WB Capstone Project

Kathryn Gadberry

June 26, 2016

```
Springboard Capstone Final: Reducing Poverty in Asia
```

Set-up

First, I will need to install the proper R packages to do my analysis. This includes the wbstats package, which delivers World Bank data direclty to R Studio.

```
suppressMessages(library(dplyr))
suppressMessages(library(tidyr))
install.packages("wbstats", repos = "http://cran.us.r-project.org")
```

```
## package 'wbstats' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\KGadberry\AppData\Local\Temp\RtmpQRhyZe\downloaded_packages
```

```
suppressMessages(library(wbstats))
```

```
## Warning: package 'wbstats' was built under R version 3.2.5
```

```
suppressMessages(library(ggplot2))
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

I'm going to single out the countries, indicators, and timeframe from the World Bank data and assign it to the object: WB_data.

```
WB_data <- wb(country = c("BD", "CN", "ID", "IN", "KH", "MM", "NP", "PH", "TH", "VN"), indicator
 = c("EG.ELC.ACCS.RU.ZS", "SH.H2O.SAFE.RU.ZS", "SH.STA.ACSN.RU", "SL.AGR.EMPL.FE.ZS", "SI.POV.RU
GP", "SI.POV.GAPS", "SP.RUR.TOTL", "SP.RUR.TOTL.ZS",  "SP.POP.TOTL"), startdate = 1985, enddate
= 2015)
```

Now, I want to inspect the data frame and rearrange it in a more logical order.

```
glimpse(WB_data)
```

```
## Observations: 1,712
## Variables: 6
## $ value       (dbl) 49.30000, 42.50000, 20.50000, 10.40000, 100.00000,...
## $ date        (chr) "2012", "2010", "2000", "1990", "2012", "2010", "2...
## $ indicatorID (chr) "EG.ELC.ACCS.RU.ZS", "EG.ELC.ACCS.RU.ZS", "EG.ELC....
## $ indicator   (chr) "Access to electricity, rural (% of rural populati...
## $ iso2c       (chr) "BD", "BD", "BD", "BD", "CN", "CN", "CN", "CN", "I...
## $ country     (chr) "Bangladesh", "Bangladesh", "Bangladesh", "Banglad...
```

```
WB_data$date <- as.numeric(WB_data$date)
str(WB_data)
```

```
## 'data.frame':    1712 obs. of  6 variables:
##  $ value      : num  49.3 42.5 20.5 10.4 100 98 95.3 92 92.9 89.4 ...
##  $ date       : num  2012 2010 2000 1990 2012 ...
##  $ indicatorID: chr  "EG.ELC.ACCS.RU.ZS" "EG.ELC.ACCS.RU.ZS" "EG.ELC.ACCS.RU.ZS" "EG.ELC.ACC
S.RU.ZS" ...
##  $ indicator  : chr  "Access to electricity, rural (% of rural population)" "Access to electr
icity, rural (% of rural population)" "Access to electricity, rural (% of rural population)" "Ac
cess to electricity, rural (% of rural population)" ...
##  $ iso2c      : chr  "BD" "BD" "BD" "BD" ...
##  $ country    : chr  "Bangladesh" "Bangladesh" "Bangladesh" "Bangladesh" ...
```

```
ASP_data <- WB_data %>%
select(date, iso2c, country, indicatorID, indicator, value) %>%
arrange(date, indicator, value)

head(ASP_data)
```

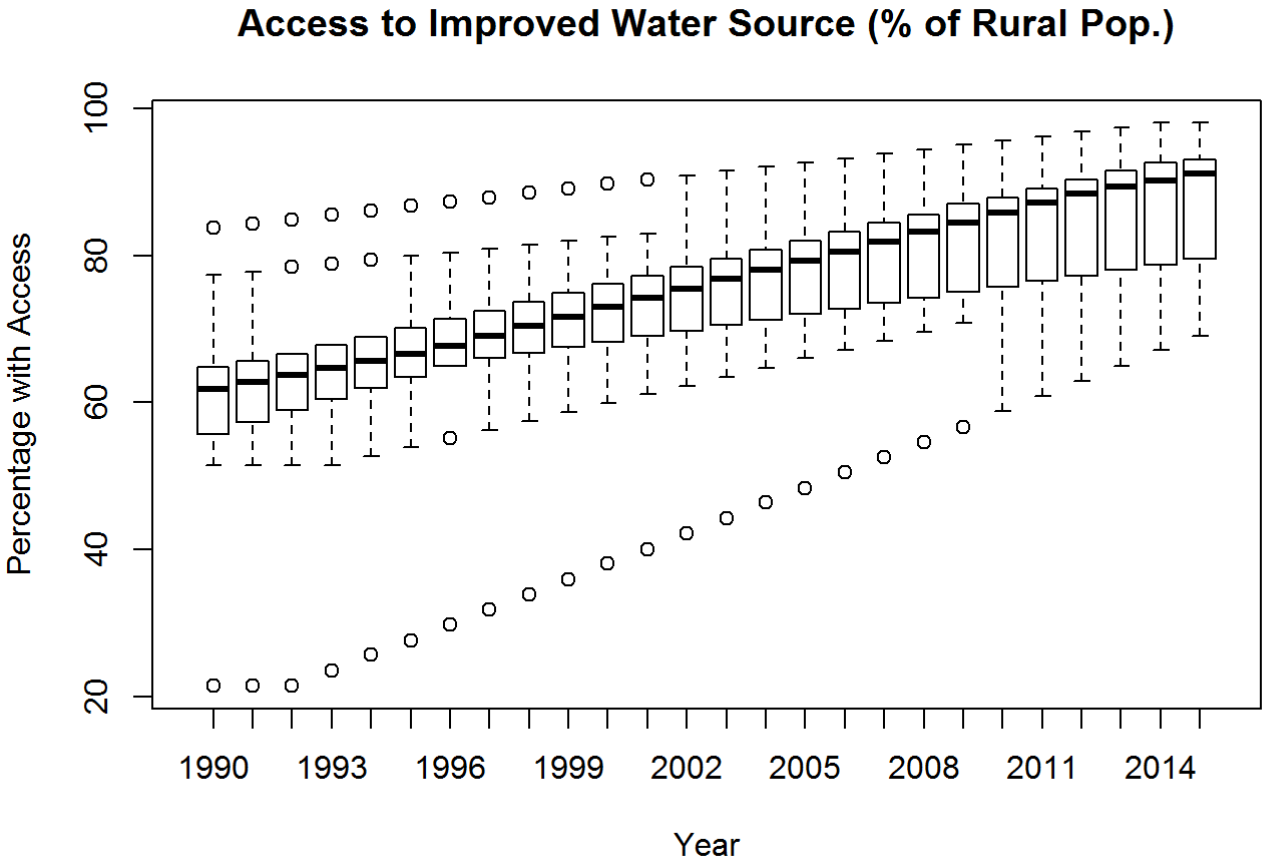
```
##   date iso2c     country       indicatorID
## 1 1985    BD  Bangladesh SL.AGR.EMPL.FE.ZS
## 2 1985    PH Philippines SL.AGR.EMPL.FE.ZS
## 3 1985    ID   Indonesia SL.AGR.EMPL.FE.ZS
## 4 1985    KH    Cambodia       SP.POP.TOTL
## 5 1985    NP       Nepal       SP.POP.TOTL
## 6 1985    MM     Myanmar       SP.POP.TOTL
##                                                    indicator      value
## 1 Employment in agriculture, female (% of female employment)        9.3
## 2 Employment in agriculture, female (% of female employment)       35.0
## 3 Employment in agriculture, female (% of female employment)       53.6
## 4                                          Population, total  7743065.0
## 5                                          Population, total 16714335.0
## 6                                          Population, total 38508821.0
```

Exploratory Data Analysis

% with Access to Improved Water Source

Access to an improved water source is the first World Bank indicator I'm going to investigate. I'll assign all data associated with this development metric to the object "ASP_water", filtered for the last 30 years. Then, I'll need to factor the date column as.numeric to plot the data. Note that all of these indicators are descriptive of the rural population % in this region.

```
ASP_water <- wb(country = c("BD", "CN", "ID", "IN", "KH", "MM", "NP", "PH", "TH", "VN"), indicat
or = c("SH.H2O.SAFE.RU.ZS"), startdate = 1985, enddate = 2015)

ASP_water$date <- as.numeric(ASP_water$date)

water_plot <- boxplot(value ~ date, data = ASP_water, main = "Access to Improved Water Source (%
 of Rural Pop.)", xlab = "Year", ylab = "Percentage with Access")
```



% with Access to Electricity

I'm going to use the same steps as I did to investigate water, but for the World Bank Indicator: percentage of the rural population with access to electricity.

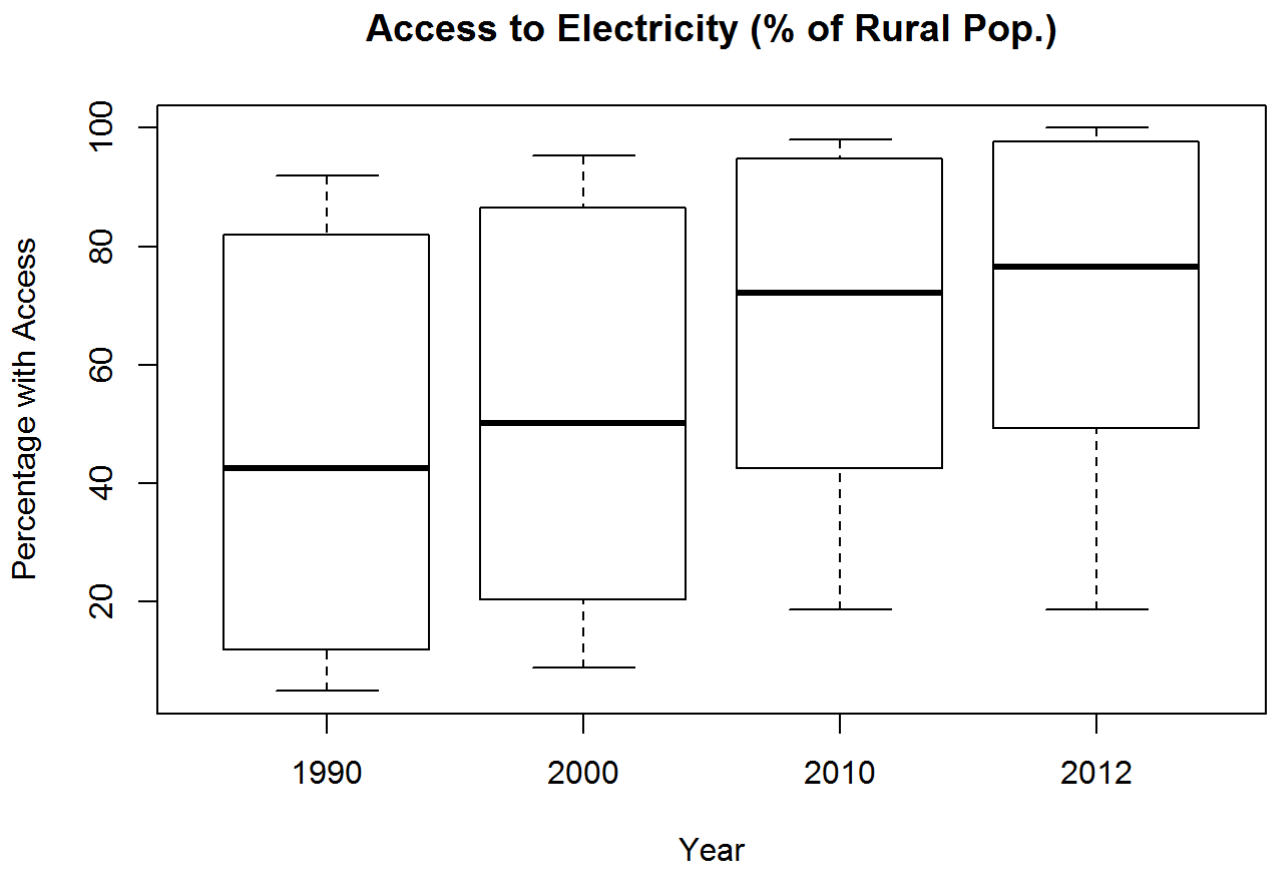
```
ASP_electric <- wb(country = c("BD", "CN", "ID", "IN", "KH", "MM", "NP", "PH", "TH", "VN"), indi
cator = c("EG.ELC.ACCS.RU.ZS"), startdate = 1985, enddate = 2015)

summary(ASP_electric)
```

```
##      value            date           indicatorID         indicator
##  Min.   :  5.00   Length:40          Length:40          Length:40
##  1st Qu.: 27.22   Class :character   Class :character   Class :character
##  Median : 68.28   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 58.65
##  3rd Qu.: 87.60
##  Max.   :100.00
##     iso2c              country
##  Length:40          Length:40
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
ASP_electric$date <- as.numeric(ASP_electric$date)
```

Using a boxplot will help me see what the minimum, maximum, and average percentages were over time for the rural asain population. This plot shows us that…

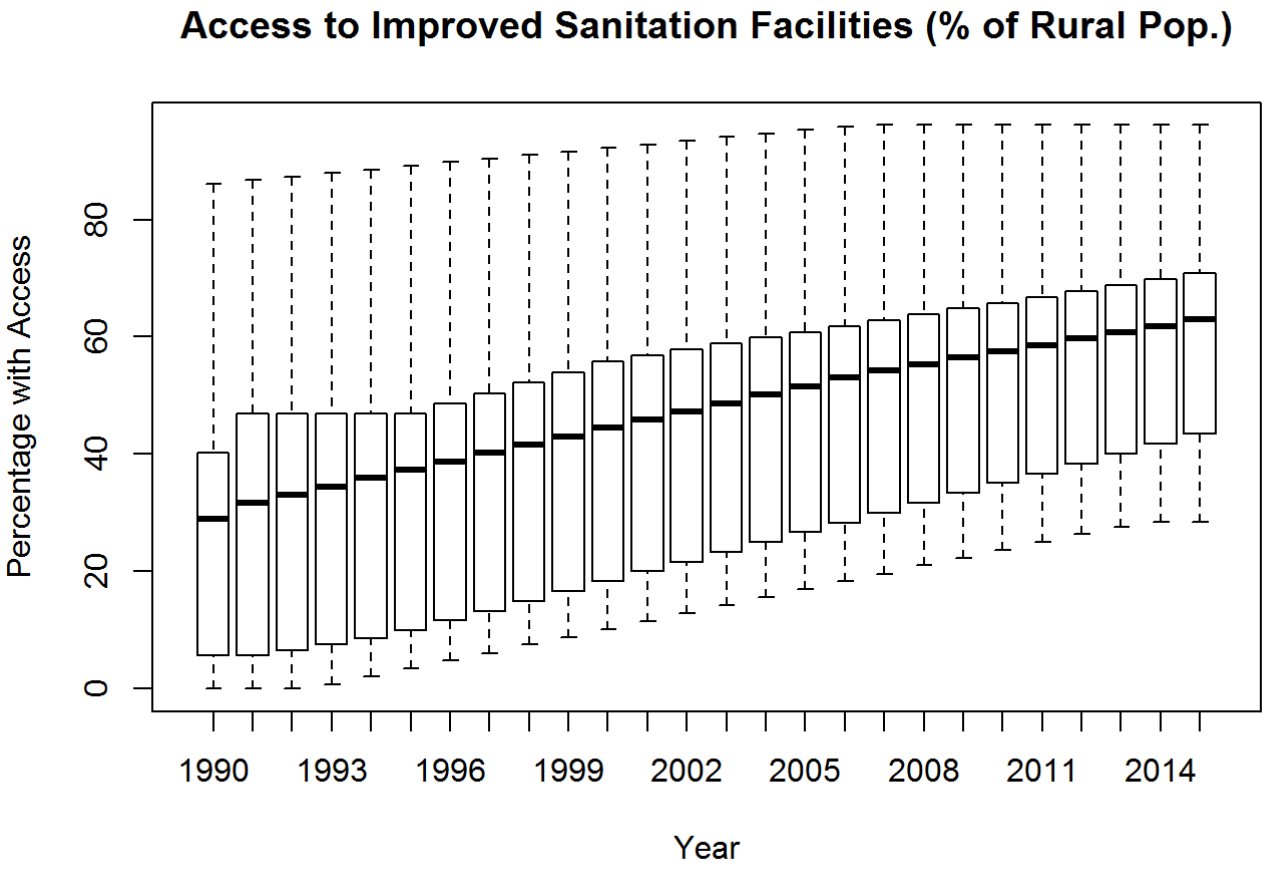
```
electric_plot <- boxplot(value ~ date, data = ASP_electric,
main = "Access to Electricity (% of Rural Pop.)",
xlab = "Year", ylab = "Percentage with Access")
```



% Access to Improved Sanitation Facilities

I'll also perform the same analysis for the third indicator: percentage of rural population with access to improved sanitation facilities.

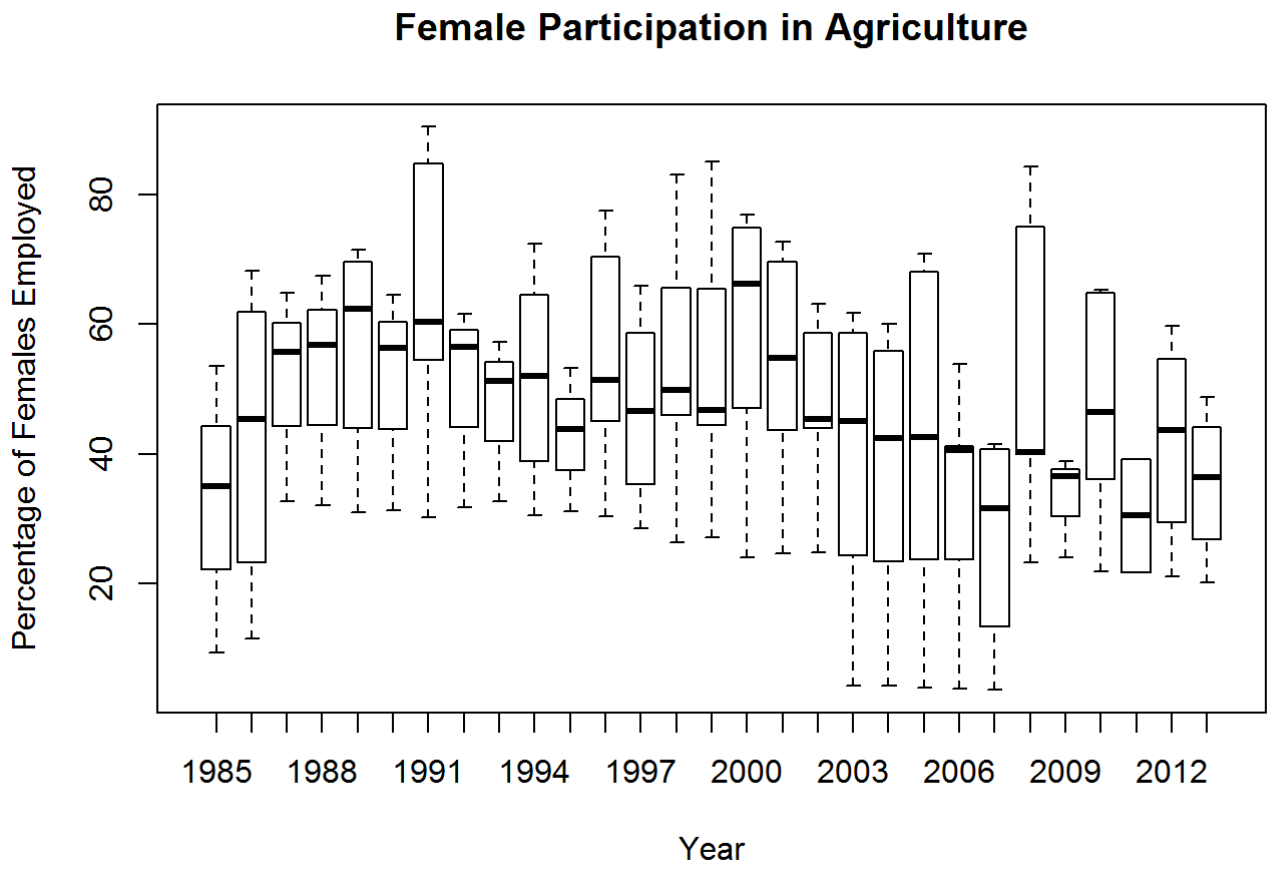
```
ASP_sanitation <- wb(country = c("BD", "CN", "ID", "IN", "KH", "MM", "NP", "PH", "TH", "VN"), in
dicator = c("SH.STA.ACSN.RU"), startdate = 1985, enddate = 2015)

ASP_sanitation$date <- as.numeric(ASP_sanitation$date)

sanitation_plot <- boxplot(value ~ date, data = ASP_sanitation, main = "Access to Improved Sanit
ation Facilities (% of Rural Pop.)", xlab = "Year", ylab = "Percentage with Access")
```



Female Participation in Agriculture

And finally, I'll end my exploratory boxplot analysis with the last indicator: percentage of women employed in agriculture.

```
ASP_women <- wb(country = c("BD", "CN", "ID", "IN", "KH", "MM", "NP", "PH", "TH", "VN"), indicat
or = c("SL.AGR.EMPL.FE.ZS"), startdate = 1985, enddate = 2015)

ASP_women$date <- as.numeric(ASP_women$date)

women_plot <- boxplot(value ~ date, data = ASP_women, main = "Female Participation in Agricultur
e", xlab = "Year", ylab = "Percentage of Females Employed")
```



Data Wrangling

In order to clean up the data for analysis, I need to re-structure ASP_data to show each indicator as its own variable (or column) with the values (averages) listed below.

```
ASP_2 <- tapply(ASP_data$value, list(date = ASP_data$date, ticker = ASP_data$indicatorID), mean)

glimpse(ASP_2)
```

```
## Observations: 31
## Variables: 9
## $ EG.ELC.ACCS.RU.ZS (dbl) NA, NA, NA, NA, NA, 44.01299, NA, NA, NA, NA...
## $ SH.H2O.SAFE.RU.ZS (dbl) NA, NA, NA, NA, NA, 59.84, 60.67, 61.49, 62....
## $ SH.STA.ACSN.RU    (dbl) NA, NA, NA, NA, NA, 29.24444, 31.84000, 32.7...
## $ SI.POV.GAPS       (dbl) 14.240000, NA, 20.610000, 11.343333, NA, 14....
## $ SI.POV.RUGP       (dbl) NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ SL.AGR.EMPL.FE.ZS (dbl) 32.63333, 42.60000, 51.06667, 52.10000, 56.7...
## $ SP.POP.TOTL       (dbl) 231935184, 236375699, 240989349, 245643934, ...
## $ SP.RUR.TOTL       (dbl) 176808593, 178987700, 181233468, 183440733, ...
## $ SP.RUR.TOTL.ZS    (dbl) 77.3098, 76.8750, 76.4319, 75.9797, 75.5133,...
```

Renaming columns from the hard to interpret Indicator IDs to simple variable names.

```
colnames(ASP_2)[1] <- "electricity"
colnames(ASP_2)[2] <- "water"
colnames(ASP_2)[3] <- "sanitation"
colnames(ASP_2)[4] <- "natl.pg.per"
colnames(ASP_2)[5] <- "rural.pg.per"
colnames(ASP_2)[6] <- "women.ag"
colnames(ASP_2)[7] <- "avg.total.pop"
colnames(ASP_2)[8] <- "avg.total.rural.pop"
colnames(ASP_2)[9] <- "avg.rural.percent.total"
```

```
summary(ASP_2)
```

```
##   electricity        water         sanitation     natl.pg.per
##  Min.   :44.01   Min.   :59.84   Min.   :29.24   Min.   : 1.075
##  1st Qu.:49.49   1st Qu.:66.26   1st Qu.:37.35   1st Qu.: 3.565
##  Median :59.67   Median :73.44   Median :44.95   Median : 7.697
##  Mean   :58.65   Mean   :73.45   Mean   :44.86   Mean   : 8.379
##  3rd Qu.:68.83   3rd Qu.:80.61   3rd Qu.:52.41   3rd Qu.:11.343
##  Max.   :71.23   Max.   :87.26   Max.   :58.95   Max.   :20.610
##  NA's   :27      NA's   :5       NA's   :5       NA's   :6
##   rural.pg.per       women.ag     avg.total.pop       avg.total.rural.pop
##  Min.   : 2.200   Min.   :27.02   Min.   :231935184   Min.   :176808593
##  1st Qu.: 3.600   1st Qu.:39.80   1st Qu.:265487650   1st Qu.:191573042
##  Median : 4.600   Median :47.03   Median :296081155   Median :196277957
##  Mean   : 5.242   Mean   :45.86   Mean   :293246083   Mean   :194131886
##  3rd Qu.: 6.400   3rd Qu.:52.54   3rd Qu.:322249256   3rd Qu.:199363821
##  Max.   :13.700   Max.   :64.06   Max.   :345881200   Max.   :200684817
##  NA's   :18       NA's   :2
##  avg.rural.percent.total
##  Min.   :62.18
##  1st Qu.:66.64
##  Median :70.82
##  Mean   :70.31
##  3rd Qu.:74.05
##  Max.   :77.31
##
```

The last re-formatting I need to do is create a data frame out of the ASP_2 data set.

```
ASP.df <- data.frame(ASP_2)
glimpse(ASP.df)
```

```
## Observations: 31
## Variables: 9
## $ electricity             (dbl) NA, NA, NA, NA, NA, 44.01299, NA, NA, ...
## $ water                   (dbl) NA, NA, NA, NA, NA, 59.84, 60.67, 61.4...
## $ sanitation              (dbl) NA, NA, NA, NA, NA, 29.24444, 31.84000...
## $ natl.pg.per             (dbl) 14.240000, NA, 20.610000, 11.343333, N...
## $ rural.pg.per            (dbl) NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ women.ag                (dbl) 32.63333, 42.60000, 51.06667, 52.10000...
## $ avg.total.pop           (dbl) 231935184, 236375699, 240989349, 24564...
## $ avg.total.rural.pop     (dbl) 176808593, 178987700, 181233468, 18344...
## $ avg.rural.percent.total (dbl) 77.3098, 76.8750, 76.4319, 75.9797, 75...
```

```
head(ASP.df)
```

```
##      electricity water sanitation natl.pg.per rural.pg.per women.ag
## 1985          NA    NA         NA    14.24000           NA 32.63333
## 1986          NA    NA         NA          NA           NA 42.60000
## 1987          NA    NA         NA    20.61000           NA 51.06667
## 1988          NA    NA         NA    11.34333           NA 52.10000
## 1989          NA    NA         NA          NA           NA 56.77500
## 1990    44.01299 59.84   29.24444    14.45333           NA 50.70000
##      avg.total.pop avg.total.rural.pop avg.rural.percent.total
## 1985     231935184           176808593                 77.3098
## 1986     236375699           178987700                 76.8750
## 1987     240989349           181233468                 76.4319
## 1988     245643934           183440733                 75.9797
## 1989     250238654           185533194                 75.5133
## 1990     254751145           187494431                 75.0377
```

Correlation Test

Water

```
cor(ASP.df$rural.pg.per, ASP.df$water, use = "pairwise.complete.obs")
```

```
## [1] -0.735621
```

```
mean.water <- mean(ASP.df$water, na.rm = T)
mean.water
```

```
## [1] 73.45385
```

Electricity

```
cor(ASP.df$rural.pg.per, ASP.df$electricity, use = "pairwise.complete.obs")
```

```
## [1] -0.999437
```

```
mean.electric <- mean(ASP.df$electricity, na.rm = T)
mean.electric
```

```
## [1] 58.64722
```

Sanitation

```
cor(ASP.df$rural.pg.per, ASP.df$sanitation, use = "pairwise.complete.obs")
```

```
## [1] -0.7393785
```

```
mean.sanitation <- mean(ASP.df$sanitation, na.rm = T)
mean.sanitation
```

```
## [1] 44.86479
```

Women in Agriculture

```
cor(ASP.df$rural.pg.per, ASP.df$women.ag, use = "pairwise.complete.obs")
```

```
## [1] 0.5275307
```

```
mean.women <- mean(ASP.df$women.ag, na.rm = T)
mean.women
```

```
## [1] 45.861
```

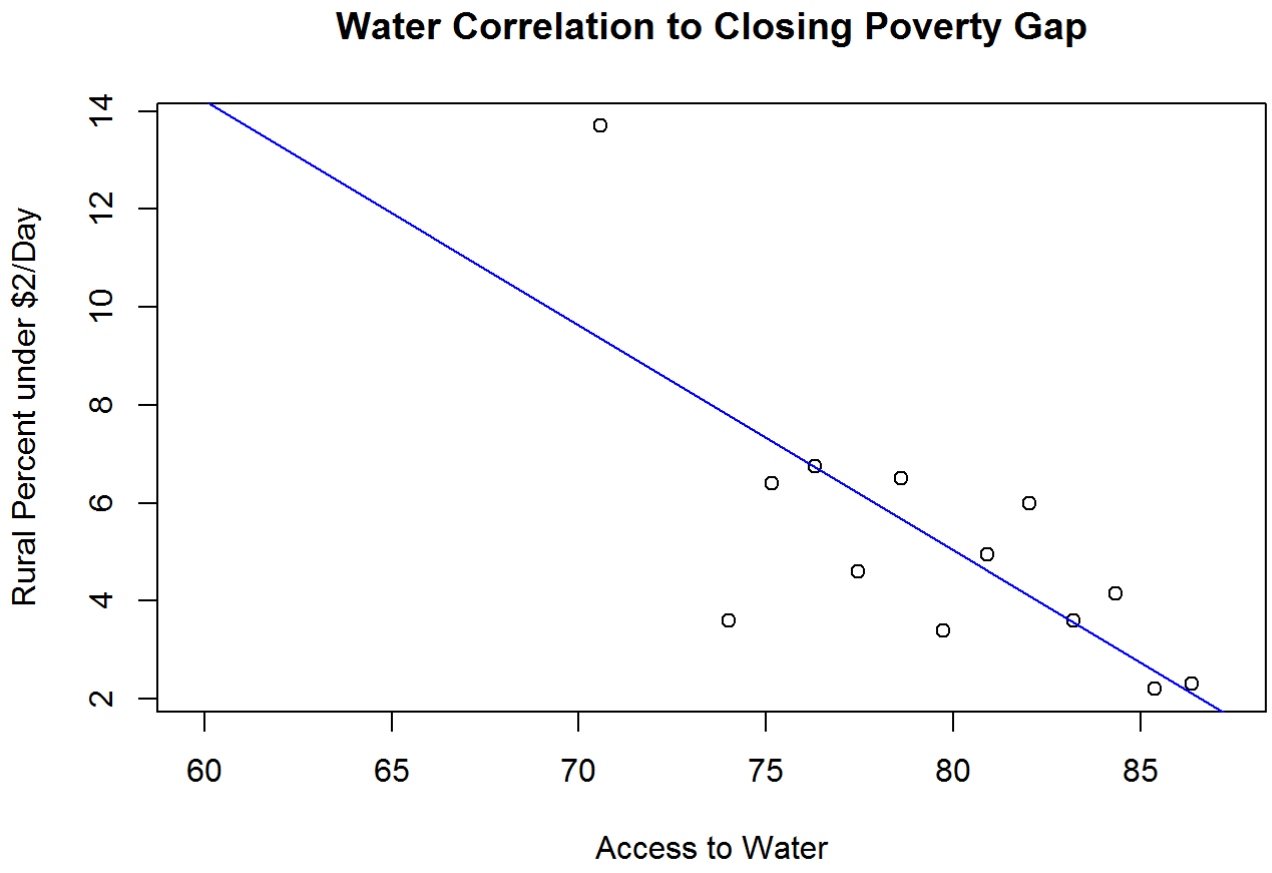
Simple Linear Regression

Water

```
model1 <- lm(rural.pg.per ~ water, data = ASP.df)
model1
```

```
##
## Call:
## lm(formula = rural.pg.per ~ water, data = ASP.df)
##
## Coefficients:
## (Intercept)        water
##     41.7457      -0.4589
```

```
water.cor.plot <- plot(ASP.df$water, ASP.df$rural.pg.per, main = "Water Correlation to Closing P
overty Gap", xlab = "Access to Water", ylab = "Rural Percent under $2/Day")
abline(h = mean.water)
abline(model1, col = "blue")
```




```
summary(model1)
```

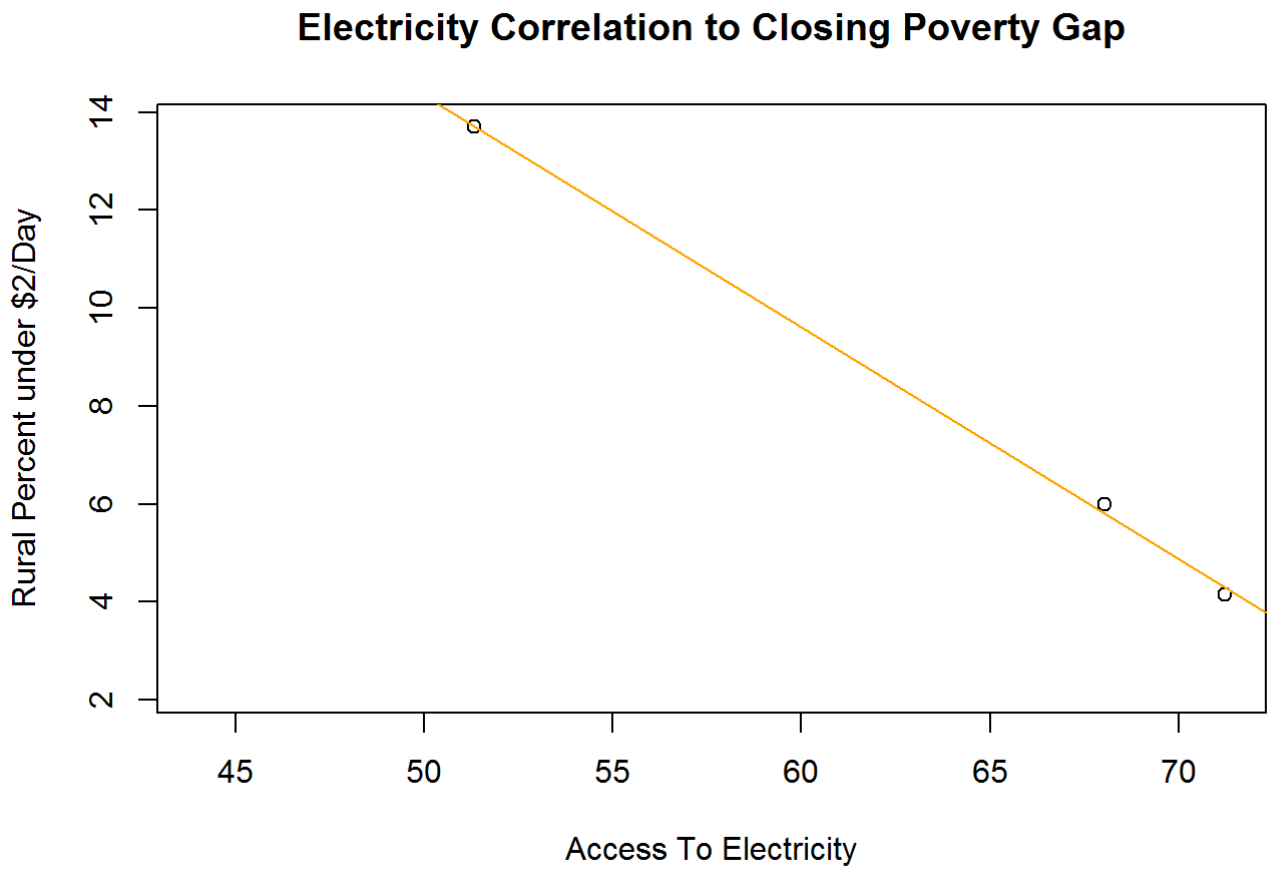
```
##
## Call:
## lm(formula = rural.pg.per ~ water, data = ASP.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1802 -0.8525  0.0414  0.8260  4.3367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.7457    10.1517   4.112  0.00172 **
## water        -0.4589     0.1274  -3.602  0.00416 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.1 on 11 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.5411, Adjusted R-squared:  0.4994
## F-statistic: 12.97 on 1 and 11 DF,  p-value: 0.004157
```

Electricity

```
model2 <- lm(rural.pg.per ~ electricity, data = ASP.df)
model2
```

```
##
## Call:
## lm(formula = rural.pg.per ~ electricity, data = ASP.df)
##
## Coefficients:
## (Intercept)  electricity
##     38.0251      -0.4734
```

```
electric.cor.plot <- plot(ASP.df$electricity, ASP.df$rural.pg.per, main = "Electricity Correlati
on to Closing Poverty Gap", xlab = "Access To Electricity", ylab = "Rural Percent under $2/Day")

abline(h = mean.electric)
abline(model2, col = "orange")
```




```
summary(model2)
```

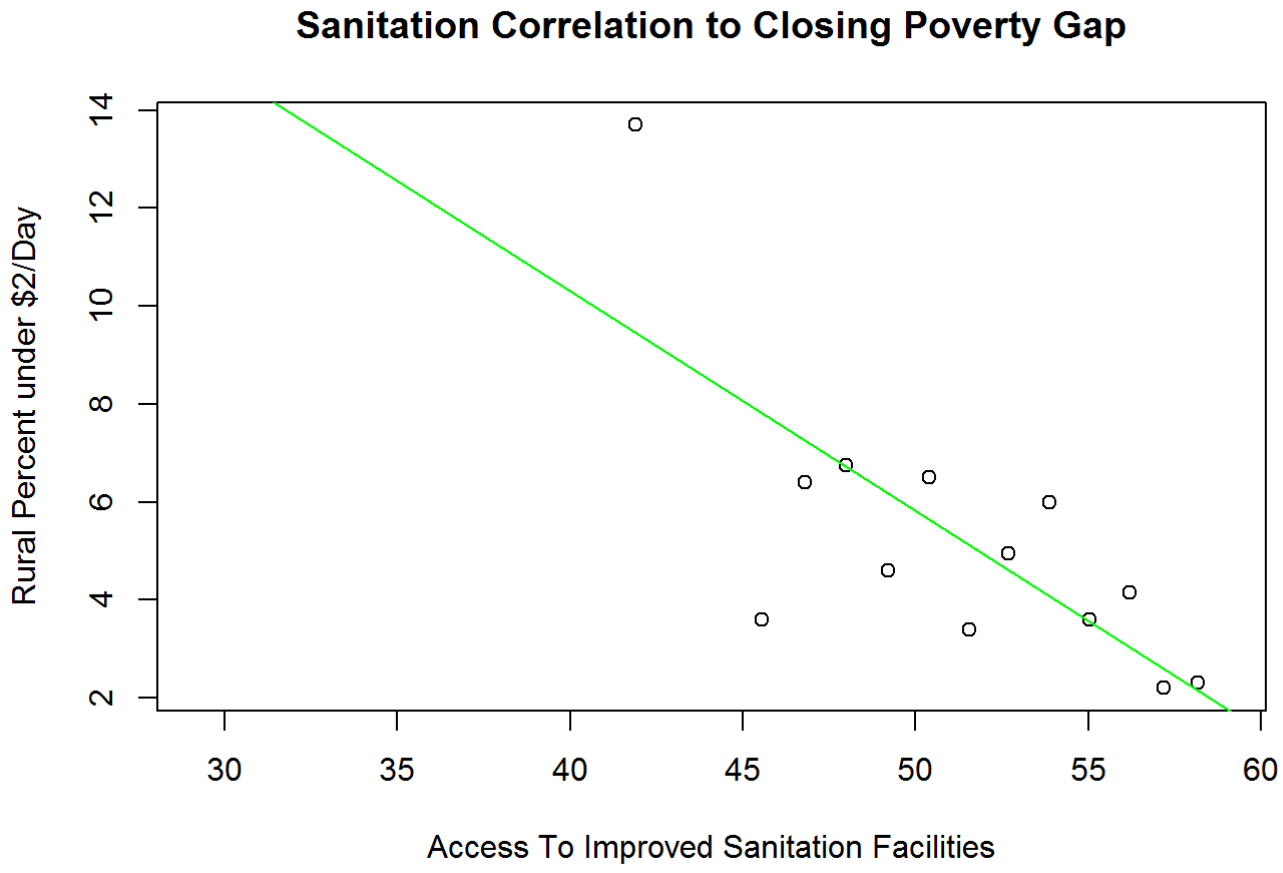
```
##
## Call:
## lm(formula = rural.pg.per ~ electricity, data = ASP.df)
##
## Residuals:
##     2000     2010     2012
## -0.02934  0.18269 -0.15334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.02515    1.01908   37.31   0.0171 *
## electricity -0.47344    0.01589  -29.79   0.0214 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2403 on 1 degrees of freedom
##   (28 observations deleted due to missingness)
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9977
## F-statistic: 887.4 on 1 and 1 DF,  p-value: 0.02136
```

Sanitation

```
model3 <- lm(rural.pg.per ~ sanitation, data = ASP.df)
model3
```

```
##
## Call:
## lm(formula = rural.pg.per ~ sanitation, data = ASP.df)
##
## Coefficients:
## (Intercept)   sanitation
##     28.2971      -0.4497
```

```
sanitation.cor.plot <- plot(ASP.df$sanitation, ASP.df$rural.pg.per, main = "Sanitation Correlati
on to Closing Poverty Gap", xlab = "Access To Improved Sanitation Facilities", ylab = "Rural Per
cent under $2/Day")
abline(h = mean.sanitation)
abline(model3, col = "green")
```




```
summary(model3)
```

```
##
## Call:
## lm(formula = rural.pg.per ~ sanitation, data = ASP.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2151 -0.8530  0.0477  0.8658  4.2437
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.2971     6.3563   4.452 0.000976 ***
## sanitation   -0.4497     0.1235  -3.642 0.003873 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.087 on 11 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.5467, Adjusted R-squared:  0.5055
## F-statistic: 13.27 on 1 and 11 DF,  p-value: 0.003873
```

Women's Employment in Agriculture

```
model4 <- lm(rural.pg.per ~ women.ag, data = ASP.df)
model4
```

```
##
## Call:
## lm(formula = rural.pg.per ~ women.ag, data = ASP.df)
##
## Coefficients:
## (Intercept)     women.ag
##     -1.2136       0.1681
```

```
women.cor.plot <- plot(ASP.df$women.ag, ASP.df$rural.pg.per, main = "Women's Empowerment Correla
tion to Closing Poverty Gap", xlab = "Female Employment in Agr.", ylab = "Rural Percent under
 $2/Day")
abline(h = mean.women)
abline(model4, col = "purple")
```


```
summary(model4)
```

```
##
## Call:
## lm(formula = rural.pg.per ~ women.ag, data = ASP.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2197 -1.7427  0.0185  0.9927  5.0470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.21358    3.49639  -0.347    0.736
## women.ag     0.16813    0.08562   1.964    0.078 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.637 on 10 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.2783, Adjusted R-squared:  0.2061
## F-statistic: 3.856 on 1 and 10 DF,  p-value: 0.07796
```

Multiple Linear Regression

```
modelA <- lm(rural.pg.per ~ water * sanitation, data = ASP.df)
summary(modelA)
```

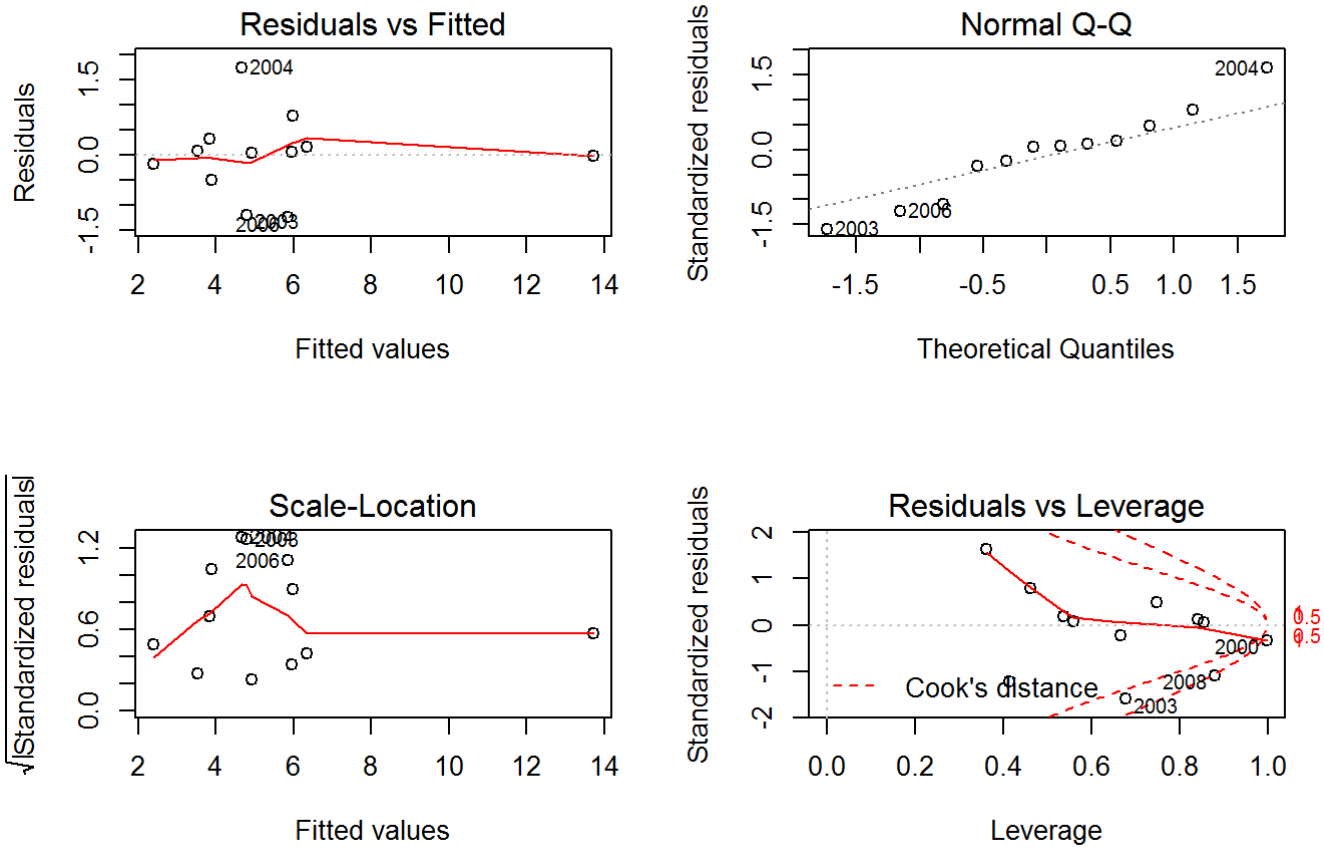
```
##
## Call:
## lm(formula = rural.pg.per ~ water * sanitation, data = ASP.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0990 -1.0381 -0.3381  1.1703  2.5301
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1551.07404 1065.84269   1.455    0.180
## water             -39.23395   28.85963  -1.359    0.207
## sanitation         18.99268   17.37883   1.093    0.303
## water:sanitation    0.14670    0.08787   1.669    0.129
##
## Residual standard error: 1.915 on 9 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.6879, Adjusted R-squared:  0.5838
## F-statistic: 6.612 on 3 and 9 DF,  p-value: 0.01183
```

```
modelB <- lm(rural.pg.per ~ water * sanitation * women.ag, data = ASP.df)
summary(modelB)
```

```
##
## Call:
## lm(formula = rural.pg.per ~ water * sanitation * women.ag, data = ASP.df)
##
## Residuals:
##     2000     2003     2004     2005     2006     2007     2008     2009
## -0.02013 -1.19852  1.73479  0.77394 -1.24745  0.15892 -0.49649  0.02667
##     2010     2011     2012     2013
##  0.06146  0.06634  0.32131 -0.18083
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.772e+03  8.570e+03   0.790    0.474
## water                      -2.056e+02  2.292e+02  -0.897    0.420
## sanitation                  1.630e+02  1.300e+02   1.254    0.278
## women.ag                   -1.870e+02  2.124e+02  -0.880    0.428
## water:sanitation            2.982e-01  7.550e-01   0.395    0.713
## water:women.ag              5.541e+00  5.670e+00   0.977    0.384
## sanitation:women.ag        -4.162e+00  3.198e+00  -1.301    0.263
## water:sanitation:women.ag -9.826e-03  1.880e-02  -0.523    0.629
##
## Residual standard error: 1.325 on 4 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.7996
## F-statistic: 7.271 on 7 and 4 DF,  p-value: 0.03694
```

```
par(mfrow=c(2,2))
plot(modelB)
```

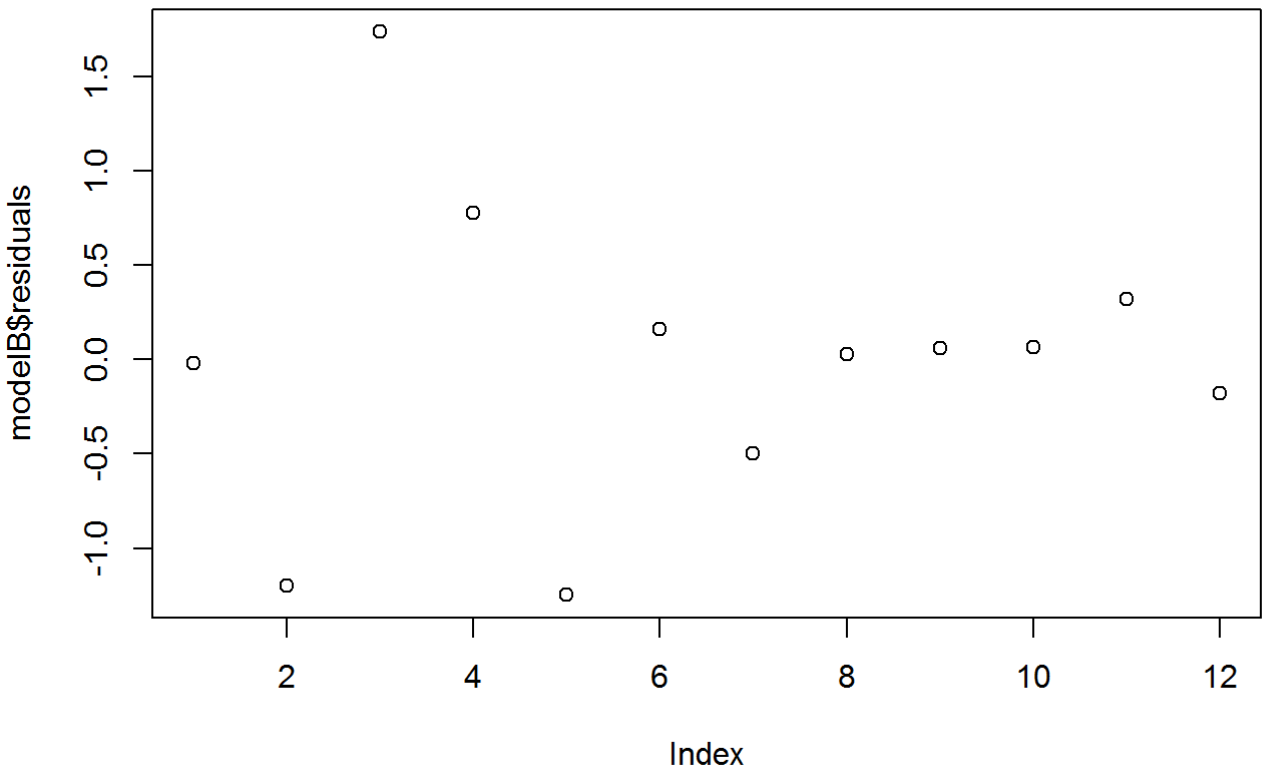
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Residuals

The residual plot shows us a pattern indicating that the model may not be perfectly fitted to the true relationships between the World Bank Indicators and the Rural Poverty Gap. Model B:

```
plot(modelB$residuals)
```




```
SSE = sum(modelB$residuals^2)
SSE
```

```
## [1] 7.018062
```

```
RMSE = sqrt(SSE/nrow(ASP.df))
RMSE
```

```
## [1] 0.4758036
```

Predicting Outcomes

```
summary(ASP.df$rural.pg.per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.200   3.600   4.600   5.242   6.400  13.700      18
```

```
povertygap1 <- mean(ASP.df$rural.pg.per, na.rm = T)
povertygap2 <- 4.0
povertygap3 <- 6.0
predict(modelB)
```

```
##      2000      2003      2004      2005      2006      2007      2008
## 13.720135  4.798525  4.665208  5.976064  5.847451  6.341076  3.896486
##      2009      2010      2011      2012      2013
##  4.923330  5.938541  3.533662  3.828694  2.380828
```

Testing the Predictions

Now, I need a new data frame with the same predictors as the original model. Back to square one — but this time, with Africa data.

```
WB_data <- wb(country = c("CM", "GH", "KE", "MW", "RW", "SN", "TZ", "UG", "ZM", "ZW"), indicator
 = c("EG.ELC.ACCS.RU.ZS", "SH.H2O.SAFE.RU.ZS", "SH.STA.ACSN.RU", "SL.AGR.EMPL.FE.ZS", "SI.POV.RU
GP", "SI.POV.GAPS", "SP.RUR.TOTL", "SP.RUR.TOTL.ZS",  "SP.POP.TOTL"), startdate = 1985, enddate
= 2015)
WB_data$date <- as.numeric(WB_data$date)

AFR_data <- WB_data %>%
select(date, iso2c, country, indicatorID, indicator, value) %>%
arrange(date, indicator, value)

AFR_2 <- tapply(AFR_data$value, list(date = AFR_data$date, ticker = AFR_data$indicatorID), mean)

colnames(AFR_2)[1] <- "electricity"
colnames(AFR_2)[2] <- "water"
colnames(AFR_2)[3] <- "sanitation"
colnames(AFR_2)[4] <- "natl.pg.per"
colnames(AFR_2)[5] <- "rural.pg.per"
colnames(AFR_2)[6] <- "women.ag"
colnames(AFR_2)[7] <- "avg.total.pop"
colnames(AFR_2)[8] <- "avg.total.rural.pop"
colnames(AFR_2)[9] <- "avg.rural.percent.total"

AFR.df <- data.frame(AFR_2)
glimpse(AFR.df)
```

```
## Observations: 31
## Variables: 9
## $ electricity             (dbl) NA, NA, NA, NA, NA, 2.82000, NA, NA, N...
## $ water                   (dbl) NA, NA, NA, NA, NA, 41.57, 42.56, 43.5...
## $ sanitation              (dbl) NA, NA, NA, NA, NA, 21.63, 21.94, 22.2...
## $ natl.pg.per             (dbl) NA, NA, 25.13000, 24.61000, 53.13000, ...
## $ rural.pg.per            (dbl) NA, NA, NA, NA, NA, NA, NA, 22.60, NA,...
## $ women.ag                (dbl) NA, NA, NA, NA, 95.90000, 56.00000, 90...
## $ avg.total.pop           (dbl) 11487274, 11895699, 12322706, 12755903...
## $ avg.total.rural.pop     (dbl) 9027163, 9305408, 9597113, 9892255, 10...
## $ avg.rural.percent.total (dbl) 77.1036, 76.7206, 76.3403, 75.9766, 75...
```

```
modelC <- lm(rural.pg.per ~ water * sanitation * women.ag, data = AFR.df)
summary(modelC)
```

```
##
## Call:
## lm(formula = rural.pg.per ~ water * sanitation * women.ag, data = AFR.df)
##
## Residuals:
##    1992    1999    2000    2001    2002    2004    2005    2009    2010
## -0.4868  1.6926 -1.3554 -0.6773  1.2180  1.1944 -0.1527 -1.6153 -1.6178
##    2011    2012    2014
##  4.8588 -6.2380  3.1797
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                6.250e+04  2.933e+04   2.131   0.1001
## water                      2.472e+03  1.121e+03   2.205   0.0921 .
## sanitation                -7.544e+03  3.467e+03  -2.176   0.0952 .
## women.ag                  -8.561e+02  4.167e+02  -2.054   0.1092
## water:sanitation          -2.553e+00  2.529e+00  -1.009   0.3699
## water:women.ag            -3.448e+01  1.572e+01  -2.194   0.0933 .
## sanitation:women.ag        1.043e+02  4.890e+01   2.132   0.1000 .
## water:sanitation:women.ag  4.225e-02  4.028e-02   1.049   0.3535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.642 on 4 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.7988, Adjusted R-squared:  0.4467
## F-statistic: 2.269 on 7 and 4 DF,  p-value: 0.2236
```

```
predict(modelC)
```

```
##      1992      1999      2000      2001      2002      2004      2005
## 23.086772  9.507447 22.555426 17.977347 11.882018 18.005623 15.852702
##      2009      2010      2011      2012      2014
##  9.215345 27.584455 20.441246 15.738007 19.720279
```

```
summary(AFR.df$rural.pg.per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    7.60   14.40   19.20   18.75   23.40   28.25      16
```

```
AFR.povertygap1 <- mean(AFR.df$rural.pg.per, na.rm = T)
AFR.povertygap2 <- 14
AFR.povertygap3 <- 23
```

Against Avg. Levels in Asia

```
povertygap.test <- predict(modelB)-predict(modelC)
SSE = sum((povertygap.test)^2)
RMSE = sqrt(SSE/nrow(AFR.df))
RMSE
```

```
## [1] 8.23985
```

```
SST = sum((povertygap1 - AFR.povertygap1)^2)
R2 = 1 - SSE/SST
R2
```

```
## [1] -10.53934
```

Against Low Levels in Asia

```
povertygap.test <- predict(modelB)-predict(modelC)
SSE = sum((povertygap.test)^2)
SST = sum((povertygap2 - AFR.povertygap2)^2)
R2 = 1 - SSE/SST
R2
```

```
## [1] -20.04749
```

Against High Levels in Asia

```
povertygap.test <- predict(modelB)-predict(modelC)
SSE = sum((povertygap.test)^2)
SST = sum((povertygap3 - AFR.povertygap3)^2)
R2 = 1 - SSE/SST
R2
```

```
## [1] -6.282868
```

The indicators performed best against other populations with higher levels of the rural population living under $2/day, but Model B would not be predictive to any of the World Bank Indicators when applied to data from Africa.