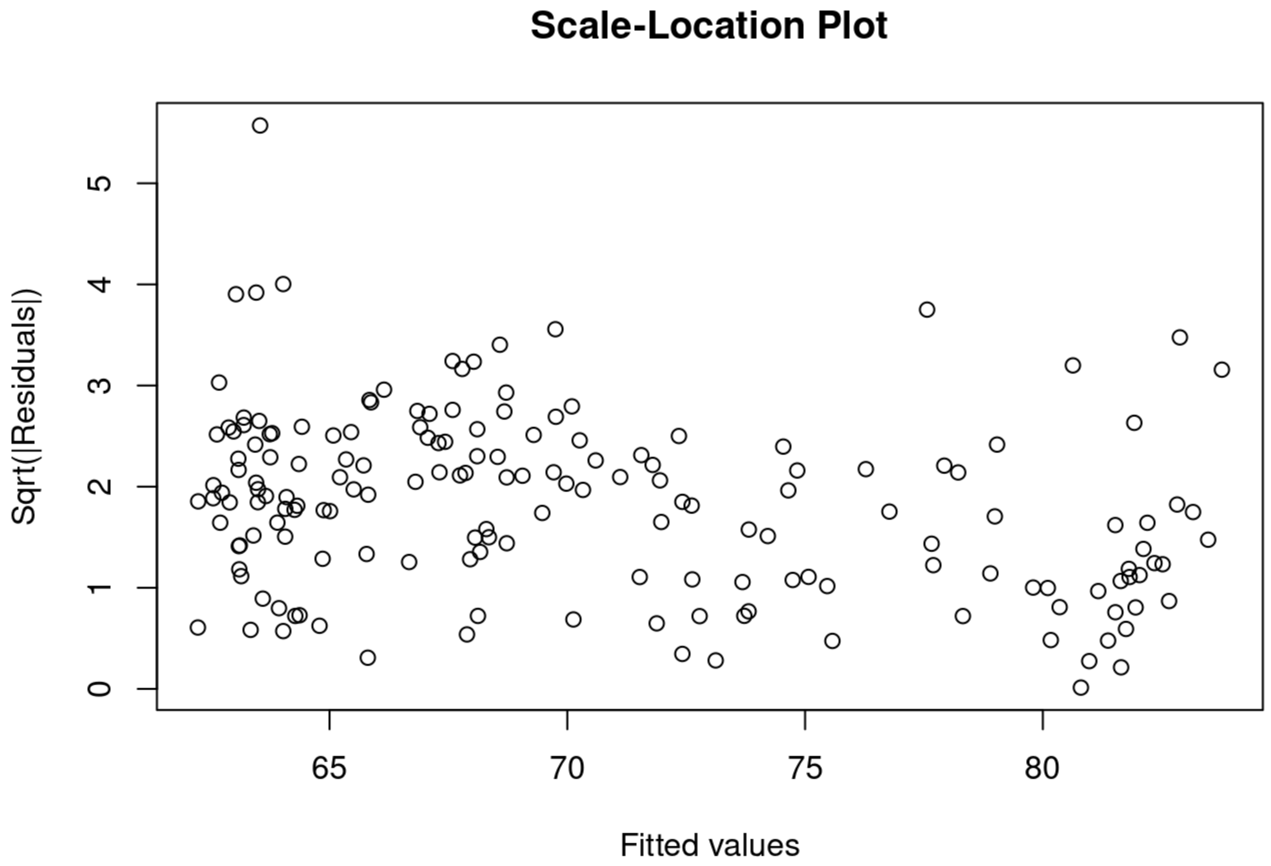
```
model_advanced1 <- lm(lifexpectancy ~ income * healthspending + I(income^2) + I(healthspending^
2), data = Countries)

plot(model_advanced1$residuals ~ model_advanced1$fitted.values,
     main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")
```



```
qqnorm(model_advanced1$residuals)
qqline(model_advanced1$residuals)
```



```
plot(model_advanced1$fitted.values, sqrt(abs(model_advanced1$residuals)),
     main = "Scale-Location Plot",
     xlab = "Fitted values", ylab = "Sqrt(|Residuals|)")
```



```
summary(model_advanced1)
```

```
## 
## Call:
## lm(formula = lifexpectancy ~ income * healthspending + I(income^2) + 
##     I(healthspending^2), data = Countries)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -31.0401  -2.6994   0.4198   4.1568  10.5121 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            6.107e+01  8.599e-01  71.014  < 2e-16 ***
## income                 6.813e-01  1.010e-01   6.746 2.44e-10 ***
## healthspending         1.851e-03  1.874e-03   0.988    0.325    
## I(income^2)           -4.700e-03  1.118e-03  -4.204 4.29e-05 ***
## I(healthspending^2)    2.221e-07  2.907e-07   0.764    0.446    
## income:healthspending -6.130e-05  4.927e-05  -1.244    0.215    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.8 on 165 degrees of freedom
## Multiple R-squared:  0.577,  Adjusted R-squared:  0.5641 
## F-statistic: 45.01 on 5 and 165 DF,  p-value: < 2.2e-16
```

A simplified version of the regression model, removing non-significant terms.

```
model_simplified2 <- update(model_advanced1, . ~ . - income:healthspending)
summary(model_simplified2)
```

```
##
## Call:
## lm(formula = lifexpectancy ~ income + healthspending + I(income^2) +
##     I(healthspending^2), data = Countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.174  -2.723   0.493   3.982  10.931
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.121e+01  8.537e-01  71.701  < 2e-16 ***
## income               6.992e-01  1.001e-01   6.984 6.59e-11 ***
## healthspending       2.939e-04  1.396e-03   0.210    0.834
## I(income^2)         -5.547e-03  8.879e-04  -6.248 3.38e-09 ***
## I(healthspending^2) -8.196e-08  1.577e-07  -0.520    0.604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.809 on 166 degrees of freedom
## Multiple R-squared:  0.573,  Adjusted R-squared:  0.5627
## F-statistic: 55.69 on 4 and 166 DF,  p-value: < 2.2e-16
```

High adjr^2, no interaction, will test with an interaction.

```
Countries$continent <- as.factor(Countries$continent)
model_full1 <- lm(lifexpectancy ~ income +
                 healthspending +
                 fertility + childmortality +
                 continent,
                 data = Countries)
get_regression_summaries(model_full1)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.837         0.829  12.5  3.53  3.63      104.       0     8   171
```

```
Countries$continent <- as.factor(Countries$continent)
model_full2 <- lm(lifexpectancy ~ healthspending +
                 co2capita+
                 wateraccess,
                 data = Countries)
get_regression_summaries(model_full2)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.682         0.676  24.4  4.94     5      119.       0     3   171
```

```
Countries$continent <- as.factor(Countries$continent)
model_full3 <- lm(lifexpectancy ~ continent+
                   gdpcapita+
                   population,
                 data = Countries)
get_regression_summaries(model_full3)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared      mse     rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>    <dbl>    <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1         1           NaN 2.90e-27 5.39e-14   NaN       NaN     NaN   170   171
```

This gives r^2 value of 0.173

```
Countries$continent <- as.factor(Countries$continent)
model_full4 <- lm(lifexpectancy ~ population,
                 data = Countries)
get_regression_summaries(model_full4)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.956         0.173  3.36  1.83  7.99      1.22   0.401   161   171
```

Final Regression Model

continent seems insignificant unless its oceania. income and health spending insignificant probably collinearity somewhere.

Update, after releveling (indicating whether oceania or not) and removing health spending regressor (explainable by income), adjR2 is 0.826 with all significant regressors.

```
test<- lm(data = logcountriesmoney, formula = lnlifex~lnincome*lnhealthspending+fertility + chil
dmortality + continent)
get_regression_table(test)
```

```
## # A tibble: 10 × 7
##    term                    estimate std_error statistic p_value lower_ci upper_ci
##    <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
##  1 intercept                  4.27      0.074    57.8     0        4.12     4.42
##  2 lnincome                   0.014     0.024     0.578   0.564   -0.033    0.061
##  3 lnhealthspending          -0.017     0.013    -1.32    0.19    -0.042    0.008
##  4 fertility                  0.026     0.007     3.80    0        0.012    0.039
##  5 childmortality            -0.003     0       -13.7     0       -0.004   -0.003
##  6 continent: Americas        0.025     0.017     1.44    0.152   -0.009    0.059
##  7 continent: Asia            0.016     0.016     0.996   0.321   -0.015    0.047
##  8 continent: Europe          0.018     0.019     0.924   0.357   -0.02     0.056
##  9 continent: Oceania        -0.055     0.021    -2.54    0.012   -0.097   -0.012
## 10 lnincome:lnhealthspen…     0.005     0.003     1.41    0.161   -0.002    0.011
```

```
get_regression_summaries(test)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared     mse   rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>   <dbl>  <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.834         0.825 0.00306 0.0553 0.057      90.2       0     9   171
```

```
newdata<- logcountriesmoney%>% mutate(oceaniaIndicator=ifelse(continent=="Oceania", 1, 0))
glimpse(newdata)
```

```
## Rows: 171
## Columns: 18
## $ X                 <int> 1, 2, 3, 6, 8, 9, 10, 12, 13, 14, 15, 16, 18, 19, 20…
## $ country           <chr> "Afghanistan", "Albania", "Algeria", "Angola", "Anti…
## $ childmortality    <dbl> 88.00, 13.30, 27.40, 120.00, 9.59, 14.40, 18.50, 4.7…
## $ co2capita         <dbl> 0.290, 2.260, 3.280, 1.240, 5.960, 4.170, 1.780, 17.…
## $ fertility         <dbl> 5.82, 1.65, 2.89, 6.16, 2.13, 2.37, 1.55, 1.93, 1.44…
## $ gdpcapita         <chr> "569", "3580", "3930", "2990", "14.4k", "13.6k", "28…
## $ healthspending    <dbl> 37.7, 241.0, 178.0, 123.0, 690.0, 742.0, 134.0, 4780…
## $ income            <dbl> 4.50, 9.77, 9.08, 6.29, 18.40, 23.10, 7.12, 61.10, 6…
## $ lifexpectancy     <dbl> 60.5, 78.1, 74.5, 60.2, 75.9, 75.9, 73.9, 82.1, 80.8…
## $ murder            <chr> "4130", "65.9", "530", "824", "5.05", "2450", "154",…
## $ population        <chr> "28.2M", "2.91M", "35.9M", "23.4M", "85.7k", "41.1M"…
## $ populationdensity <dbl> 43.40, 106.00, 15.10, 18.70, 195.00, 14.70, 104.00, …
## $ wateraccess       <dbl> 48.8, 91.4, 92.3, 50.4, 98.4, 98.4, 98.1, 99.9, 100.…
## $ continent         <chr> "Asia", "Europe", "Africa", "Africa", "Americas", "A…
## $ lnlifex           <dbl> 4.102643, 4.357990, 4.310799, 4.097672, 4.329417, 4.…
## $ lnincome          <dbl> 1.5040774, 2.2793165, 2.2060742, 1.8389611, 2.912350…
## $ lnhealthspending  <dbl> 3.629660, 5.484797, 5.181784, 4.812184, 6.536692, 6.…
## $ oceaniaIndicator  <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```

```
model5<- lm(data=newdata, formula = lnlifex~lnincome+fertility+
              childmortality+oceaniaIndicator)
get_regression_table(model5)
```

```
## # A tibble: 5 × 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           4.20      0.027    153.         0    4.14     4.25
## 2 lnincome            0.039     0.007      5.71       0    0.025    0.052
## 3 fertility           0.026     0.006      4.41       0    0.014    0.037
## 4 childmortality     -0.003     0        -14.1        0   -0.004   -0.003
## 5 oceaniaIndicator   -0.07      0.019     -3.66       0   -0.108   -0.032
```

```
get_regression_summaries(model5)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared     mse   rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>   <dbl>  <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1      0.83         0.826 0.00314 0.0560 0.057      203.       0     4   171
```

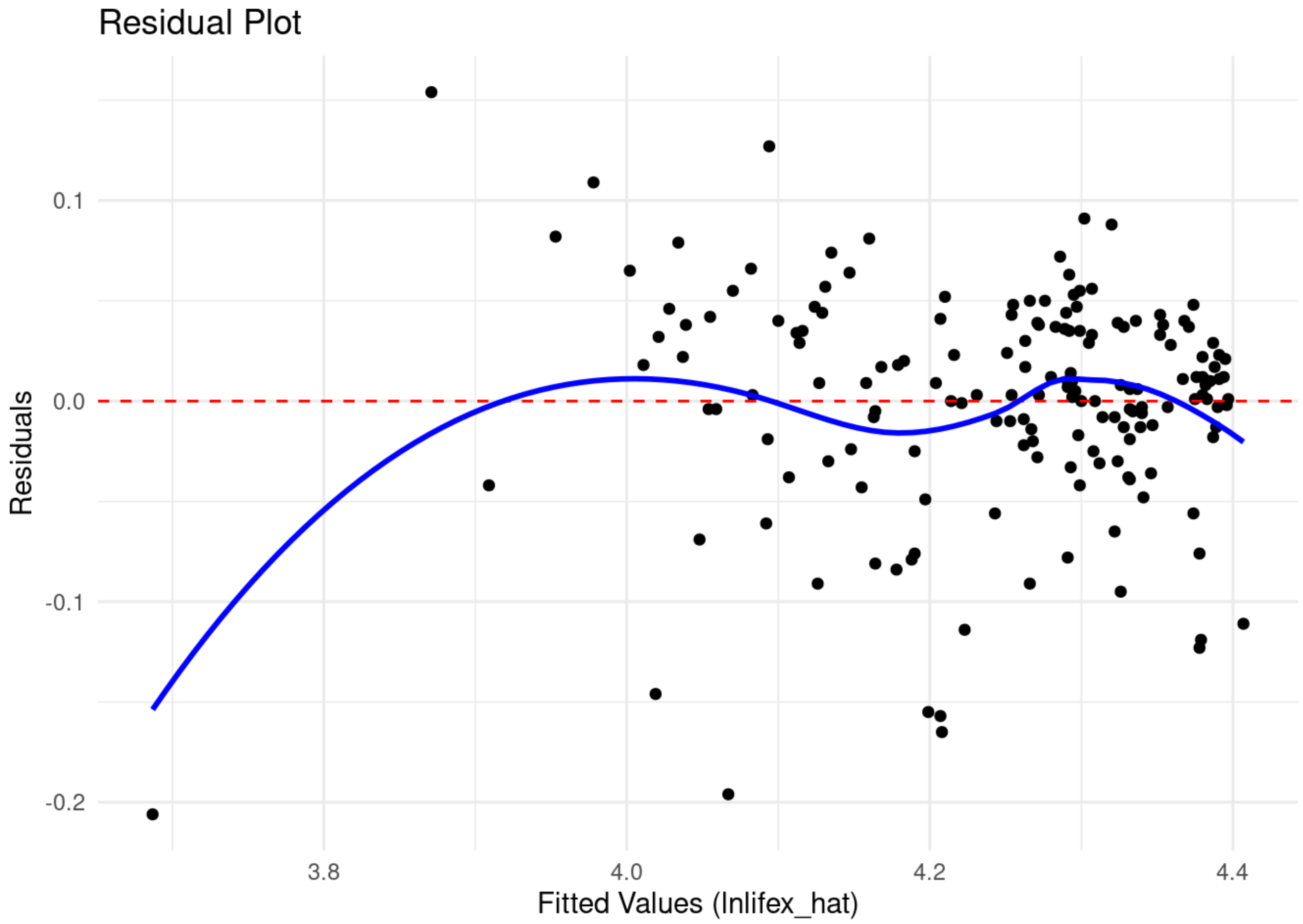
```
finalpoints<- get_regression_points(model5)
```

After multiple regressors, we found out that the variables surrounding child mortality, fertility and income and significance towards life expectancy. Oceania seemed to have a significant impact on life expectancy so, we created an indicator for whether the country was in Oceania or not. Variables, such as population served little to no significance in our models. In terms of initial exploration of regressors, I tried creating a smaller multiple regression model with ln life expectancy, ln income, and ln health spending. The interaction term was significant but when merged with other regressors such as child mortality, fertility, etc. the interaction was insignificant which implies colliniearity so, we took out health spending and left just income.

The residual plot we got below is the best fit for the regression model we created. The model below shows some explanatory power. The points are some what random, which suggests that the predictors might not be capturing the non lineartires present in the interaction between predictors and the response variables we chose.

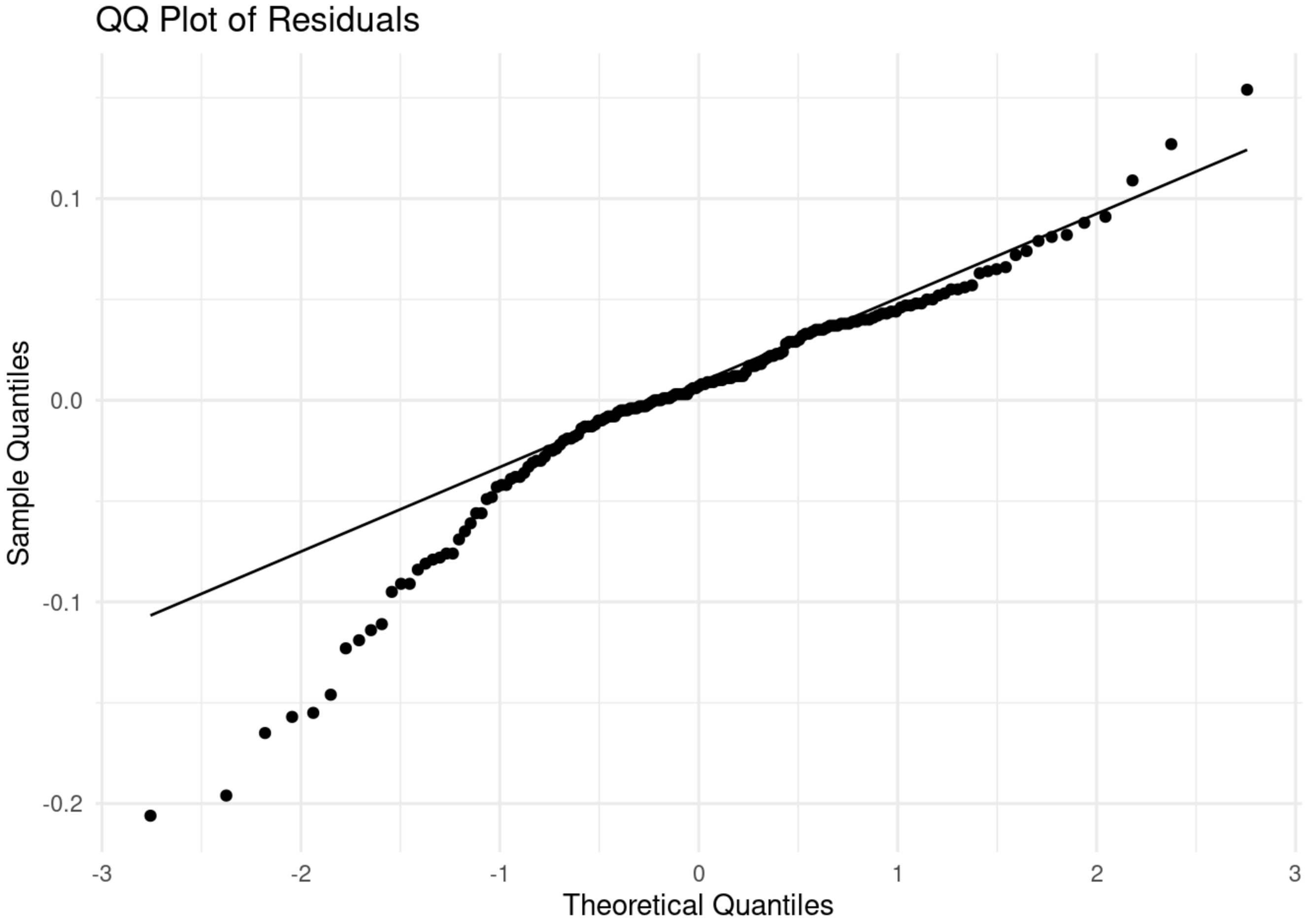
```
ggplot(data = finalpoints, mapping = aes(x = lnlifex_hat, y = residual)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  geom_smooth(se = FALSE, color = "blue") +
  labs(title = "Residual Plot",
       x = "Fitted Values (lnlifex_hat)",
       y = "Residuals") +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The points follow the line closely and suggests that the residuals are approximately normally distributed around the mean. However, the clear deviations on both ends of the tail suggest that there are more anomalies. The qq plot below shows that the points stray away from the mean which hints at the potential existence of outliers or influential points.

```
qq_plot <- ggplot(finalpoints, aes(sample = residual)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "QQ Plot of Residuals",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal()

print(qq_plot)
```


QQ Plot of Residuals
Sample Quantiles
0.1
0.0
-0.1
-0.2
Theoretical Quantiles
-3
-2
-1
0
1
2
3