STAT5120, Assignment 13, Allen Baumgarten (NOT an FSU fan)

Brief Description of the Data:

The *infmort* dataset in the Faraway package contains four variables documenting infant mortalities (presumably counts of mortalities?) in five different major regions of the world. Per-capita income and oil exportation are two of these variables that may be related to infant mortality, or at least associated with it.

Our task will be to see if, at minimum, there is a statistically significant association between *region*, *oil* (export), and *income* with the counts of infant *mortalities* and if so, can the changes in these three variables be said to in some way 'explain' or 'be associated with' the increase(decrease) in infant mortalities. We must be cautious in assuming that any of these three variables explain what causes infant mortalities though it is a good bet that income levels would be associated with mortalities. Regions may also be associated with infant mortality as well if certain regions can be said to have more internal conflict or economic hardship than others, though again, this data may be too high of a level to make any strong inferences.

We will begin with a summary examination of the data itself by looking at univariate and bivariate relationships, both mathematically and graphically.

Exploratory Data Analysis

<u>Statistical Summary</u>

A statistical summary of the four *regions* reveals that Africa accounts for the most observations with Asia coming in with the second-most. The per-capita income levels are skewed to the right with a few regions with very high income such as the United States, Sweden, Denmark, and Germany. Thus, our *income* variable is decidedly skewed to the right. Moreover, the Americas, Asia, and Africa regions all seem to have similar income levels while Europe appears to have more varied income levels (see boxplot following).

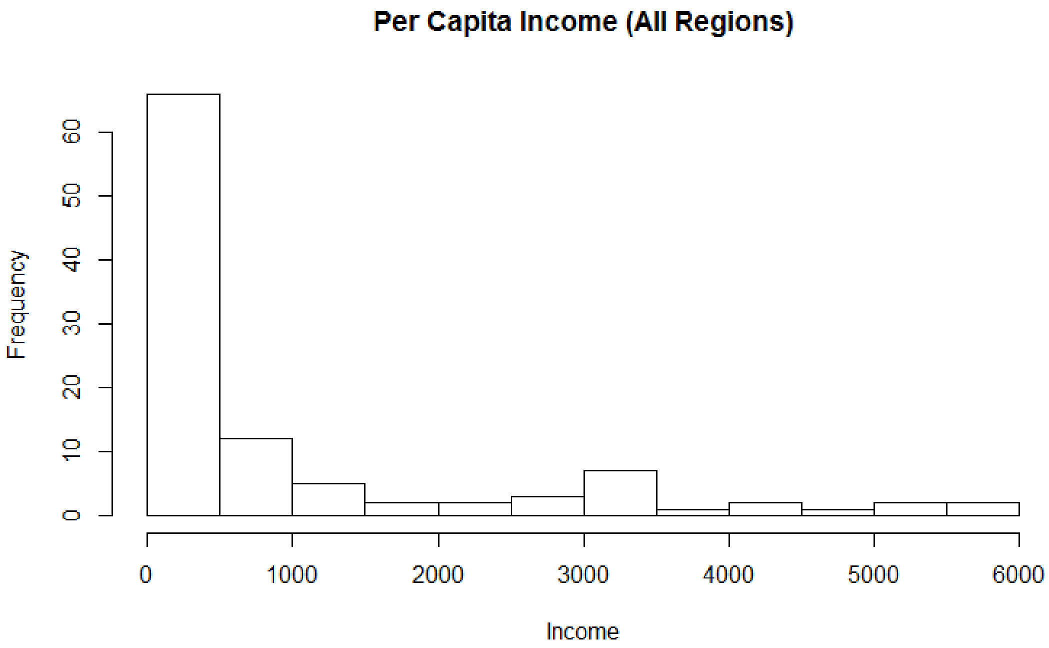
Infant *mortalities* are also skewed with a median mortality (rate?) of 60.6 and a maximum of 650 in Saudi Arabia. This distribution is also skewed right with a couple of outliers to consider.

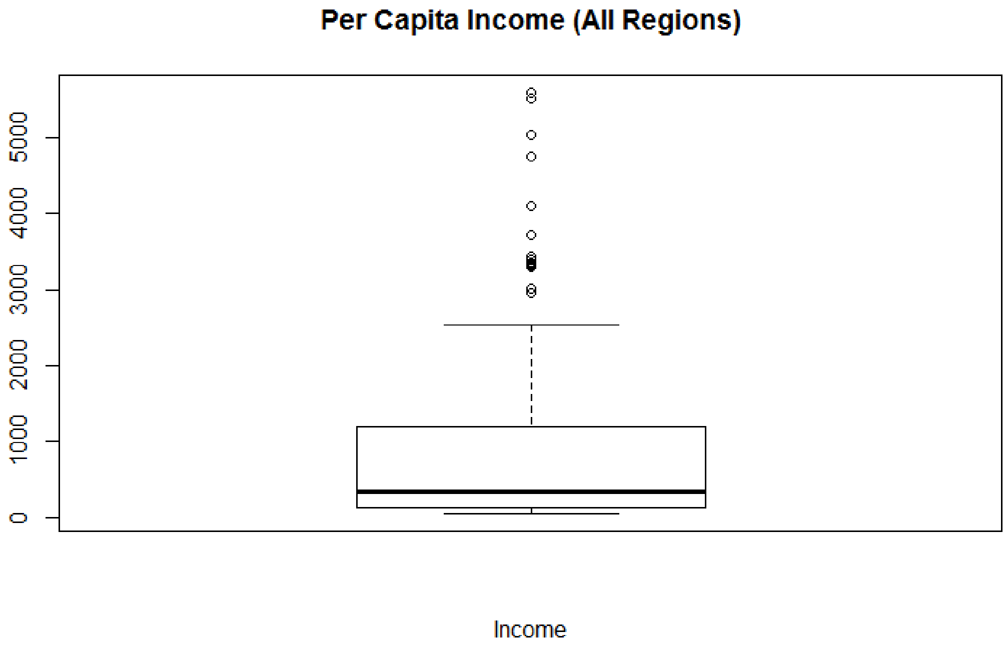
Bivariate relationships are also examined between income and mortality, respectively. A scatter plot of all mortalities vs. income shows clustering at the lower income levels and trails off to the right wit increases. At first glance, we do not appear to see a 'clean' straight (linear) relationship between these two numeric variables.

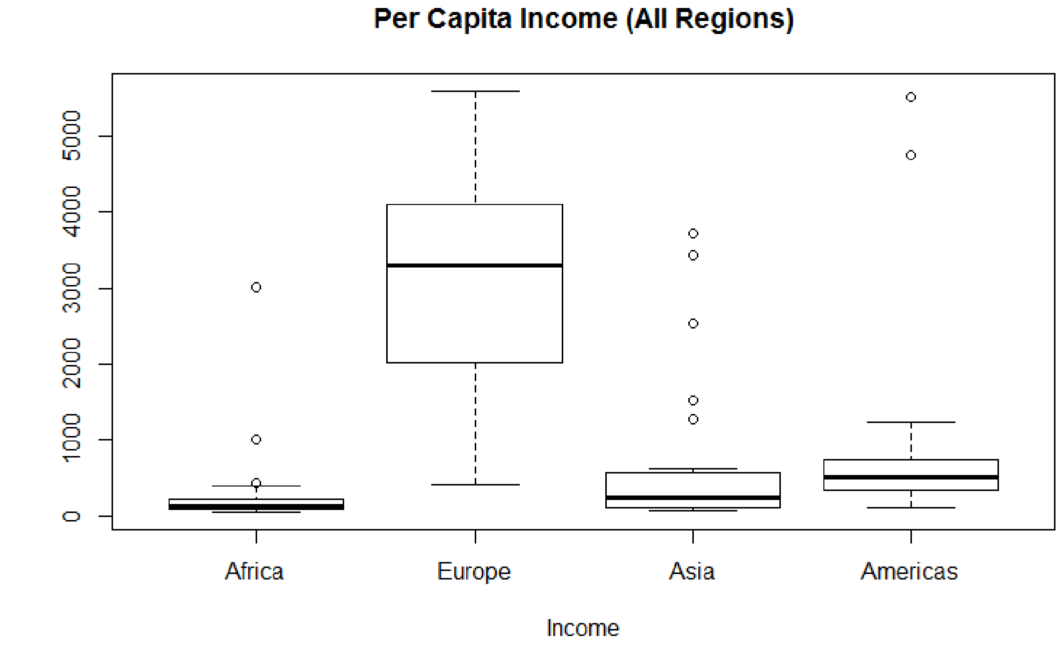
```
region          income              mortality           oil
 Africa:    34   Min.:      50.0    Min.:       9.60    oil exports:      9
 Europe:    18   1st Qu.:  123.0    1st Qu.:   26.20    no oil exports:  96
 Asia:      30   Median:   334.0    Median:    60.60
 Americas:  23   Mean:     998.1    Mean:      89.05
                 3rd Qu.: 1191.0    3rd Qu.:  129.40
                 Max:     5596.0    Max.:     650.00
                                    NA's:        4
```

Graphical Examination

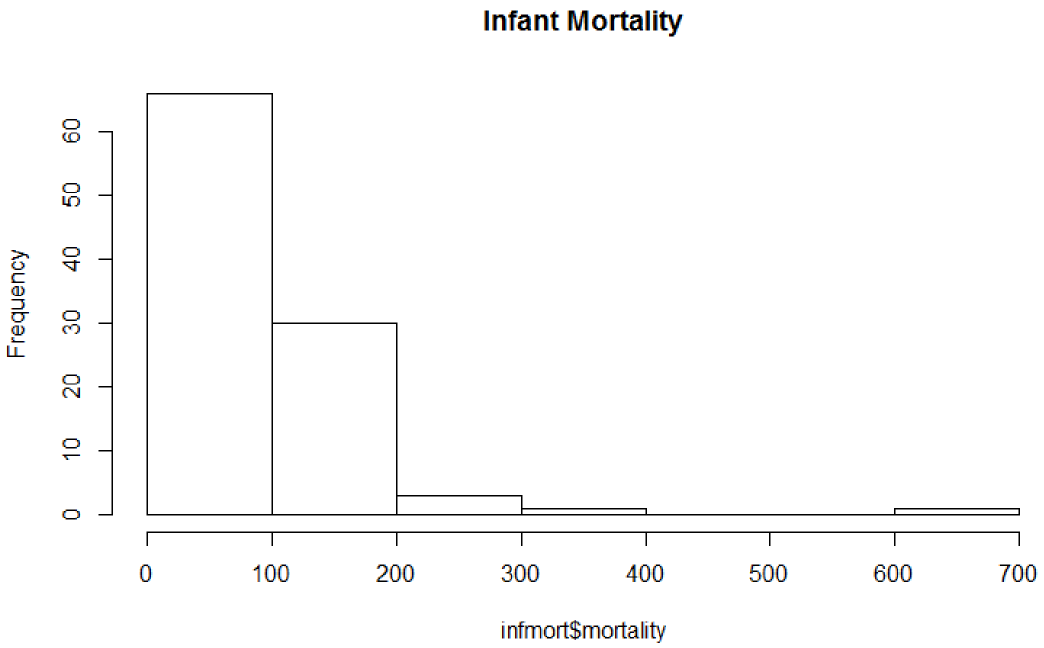
A graphical examination of the data looked at per-capita income levels in total and by region.

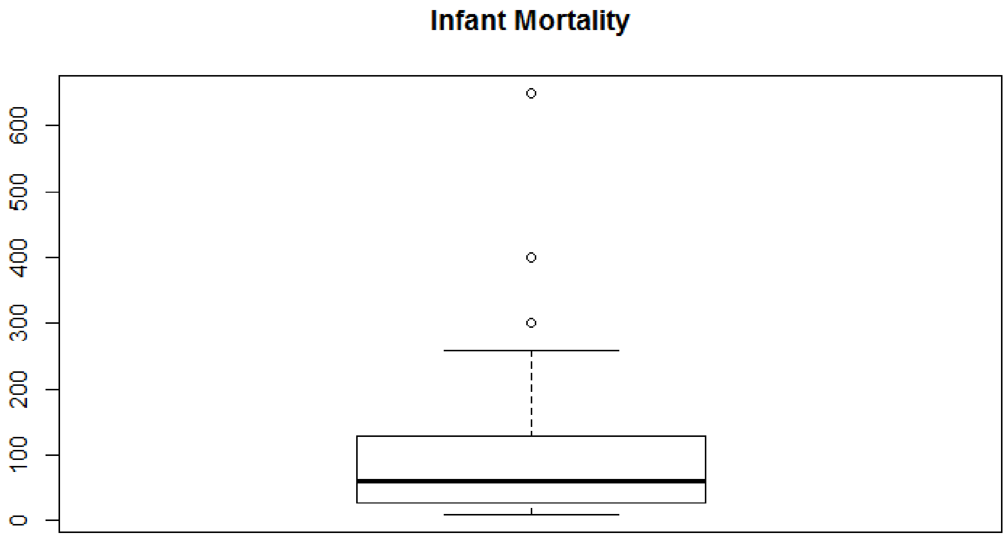


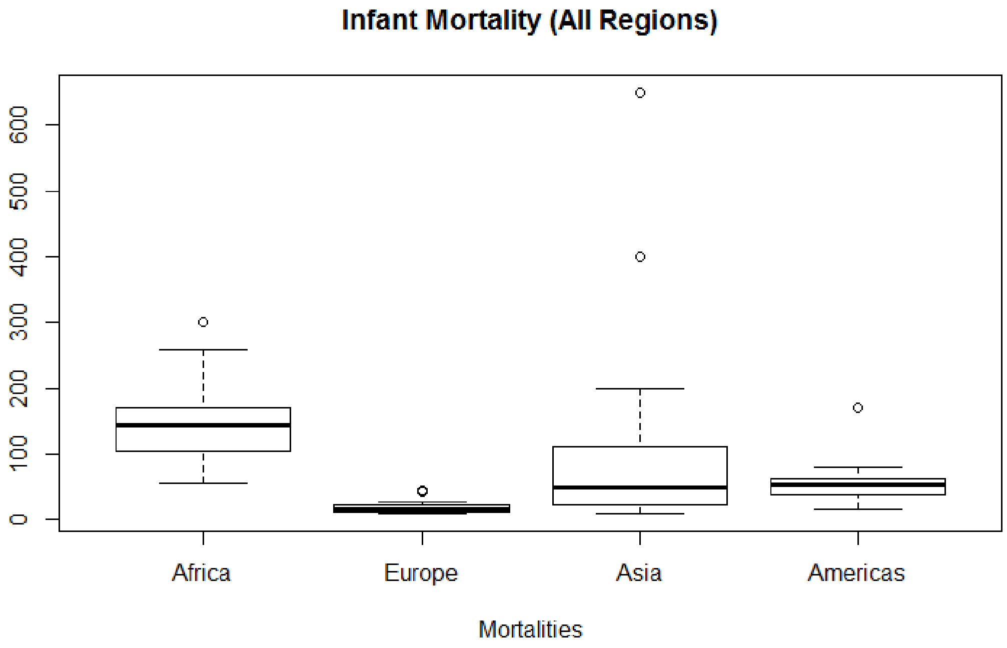




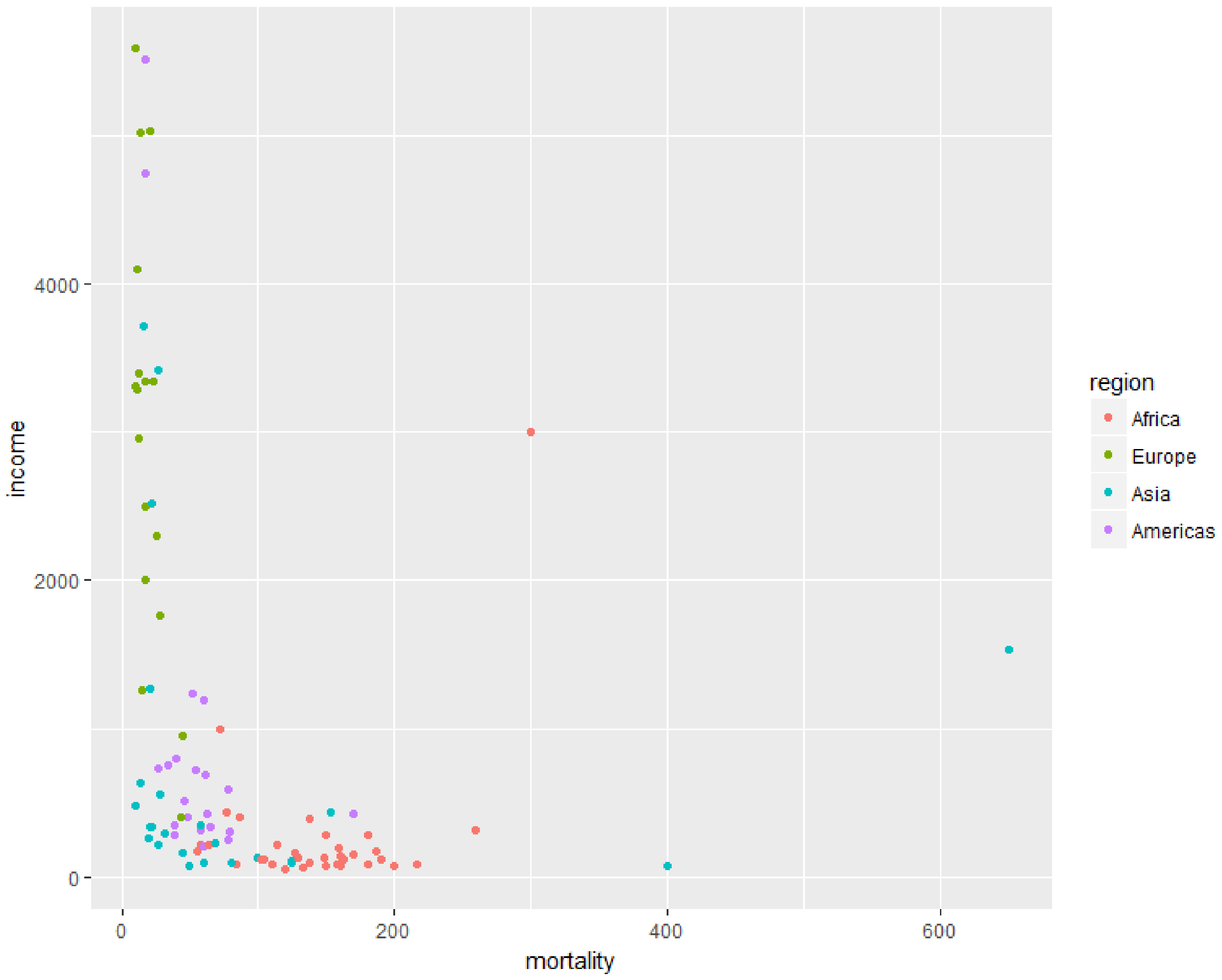
Infant mortality graphics:




Infant Mortality
0
100
200
300
400
500
600


Infant Mortality (All Regions)
0
100
200
300
400
500
600
Africa
Europe
Asia
Americas
Mortalities

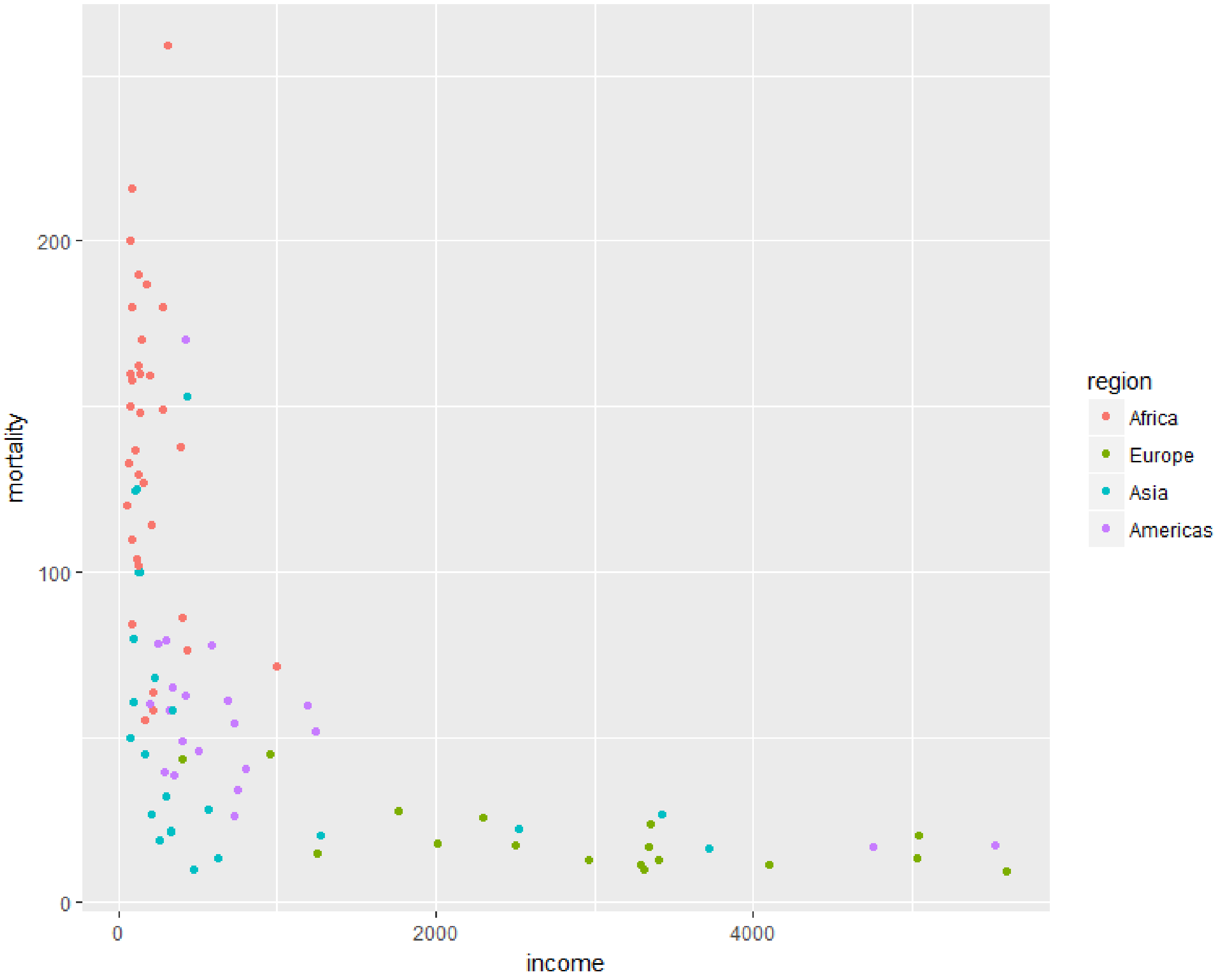
Bivariate graphics:



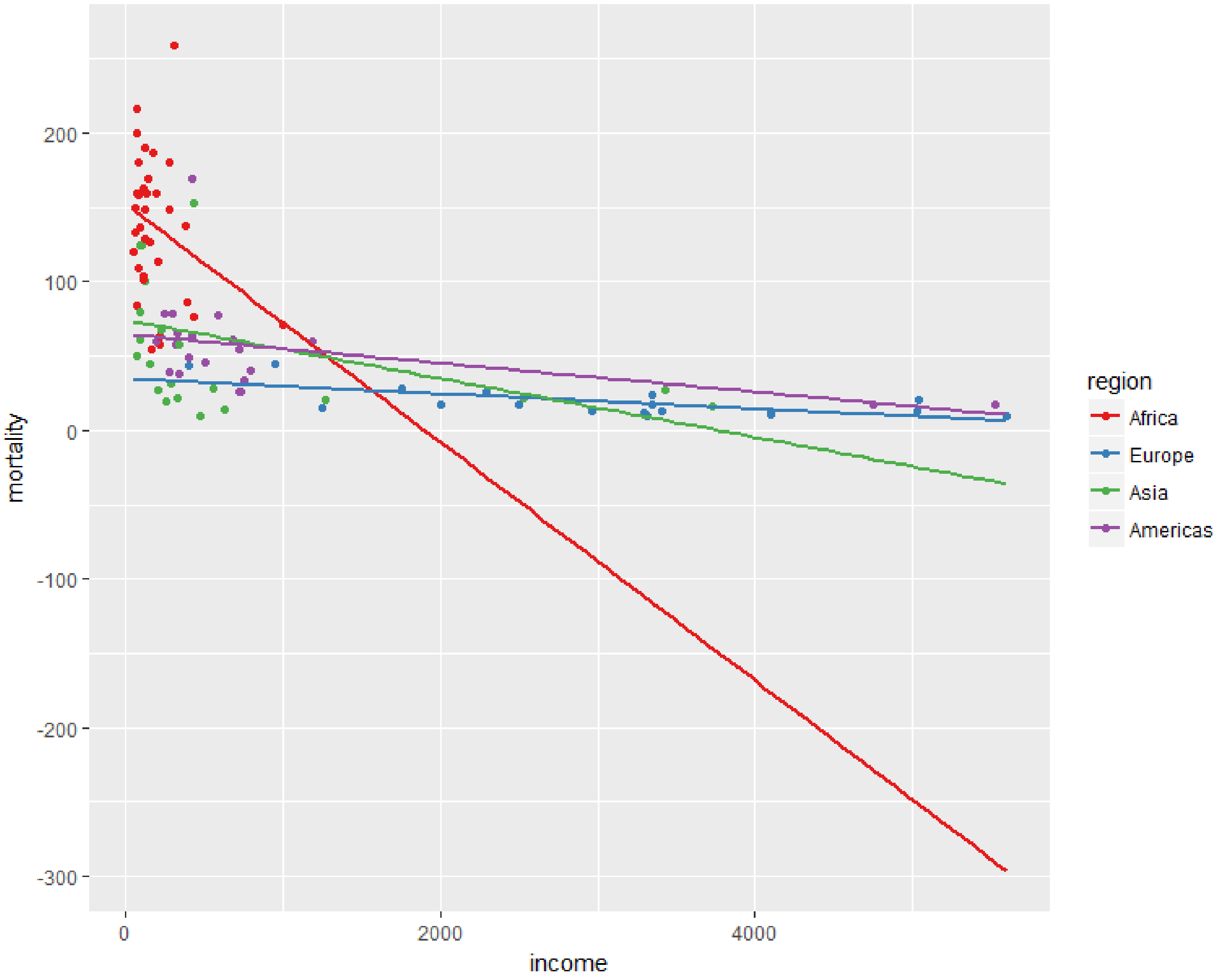
Potential outliers were removed through sub-setting (see R script) for the purpose of examining a better fit:

```
Libya                 Africa   3010     300.0    oil exports (row 26)
Saudi_Arabia            Asia   1530     650.0    oil exports (row 28)
Afganistan              Asia     75     400.0 no oil exports
```

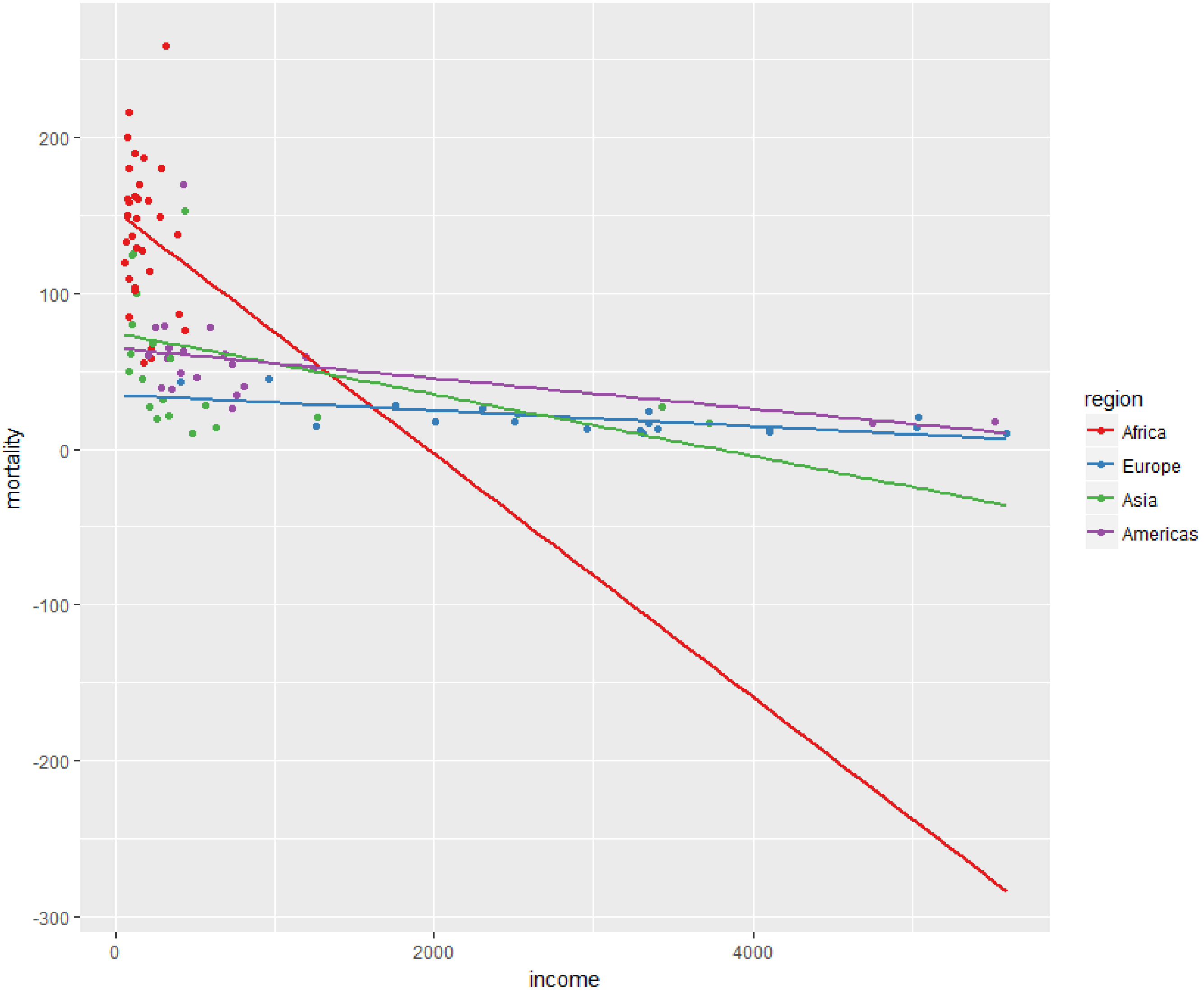
This subset is shown below. We see that despite removing these three "outliers," these two variables still appear to be non-linear:



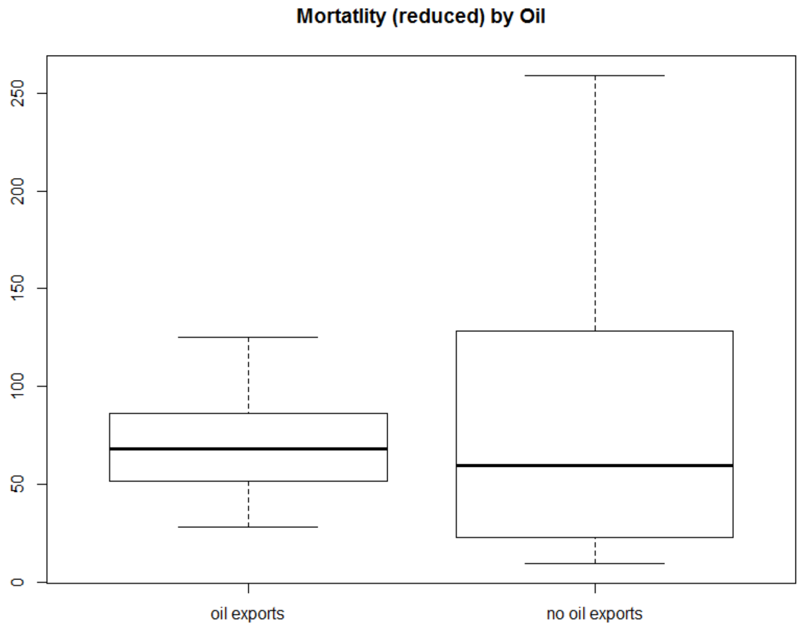
Taking the same reduced set and passing least squares lines through the data according which of the four regions they belong to was done to examine linear relationship. It would appear that Africa shows mortality almost "vertically" related to income with the exception of one additional "outlier," viz., Africa, row 16:



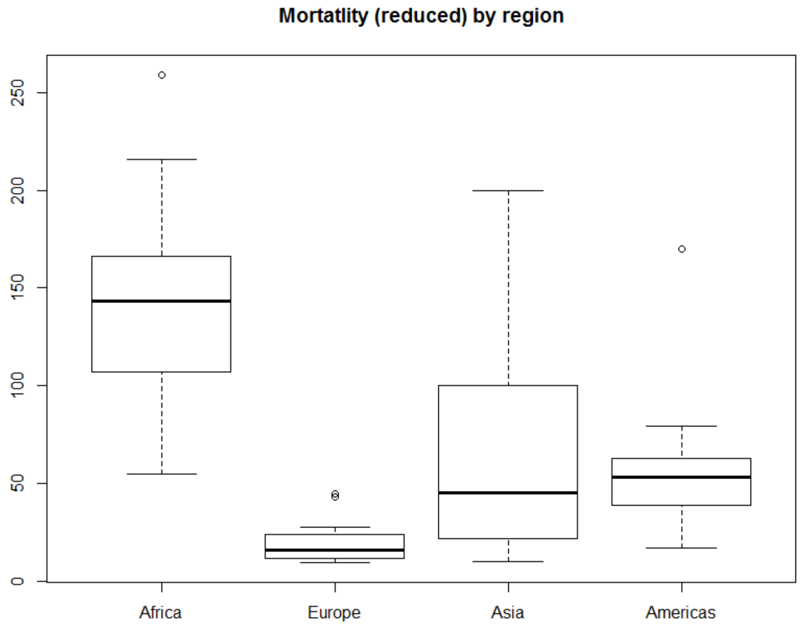
A reduced dataset was produced to examine the relationship of these two variables by removing row 16. These points with least squares fits are shown below:



We see, finally, by examining mortality per the *oil* category, there appears to be much more variation in *mortality* in region with no oil exports:



We examine also the spread of mortality by region with our reduced dataset and see that variation has been brought down some:



Having removed outliers to achieve a tighter set of representative observations, we fit a glm to investigate the effects of categories. We will use a backward-elimination approach by starting with all variables and then reducing them if necessary.

Our first blush at a model suggests that the *region* and *income* variables are significant while, perhaps surprisingly given the boxplot, the *oil* variable is not.

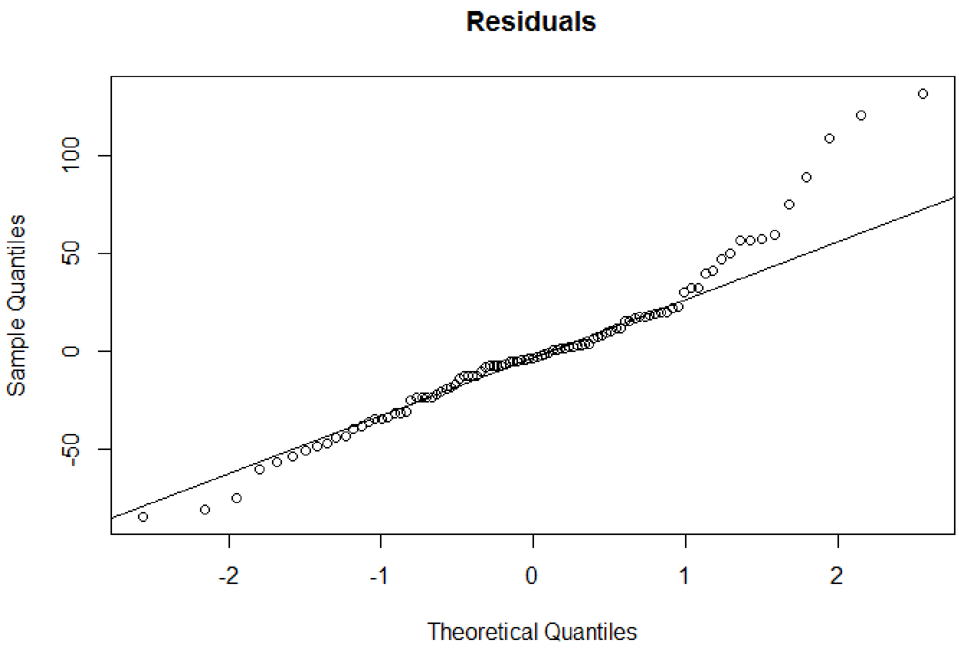
```
Deviance Residuals:
   Min       1Q     Median     3Q       Max
-85.596  -23.411  -4.787    16.485   130.515

Coefficients:
                                        Estimate      Std. Error    t value   Pr(>|t|)
(Intercept)                           125.414934   17.211361     7.287    1.12e-10 ***
infmort_reduced2$income                 -0.010945    0.003938   -2.779    0.00662 **
infmort_reduced2$regionEurope          -89.913820   16.241657   -5.536    2.96e-07 ***
infmort_reduced2$regionAsia            -72.161436   10.710814   -6.737    1.43e-09 ***
infmort_reduced2$regionAmericas        -75.069468   11.406600   -6.581    2.92e-09 ***
infmort_reduced2$oilno oil exports      17.030434   16.791198    1.014    0.31315
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1558.274)

Null deviance:        356627  on 96  degrees of freedom
Residual deviance:  141803  on 91  degrees of freedom
(4 observations deleted due to missingness)
AIC:  996.16
Number of Fisher Scoring iterations:  2
```

Re-running the model without oil at all produced a model with an AIC very close to our first model but because the AIC was slightly higher in our first model, albeit with the non-significant oil variable, we will use that one.

```
Deviance Residuals:
   Min       1Q      Median      3Q        Max
-84.532  -23.326  -3.278     16.669    131.955

Coefficients:
                                      Estimate       Std. Error    t value      Pr(>|t|)
(Intercept)                         141.358736     7.009508     20.167    < 2e-16 ***
infmort_reduced2$income               -0.010811     0.003937     -2.746     0.00725 **
infmort_reduced2$regionEurope        -89.236794   16.230465     -5.498     3.41e-07 ***
infmort_reduced2$regionAsia          -72.524112   10.706513     -6.774     1.16e-09 ***
infmort_reduced2$regionAmericas      -75.662828   11.393364     -6.641     2.14e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1558.761)

Null deviance:        356627  on 96  degrees of freedom
```

```
Residual deviance:  143406  on 92  degrees of freedom
(4 observations deleted due to missingness)
AIC:  995.25

Number of Fisher Scoring iterations:  2
```

We examine the residuals for equal variance and see with our qq-plot that most of them follow a normal distribution though with some skew to the right:



Our proposed model can be summarized as:

Mortality = 125.4 + -0.01(income) + -89.9(Europe) + -72.2(Asia) + -75.1(Americas)

We can thus interpret our model as saying that for every one-unit decrease in income by -.01, and all other variables held constant, mortality will increase accordingly. The categorical factors are all shown to have negative impacts as well with Europe's impact being the greatest.

APPENDIX: R SCRIPTS

```
> summary(infmort)
     region        income          mortality            oil
 Africa  :34    Min.   :  50.0    Min.   :  9.60    oil exports   : 9
 Europe  :18    1st Qu.: 123.0    1st Qu.: 26.20    no oil exports:96
 Asia    :30    Median : 334.0    Median : 60.60
 Americas:23    Mean   : 998.1    Mean   : 89.05
                3rd Qu.:1191.0    3rd Qu.:129.40
                Max.   :5596.0    Max.   :650.00
                                  NA's   :4

> hist(infmort$income, main="Per Capita Income (All Regions)", xlab="Income")
> boxplot(infmort$income, main="Per Capita Income (All Regions)", xlab="Income")
> boxplot(infmort$income ~ infmort$region, main="Per Capita Income (All Regions)", xlab="Income")
> hist(infmort$mortality, main="Infant Mortality")
> boxplot(infmort$mortality, main="Infant Mortality")
> plot(infmort$mortality ~ infmort$income, main="Mortality vs. Income", xlab="income", ylab = "mortality")

> infmort_reduced <- infmort[-c(26,28,73),]
> ggplot(infmort_reduced, aes(x=income, y=mortality, colour=region)) + geom_point()
> sps <- ggplot(infmort_reduced2, aes(x=income, y=mortality, colour=region)) + geom_point() +
scale_colour_brewer(palette="Set1")
> sps + geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
> boxplot(infmort_reduced2$mortality ~ infmort_reduced2$oil, main="Mortatlity (reduced) by Oil")
> boxplot(infmort_reduced2$mortality ~ infmort_reduced2$region, main="Mortatlity (reduced) by region")

> lmod_infmort <- glm(infmort_reduced2$mortality ~ infmort_reduced2$income + infmort_reduced2$region +
infmort_reduced2$oil, data=infmort_reduced2, family = gaussian())
> summary(lmod_infmort)

Call:
glm(formula = infmort_reduced2$mortality ~ infmort_reduced2$income + infmort_reduced2$region +
infmort_reduced2$oil, family = gaussian(), data = infmort_reduced2)

Deviance Residuals:
    Min       1Q    Median      3Q      Max
-85.596  -23.411   -4.787   16.485  130.515

Coefficients:
                                  Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)                     125.414934   17.211361    7.287   1.12e-10 ***
infmort_reduced2$income          -0.010945    0.003938   -2.779   0.00662 **
infmort_reduced2$regionEurope   -89.913820   16.241657   -5.536   2.96e-07 ***
infmort_reduced2$regionAsia     -72.161436   10.710814   -6.737   1.43e-09 ***
infmort_reduced2$regionAmericas -75.069468   11.406600   -6.581   2.92e-09 ***
infmort_reduced2$oilno oil exports 17.030434 16.791198   1.014   0.31315
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1558.274)

Null deviance:        356627  on 96  degrees of freedom
Residual deviance:  141803  on 91  degrees of freedom
(4 observations deleted due to missingness)
AIC:  996.16
Number of Fisher Scoring iterations:  2

> summary(lmod_infmort2 <- glm(infmort_reduced2$mortality ~ infmort_reduced2$income +
infmort_reduced2$region, data=infmort_reduced2, family = gaussian()))

Call:
glm(formula = infmort_reduced2$mortality ~ infmort_reduced2$income +
    infmort_reduced2$region, family = gaussian(), data = infmort_reduced2)

Deviance Residuals:
    Min        1Q      Median      3Q        Max
-84.532   -23.326   -3.278     16.669    131.955

Coefficients:
                                        Estimate      Std. Error    t value     Pr(>|t|)
(Intercept)                            141.358736    7.009508     20.167     < 2e-16 ***
infmort_reduced2$income                  -0.010811    0.003937     -2.746      0.00725 **
infmort_reduced2$regionEurope         -89.236794   16.230465     -5.498      3.41e-07 ***
infmort_reduced2$regionAsia           -72.524112   10.706513     -6.774     1.16e-09 ***
infmort_reduced2$regionAmericas   -75.662828   11.393364     -6.641     2.14e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1558.761)

Null deviance:        356627  on 96  degrees of freedom
Residual deviance:  143406  on 92  degrees of freedom
(4 observations deleted due to missingness)
AIC:  995.25

Number of Fisher Scoring iterations:  2

> qqnorm(residuals, main="Residuals")
> qqline(residuals)
```