World Bank Data on Fertility Rates, Infant Mortality Rates, and GNI per Capita

Mohamed Shatry and Gabriel Mazuera

December 18, 2018

For this exercise, we’re going to look at World Bank Data and see what we can learn from the data

```
WBData <- read.csv("/Users/mohamed_shatry/Desktop/World-Bank-Analysis/World Bank Data.csv")
names(WBData)
```

```
##  [1] "Country.Name"            "Country.Code"
##  [3] "Region"                  "Rural"
##  [5] "CO2.per.Capita"          "GNI.per.capita"
##  [7] "Energy.Use.2011"         "Life.Expectancy"
##  [9] "Fertility.Rate"          "Infant.Mortaility.Rate"
## [11] "Population"              "Primary.Rate.Female"
## [13] "Primary.Rate.Male"       "Exports"
## [15] "Imports"                 "Cell.Phones.per.100"
## [17] "Organic.Pollutants"      "Rural.Access.to.Water.."
## [19] "log.Exports"             "log.Imports"
## [21] "log.GNI"                 "log.Energy.2011"
## [23] "new.forest"              "log.CO2"
```

We decided to use data from the World Bank. We’re looking at relationships between Infant Mortality Rates and Fertility Rates. We’re also going to examine the role of GNI per Capita in affecting this relationship.

```
sWBData <- WBData[, c(2, 3, 9, 10)]
sWBData <- sWBData[complete.cases(WBData[, c(9, 10)]), ]
```

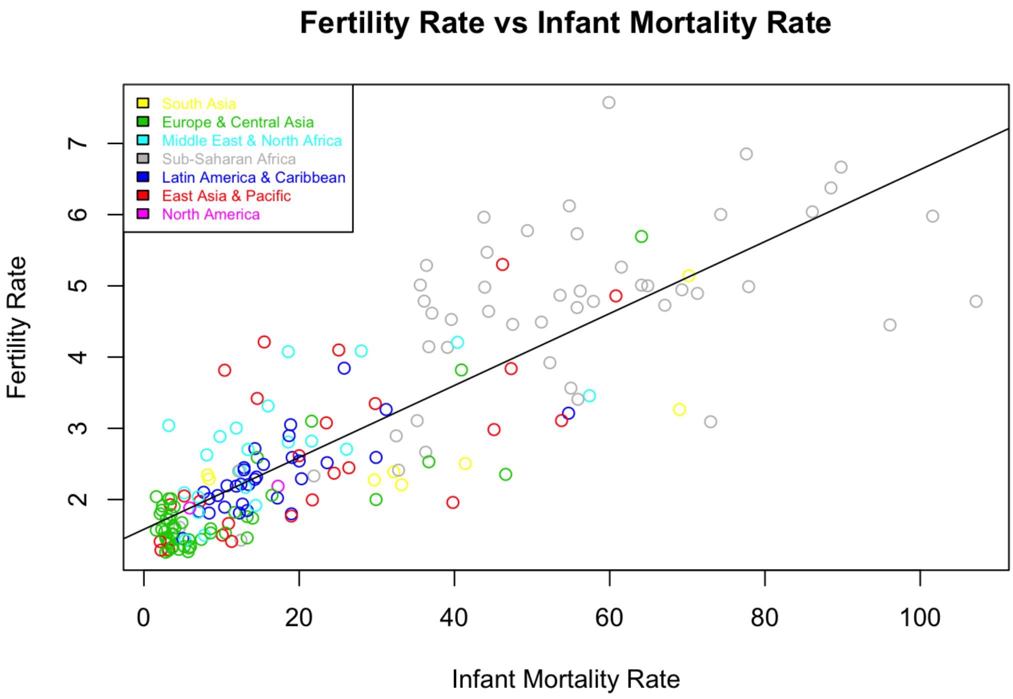
Understanding the Data

We first plotted the raw values for Infant Mortality Rates vs Fertility Rates. We tried to fit a linear regression model onto the data.

```
plot(sWBData$Infant.Mortaility.Rate, sWBData$Fertility.Rate, xlab ="Infant Mortality Rate", ylab = "Fertilit
y Rate", main = "Fertility Rate vs Infant Mortality Rate", col = as.numeric(sWBData$Region)+1)
legend("topleft", legend = unique(sWBData$Region), text.col = as.numeric(unique(sWBData$Region))+1, fill = a
s.numeric(unique(sWBData$Region))+1, cex = .6)
fit5 <- lm (sWBData$Fertility.Rate ~ sWBData$Infant.Mortaility.Rate)
summary(fit5)
```

```
## 
## Call:
## lm(formula = sWBData$Fertility.Rate ~ sWBData$Infant.Mortaility.Rate)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.21441 -0.44544 -0.06382  0.30432  2.96682 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    1.581503   0.084759   18.66   <2e-16 ***
## sWBData$Infant.Mortaility.Rate 0.050512   0.002393   21.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7839 on 184 degrees of freedom
## Multiple R-squared:  0.7077, Adjusted R-squared:  0.7061 
## F-statistic: 445.5 on 1 and 184 DF,  p-value: < 2.2e-16
```

```
abline(fit5)
```



However, there are a few problems with this plot. We want our plot to be easy to read and to be somewhat normally distributed, and we wanted to see if we could get a higher r-squared value.

Log-Adjusted Plot

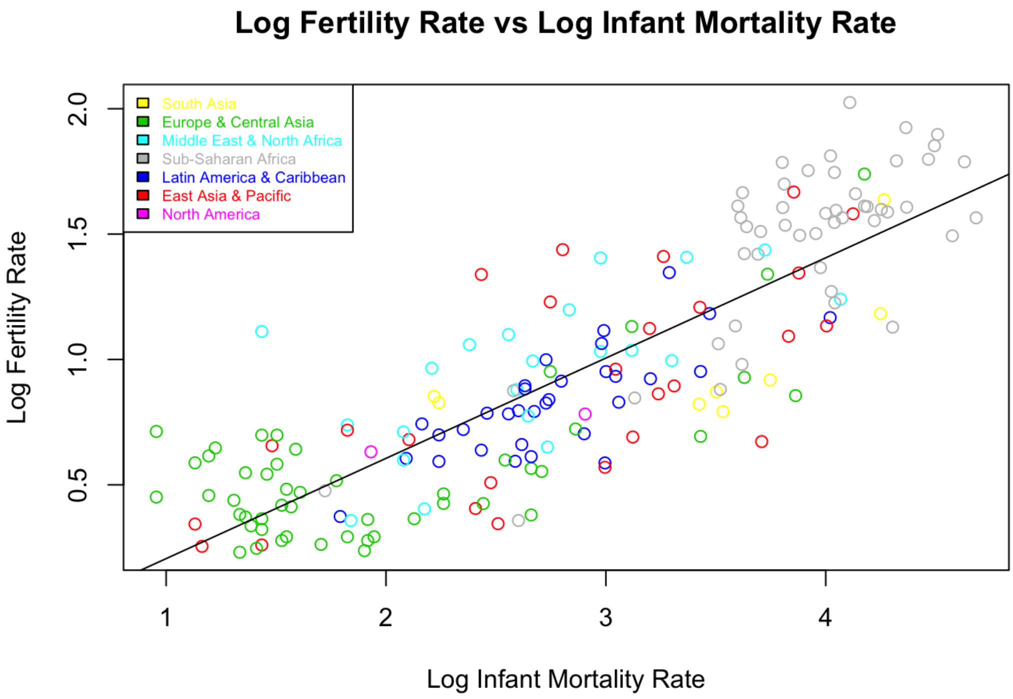
To solve these problems, we decided to transform the data by taking the log of the values, which resulted in the following plot, again with a linear regression model:

```
sWBData$logInfant <- log(1+sWBData$Infant.Mortaility.Rate)
sWBData$logFertility <- log(sWBData$Fertility.Rate)


# Plot of Log Infant Mortality sv Log Fertility Rate.
## Regression line is linear
plot(sWBData$logInfant, sWBData$logFertility, xlab ="Log Infant Mortality Rate", ylab = "Log Fertility Rate"
, main = "Log Fertility Rate vs Log Infant Mortality Rate", col = as.numeric(sWBData$Region)+1)
legend("topleft", legend = unique(sWBData$Region), text.col = as.numeric(unique(sWBData$Region))+1, fill = a
s.numeric(unique(sWBData$Region))+1, cex = .6)
fit0 <- lm(sWBData$logFertility ~ sWBData$logInfant)
summary(fit0)
```

```
##
## Call:
## lm(formula = sWBData$logFertility ~ sWBData$logInfant)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61694 -0.19908  0.00167  0.17474  0.73216
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.19449    0.05662  -3.435 0.000733 ***
## sWBData$logInfant  0.40011    0.01871  21.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2511 on 184 degrees of freedom
## Multiple R-squared:  0.7132, Adjusted R-squared:  0.7116
## F-statistic: 457.5 on 1 and 184 DF,  p-value: < 2.2e-16
```

```
abline(fit0)
```

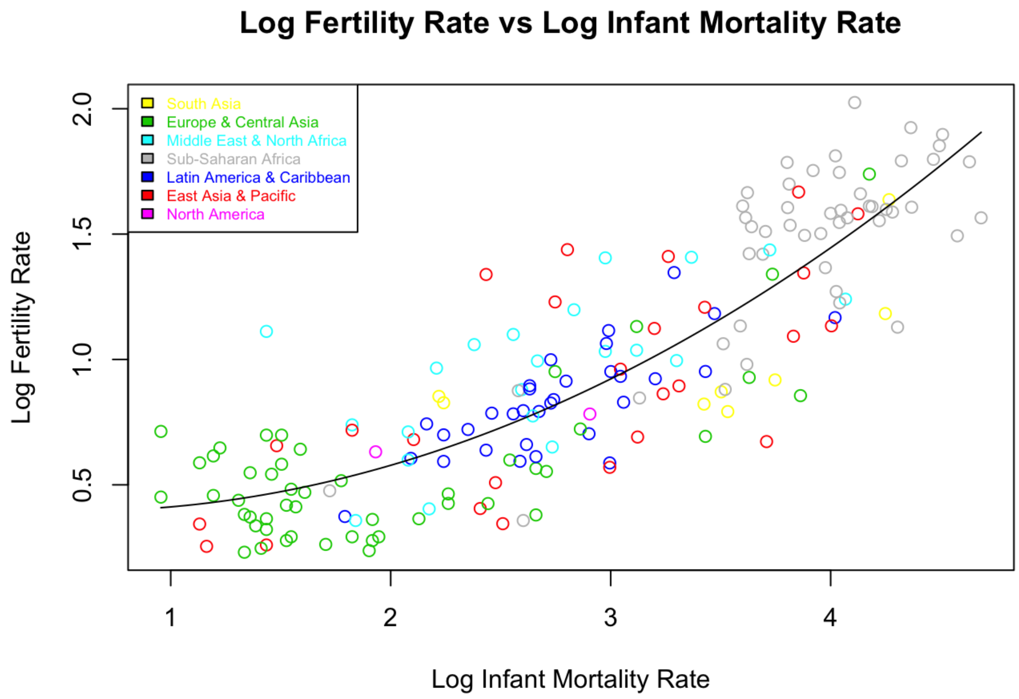


Here the data is much easier to read, and log-adjusting makes the distribution more normal. We wanted to see if a quadratic model was a better fit for a least-squared residual line:

```
plot(sWBData$logInfant, sWBData$logFertility, xlab ="Log Infant Mortality Rate", ylab = "Log Fertility Rate"
, main = "Log Fertility Rate vs Log Infant Mortality Rate", col = as.numeric(sWBData$Region)+1)
legend("topleft", legend = unique(sWBData$Region), text.col = as.numeric(unique(sWBData$Region))+1, fill = a
s.numeric(unique(sWBData$Region))+1, cex = .6)
fit <- lm(sWBData$logFertility ~ sWBData$logInfant + I(sWBData$logInfant^2))
summary(fit)
```

```
##
## Call:
## lm(formula = sWBData$logFertility ~ sWBData$logInfant + I(sWBData$logInfant^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60203 -0.18051  0.00923  0.15277  0.64968
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.42456    0.14337   2.961  0.00347 **
## sWBData$logInfant      -0.10190    0.10925  -0.933  0.35220
## I(sWBData$logInfant^2)  0.08927    0.01917   4.657 6.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2381 on 183 degrees of freedom
## Multiple R-squared:  0.7435, Adjusted R-squared:  0.7407
## F-statistic: 265.3 on 2 and 183 DF,  p-value: < 2.2e-16
```

```
grid <- seq(min(sWBData$logInfant, na.rm=TRUE), max(sWBData$logInfant, na.rm=TRUE), length.out=100)
quad <- fit$coefficients[1] + fit$coefficients[2]*grid + fit$coefficients[3]*grid^2
lines(grid, quad, type="l")
```



The new line of best fit seems more appropiate for the pattern that the data shows, and it has a higer r-squared value of 0.7435, compared to 0.7132 for the previous line.

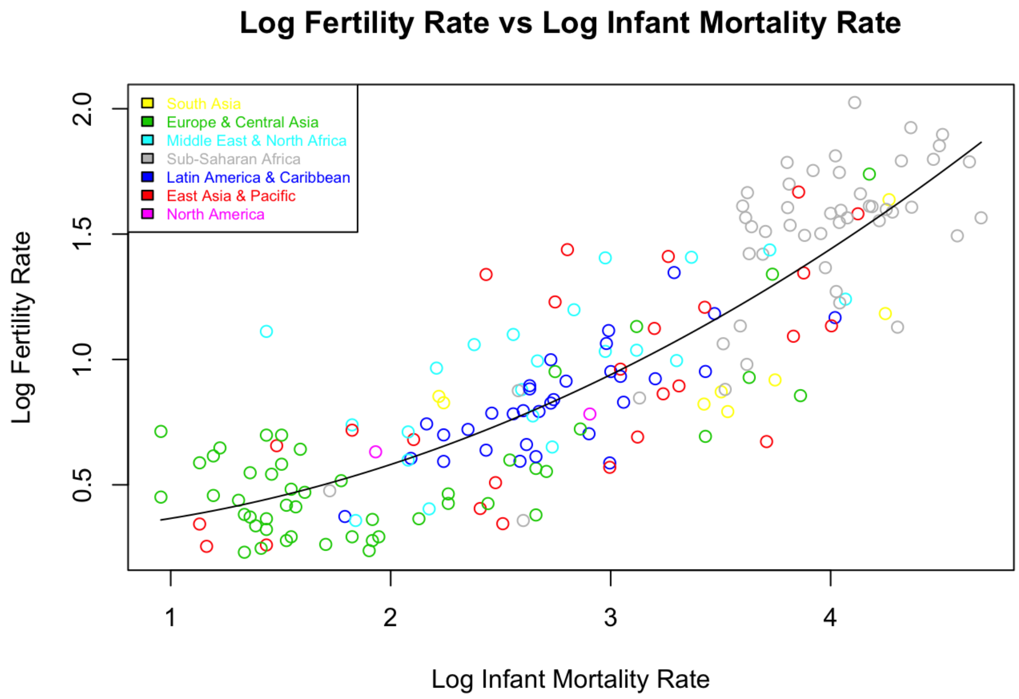
Adjusted R-squared values

Looking at the summary statistics for our new line of best fit, we see that there's one significance value that stands out. The p-value for the coefficient of the linear term (sWBData$logInfant) being zero is surprisingly high, at 0.35220. Since we would fail to reject the null hypothesis that this coefficient is zero at any reasonable alpha-level, we decided to see what would change if we removed this term.

```
plot(sWBData$logInfant, sWBData$logFertility, xlab ="Log Infant Mortality Rate", ylab = "Log Fertility Rate"
, main = "Log Fertility Rate vs Log Infant Mortality Rate", col = as.numeric(sWBData$Region)+1)
legend("topleft", legend = unique(sWBData$Region), text.col = as.numeric(unique(sWBData$Region))+1, fill = a
s.numeric(unique(sWBData$Region))+1, cex = .6)
fit2 <- lm(sWBData$logFertility ~ I(sWBData$logInfant^2))
summary(fit2)
```

```
##
## Call:
## lm(formula = sWBData$logFertility ~ I(sWBData$logInfant^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60725 -0.18409  0.00784  0.14857  0.66982
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.294531   0.033421   8.813 9.08e-16 ***
## I(sWBData$logInfant^2) 0.071625   0.003111  23.023  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.238 on 184 degrees of freedom
## Multiple R-squared:  0.7423, Adjusted R-squared:  0.7409
## F-statistic: 530.1 on 1 and 184 DF,  p-value: < 2.2e-16
```

```
grid <- seq(min(sWBData$logInfant, na.rm=TRUE), max(sWBData$logInfant, na.rm=TRUE), length.out=100)
quad <- fit2$coefficients[1] + fit2$coefficients[2]*grid^2
lines(grid, quad, type="l")
```

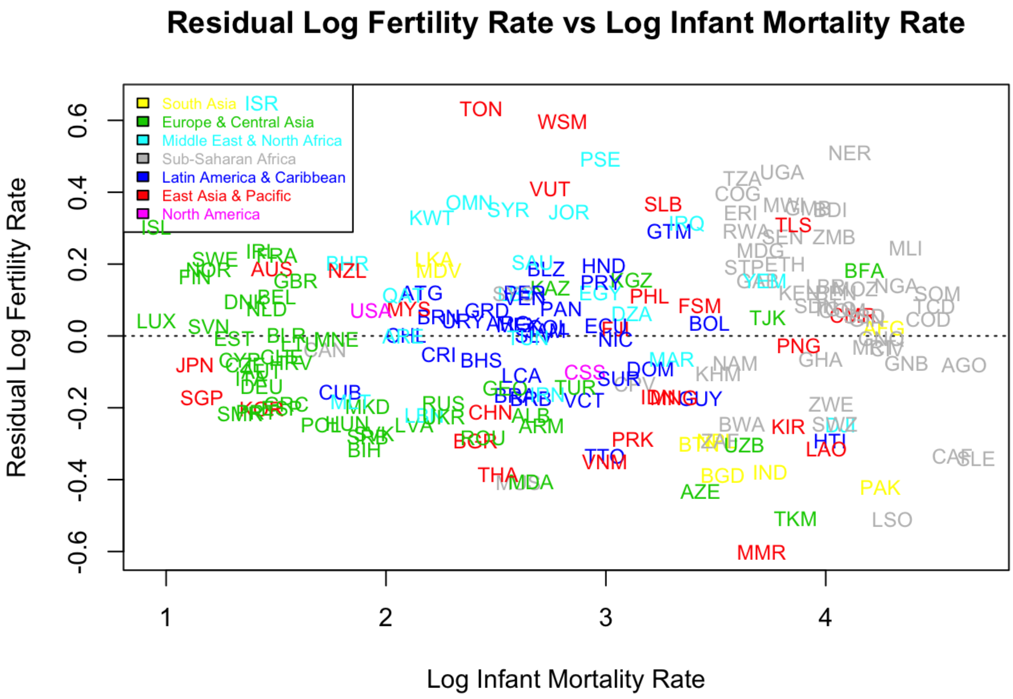


When we looked at the summary statistics for the line without the aforementioned term, we saw a slight increase in the adjusted r-squared value, but a drop in the r-squared value. This is because the adjusted r-squared value takes into account the fact that it is easier to reach higher r-squared values by simply adding more terms to the formula of the line of best fit. For example, our line with three terms will always have a r-squared value greater than or equal to that of our line with two terms, because there's always the option of having the linear term's coefficient be zero. The r-squared value takes this into account, so the fact that we got a higher adjusted r-squared value after removing the linear term means that the linear term wasn't adding enough significance to justify having that term. This conclusion lines up with the fact that the p-value was so high that we couldn't reject the null hypothesis of the coefficient being zero. The differences are minimal but telling.

Residual plot and sub-Saharan Africa

Moving on, we plotted the Log Infant Mortality Rates vs the residuals of Log Fertility Rate as described by Log Infant Mortality Rates. We kept the three-term line of best fit in this case. The countries are labeled by the country code assigned to them by the World Bank.

```
plot(sWBData$logInfant, fit$residuals, type = "n", xlab ="Log Infant Mortality Rate", ylab = "Residual Log Fertility Rate", main = "Residual Log Fertility Rate vs Log Infant Mortality Rate", col = as.numeric(sWBData$Region)+1)
legend("topleft", legend = unique(sWBData$Region), text.col = as.numeric(unique(sWBData$Region))+1, fill = as.numeric(unique(sWBData$Region))+1, cex = .6)
abline(h=0, lty=3)
text(sWBData$logInfant, fit$residuals, labels=sWBData$Country.Code, col = as.numeric(sWBData$Region)+1, cex=0.8)
```



Most regions seem to be spread out reasonably normally, but Sub-Saharan African countries seem to be mostly above the residual line– which means that they have a higher fertility rate than what was predicted by the Infant Mortality Rate. There are a few countries that don’t follow this pattern, for example, Lesotho is below the residual line and the lowest of all the African countries.

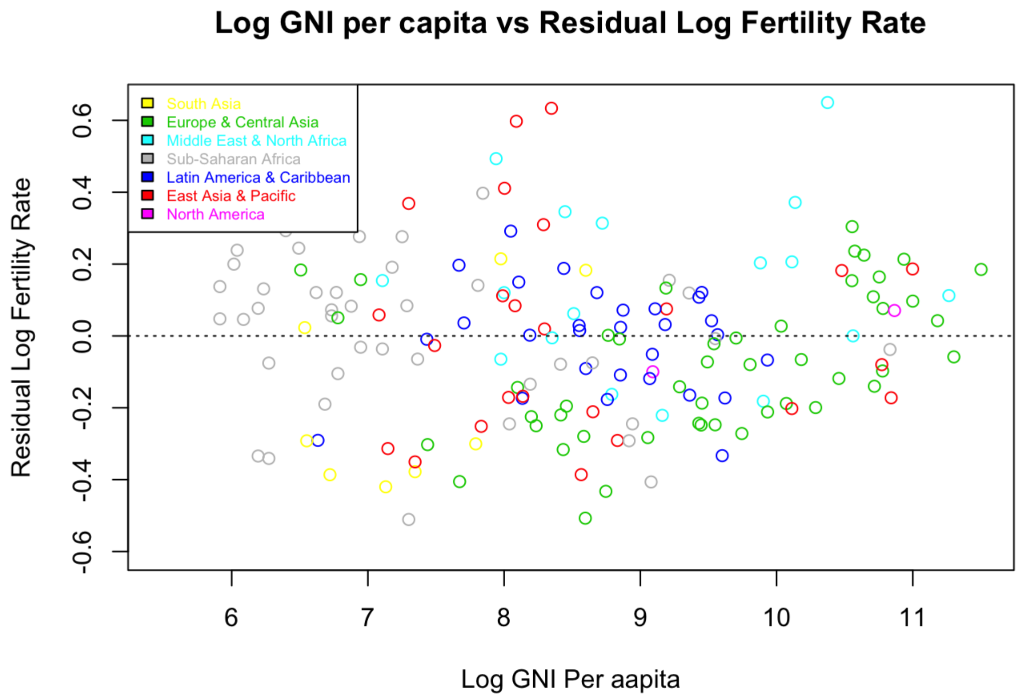
We did research on Lesotho, and why it has a relatively lower fertility rate in terms of its mortality rate. We think that this can give us some insight as to why some African countries don’t follow as the same pattern as most of the continent. We learned that Lesotho has a high prevalence of HIV, which has led to increased government efforts to control the spread of this disease. The government launced an initiative to increase the use of contraceptives in order to stop HIV from being passed on. The rate of contraceptive use rose to 47% by 2011, which likely lowered the fertility rate as well. This is probably one of the biggest reasons why Lesotho has a lower fertility rate relative to its infant mortality rate than both the line of best fit’s prediction and its sub-Saharan African neighbors. (World Bank Lesotho Country Brief)

On GNI per Capita

Finally, we wanted to see what impact GNI per capita would have on this relationship between fertility and infant mortality rates. We know that at face value, GNI per capita has a significant correlation with both fertility and infant mortality. However, we cannot truly establish a line of causation without experimentation, so we won’t know if it’s GNI per capita affecting fertility and infant mortality, or GNI affecting infant mortality which in turn affects fertility.

To see whether GNI per capita had a significant effect on the relationship between fertility rates and infant mortality rates, we plotted GNI per capita vs the residuals of Log Fertilty rates as predicted by Log Infant Mortality rates.

```
z <- WBData[complete.cases(WBData[, c(9, 10)]), ]
plot(log(z$GNI.per.capita), fit$residuals, xlab ="Log GNI Per aapita", ylab = "Residual Log Fertility Rate",
main = "Log GNI per capita vs Residual Log Fertility Rate", col = as.numeric(sWBData$Region)+1)
legend("topleft", legend = unique(sWBData$Region), text.col = as.numeric(unique(sWBData$Region))+1, fill = a
s.numeric(unique(sWBData$Region))+1, cex = .6)
abline(h=0, lty=3)
```



```
##text(log(z$GNI.per.capita), fit$residuals, labels=sWBData$Country.Code, col = as.numeric(sWBData$Region)+1
, cex=0.8)
```

This residual plot shows no significant relationship or pattern, so we concluded that once we account for infant mortality, GNI per capita has no real impact on fertility rates. This is interesting because with raw values, GNI is a useful predictor for both fertility rates and infant mortality rates.

Conclusions

Our conclusion is that infant mortality rate seems to account for about 75% of the variation in fertility rates for countries. In the case of Lesotho, we saw how specific factors can influence a country’s fertility rate compared to its neighbors. We believe there are many other lurking factors that affect fertility, such as culture which is difficult to categorize or quantify.

Works Cited

World Bank, Lesotho: Country Brief. http://go.worldbank.org/3D2PCW5DC0