Springboard Final Capstone Report
Reducing Poverty in Asia by Kathryn Gadberry
June 2016

Problem

I currently work in an analyst role at Heifer Project International, one of the most capable and longstanding community development organizations in the industry. Our mission is to end hunger and poverty while caring for the earth. One of the greatest weaknesses of the non-profit industry today is the historical lack of measurement and management around implementing effective change. Utilizing methodology and software (R programming language specifically) in this new age of big data and technology has enabled Non-Governmental Organizations to statistically measure sustainability techniques in a quick and reproducible fashion.

The client I want my analysis to create value for is either an international development organization designing programs or a donor interested in investing in an organization to reduce poverty. I will explain how to measure which community development indicators would be the most predictive to ensure sustainable impact and mission success after project completion.

The problem this example will explore is how to measure and predict success of key sustainability indicators like access to clean water, access to electricity, access to improved sanitation facilities and women's empowerment/participation in the labor force in the Asia Pacific region. Rural farmer's basic amenities are very limited and women have become the backbone of many smallholder farms. About 43% of the world's agricultural labor force (up to 70% in some countries) is comprised of female workers. Unfortunately, these women are also limited by ownership inequality and decision making power. Asia's significance in any world demographic analysis is clear to see just by it's sheer size. Almost two-thirds of the world's population lives in Asia; that's 4 billion people, covering 17% of the earth's surface. Significantly reducing poverty in a large population is crucial to achieving the United Nation's primary Millennium Development Goal of halving poverty by 2015, which trickles down to many similar incidental goals made by other NGOs as well.

Lack of infrastructure, geographic barriers and educational, as well as social inequalities are just a few of the many reasons why poverty has become a "rural problem". It is estimated that between 80-90% of people living below the poverty line (less than $2 income/day) are residing within rural regions in a majority of these emerging states. A developing country is defined as a nation with a poor agricultural society that is seeking to become more socially and economically advanced. Ten of the most populous developing countries in the Asian Pacific are Bangladesh, Cambodia, China, India, Indonesia, Myanmar, Nepal, the Philippines, Thailand, and Vietnam. It is these ten countries I will base my analysis off of. Which sustainability measurements (water, electricity, sanitation, and women's empowerment) have been the most successful, and will continue to be, in reducing rural poverty in Asia?

Exploratory Data Analysis

Approach

I would have like to have used an annual project outcomes report that Heifer sends to donors, detailing how adoption of our sustainable living techniques significantly changes a family's life by allowing them basic amenities and means, like income or training, to lift themselves out of poverty. This annual report turned out to be unsuitable for my purposes since the data recorded in this type of review has only been in existence for a fraction of the time as another similar data set I found online at the World Bank. I did not anticipate how difficult it would be to find a complete data set on this topic.The World Bank tracks several identical key indicators that Heifer Project International uses in our Mission Effectiveness reporting. The indicators my analysis will look at include:

- **Access to an improved water source (% or the rural population)**
- **Access to electricity (% or the rural population)**
- **Access to improved sanitation facilities (% or the rural population)**
- **% Females employed in Agriculture**

These four independent variables will be studied against the dependent variable:

- **Rural Poverty Gap (at national poverty lines (% under 1.90/day)**

In later analysis, we would be able to grade how these sustainability factors impact an impoverished region over time. My data sources include the World Bank site (http://data.worldbank.org/country), the 'WDI' - R - Cran - Package printout (https://cran.r-project.org/web/packages/WDI/WDI.pdf), and the Rural Poverty Portal (http://www.ruralpovertyportal.org/region/home/tags/asia).

This data set proved to be thorough and meaningful for this subject, yet still easy enough to acquire and clean for analysis.

The approach I took for this analysis began with installing necessary packages and loading the appropriate libraries to operate within my Rstudio workspace. I filtered out the countries, indicators, and timeframe from the World Bank data and assigned it to the object: WB_data. Preliminary research of the raw data revealed that complete data had only been collected for all the variables since 1985, about 30 years ago. I inspected the data frame using the glimpse() and str() functions from the dplyr package. I had to factor out the "Date" data and use as.numeric() for WB_data$date in order to change the value from a character to an integer. I continued to use dplyr functions to select and arrange the data in a more logical order and assigned this new format to the object "ASP_data". After inspecting the ASP_data using the head() function, I was ready to begin the exploratory data analysis.

My exploratory investigation began by separately looking at each World Bank indicator. I created an object for each sustainability indicator (ASP_water, ASP_electricity, ASP_sanitation and ASP_women), then set those objects equal to the WB_data for that indicator ID, and repeated the as.numeric() step for each data sets “date” variable. I found the best visual method for evaluating each indicator was using box and whisker plots. This plot type shows the minimum, 1st quartile average, set average, 3rd quartile average, and maximum for each measure of sustainability over a 30 year period.

Findings

Box and whisker plots turned out to be very telling of the World Bank data for each variable. I found that the percentage of the rural population with access to water had steadily increased from an average of about 30% or 1% a year. The narrowing of the range of the data was the most dramatic finding. Some rural populations had up to 80% without access to improved sources in 1990, which decreased to about 30% in 2015 (see Figure 1.1).

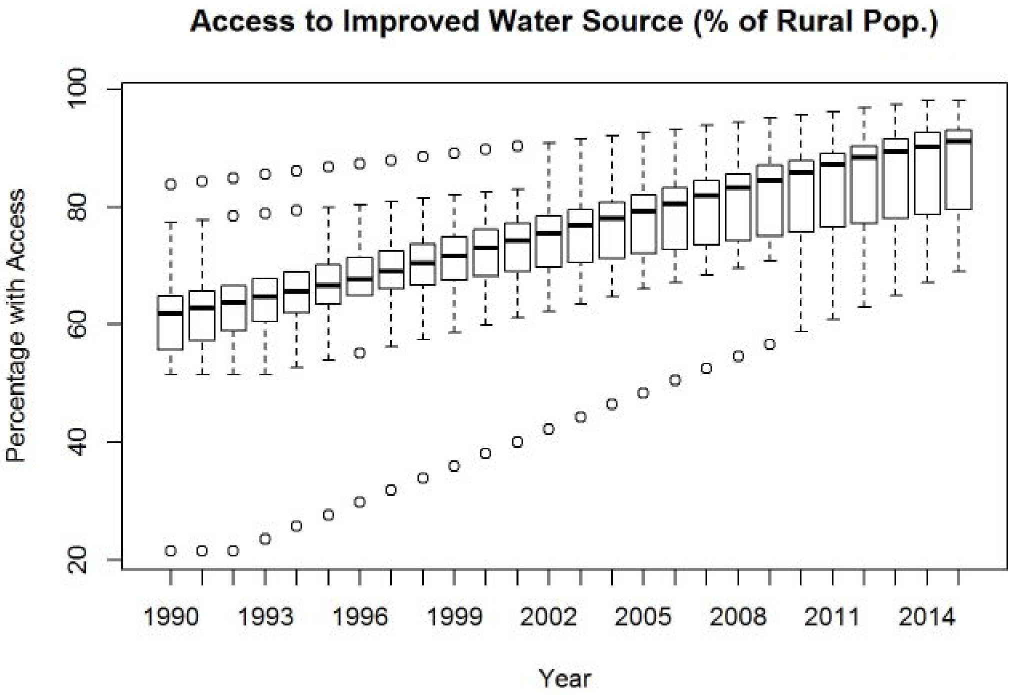


Figure 1.2

The access to Electricity variable had much less data available comparative to water, and therefore only had 4 box and whisker plots for 1990, 2000, 2010, and 2012. I was still able to recognize an upward trend from a little over 40% of the rural population with electricity to almost 80% on average over this time period (Figure 1.2).

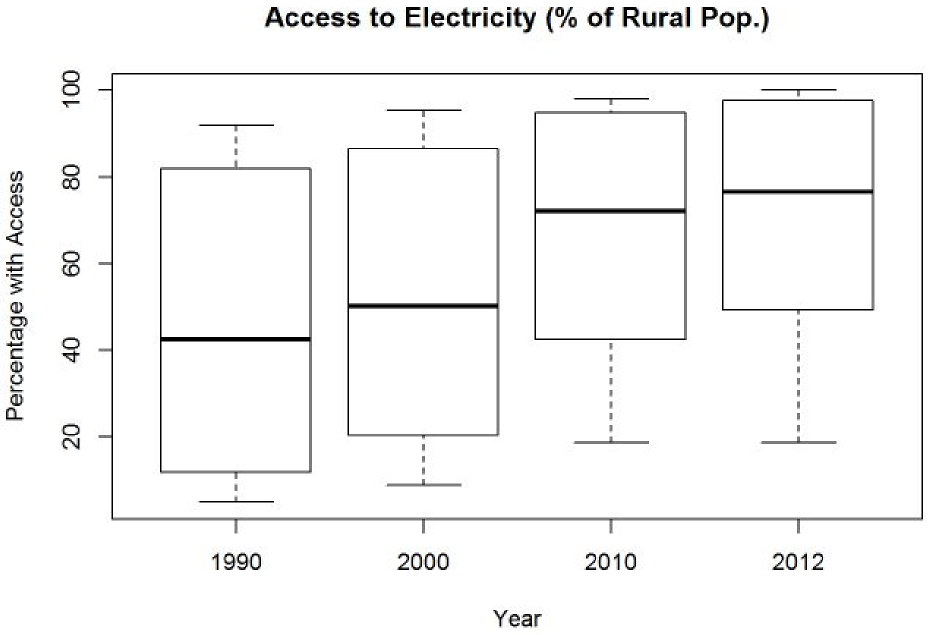


Figure 1.2

I performed the same analysis for sanitation, which similarly increased on average along with a narrowing range (Figure 1.3).

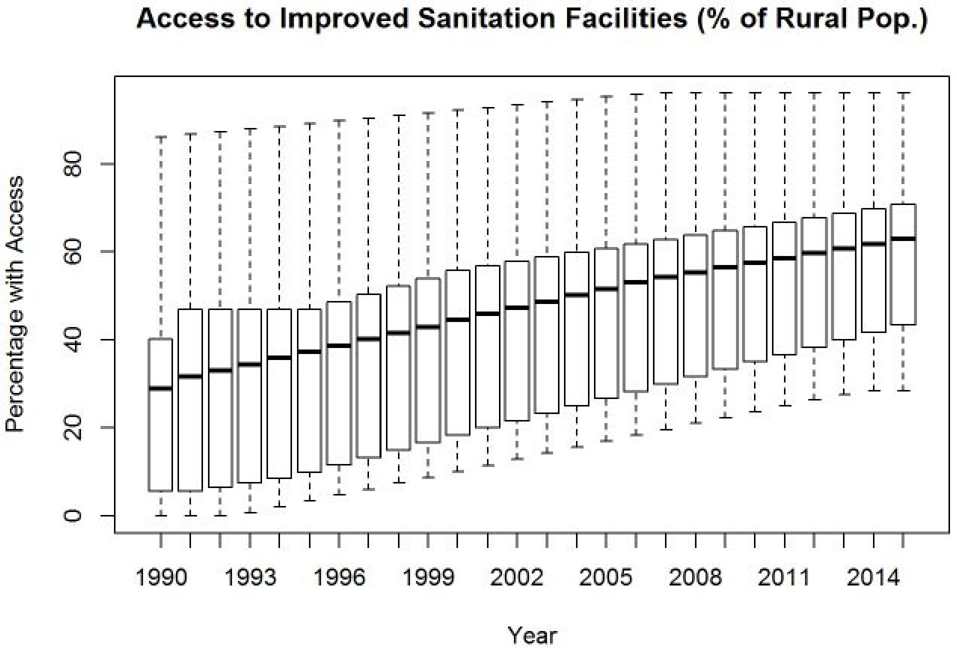


Figure 1.3

The boxplot analysis on the women in agriculture was the most unique by far. Female participation began to gain traction in the 80s, moving upward, then dipping the the late 90s, peaking in the early 2000s, and then coming to an all time average low in 2007. Since then, it has seemed to stabilize, but still only about 40-45% of women are participating in the agricultural labor force of these developing countries (Figure 1.4).

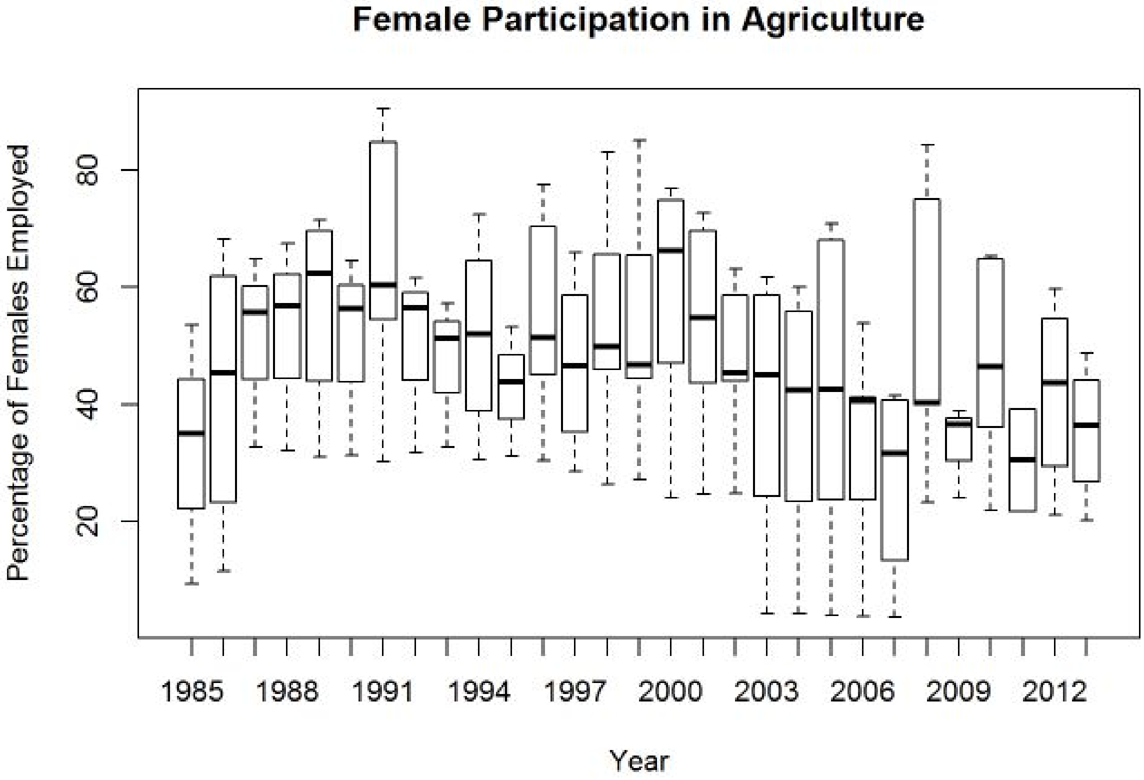


Figure 1.4

Testing for Linear Relationships

<u>Approach</u>

Before I could begin any predictive analysis, I needed to wrangle and clean the data for hypothesis testing. The current structure of the data frame listed each indicator under one variable [indicator/indicatorID] and all values for those indicators as one variable [value]. I separated out each indicator and created a column for it containing the values for each year's observation.

For example: I moved the indicator observation, EG.ELC.ACCS.RU.ZS ("access to electricity"), to a separate column with the values listed below for the average 1990 observation. I created the columns "Year | Electricity (indicator1) | Water (2) | Sanitation (3)" with values listed below each indicator relating to average population count or percentage that had access to that resource during 1990. I accomplished this by using the function call: tapply(ASP_data$value, list(date = ASP_data$date, ticker = ASP_data$indicatorID), mean) and assigning it to the object ASP_2. Next, I renamed the columns from the difficult to interpret indicator IDs to simple variable names using the colnames(ASP-2)[ ] function. Finally, after I restructured and renamed variables in ASP_2, I reformatted the whole thing into a data frame and set it equal to ASP.df.

I was now ready to begin correlation testing between the variables. After removing the missing NA values, I use the cor() function for each independent variable against the dependent, rural percentage living under $2 day. I ran a correlation test for each indicator, and at the same time, calculated the mean in order to compare the two.

Findings

The correlation test came back negative (0 = no correlation, -1 = highest correlation), because we wanted an inverse relationship between a high percentage of access to resources and employment vs. a low percentage of the population living in poverty to show that the sustainability indicator had an impact towards reducing poverty. The results for each variable helped me formulate a hypothesis: that access to clean water was highly correlated and more prevalent on average for the rural population in Asia that was rose out of poverty since 1985 (Figure 2.1).

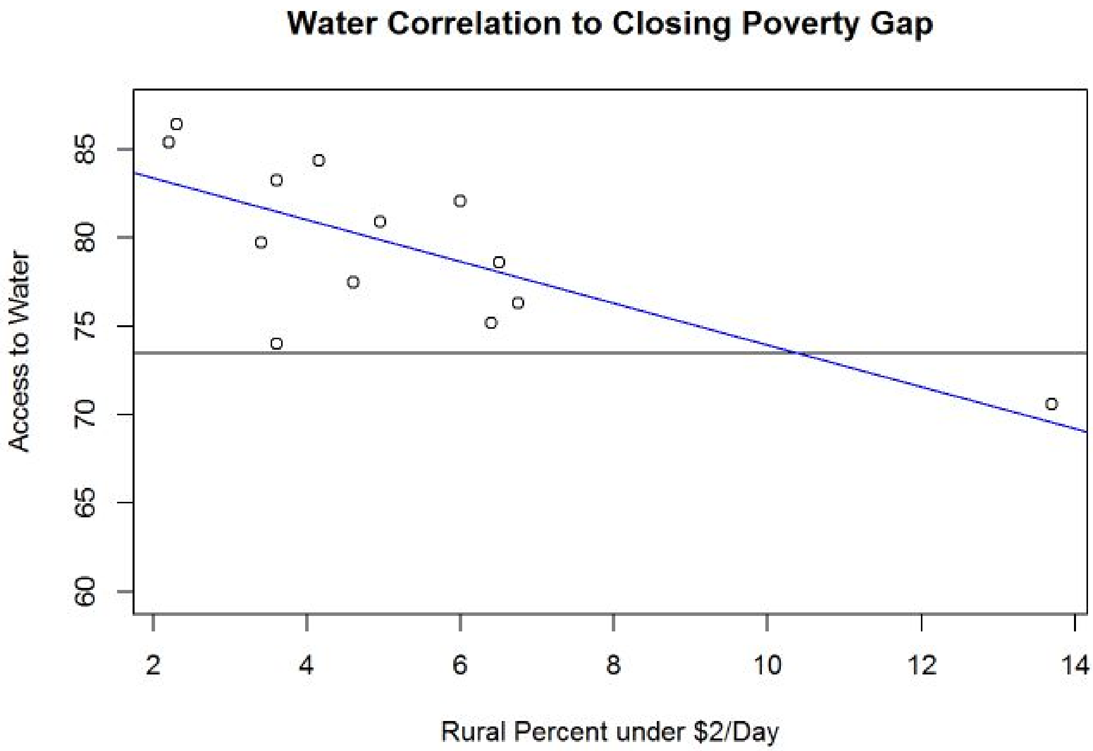


Figure 2.1

Electricity had the highest correlation, but fewer data points to draw from. This is still a significant factor, but not as defined as it may seem initially (Figure 2.2).

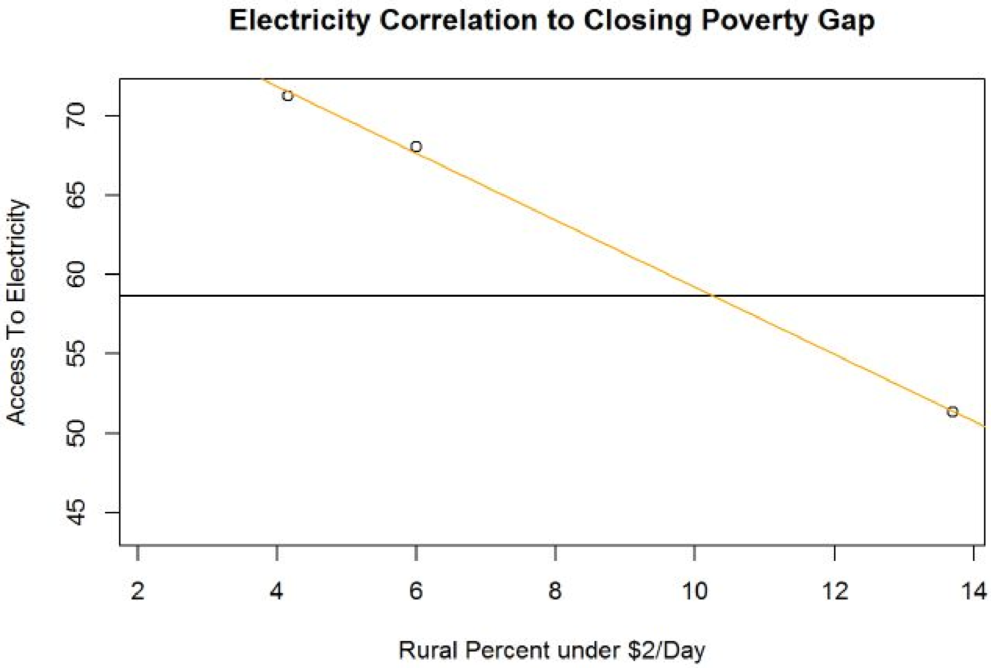


Figure 2.2

It also seemed like, on average, there was a similar amount of women employed in agriculture as there was access but improved sanitation facilities, but the sanitation variable was much more significantly correlated with reducing rural percentage of the poverty gap (Figure 2.3).

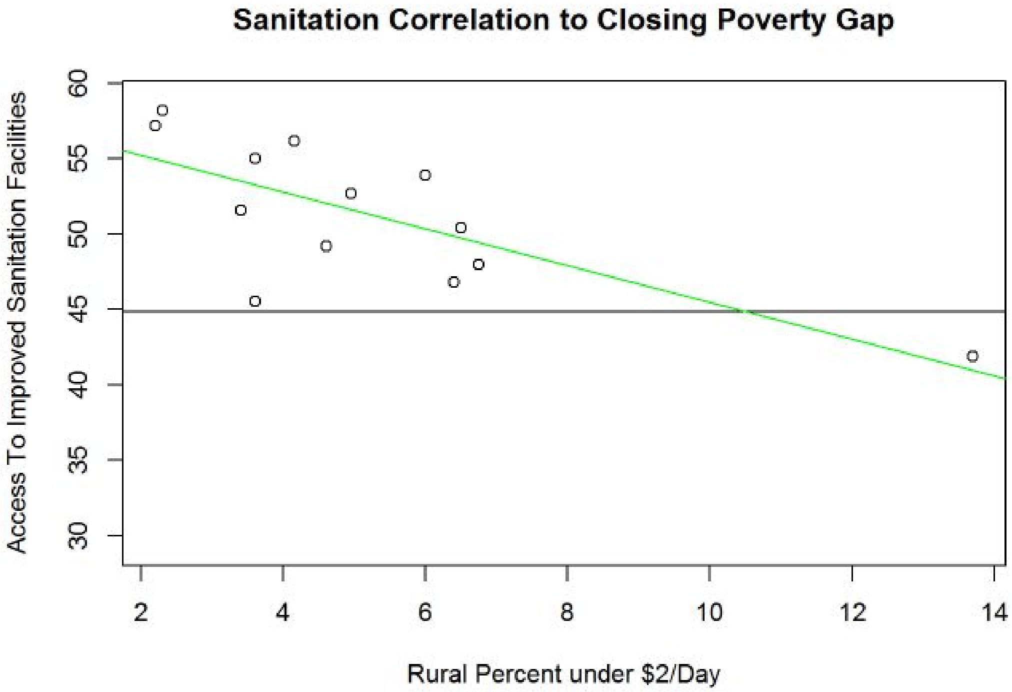


Figure 2.3

Surprisingly, women employed in agriculture leaned towards a positive correlation. This variable ranks as the least significant out of the four in reducing poverty, but could yield an interesting result in the analysis of its effect on the dependent variable (Figure 2.4).

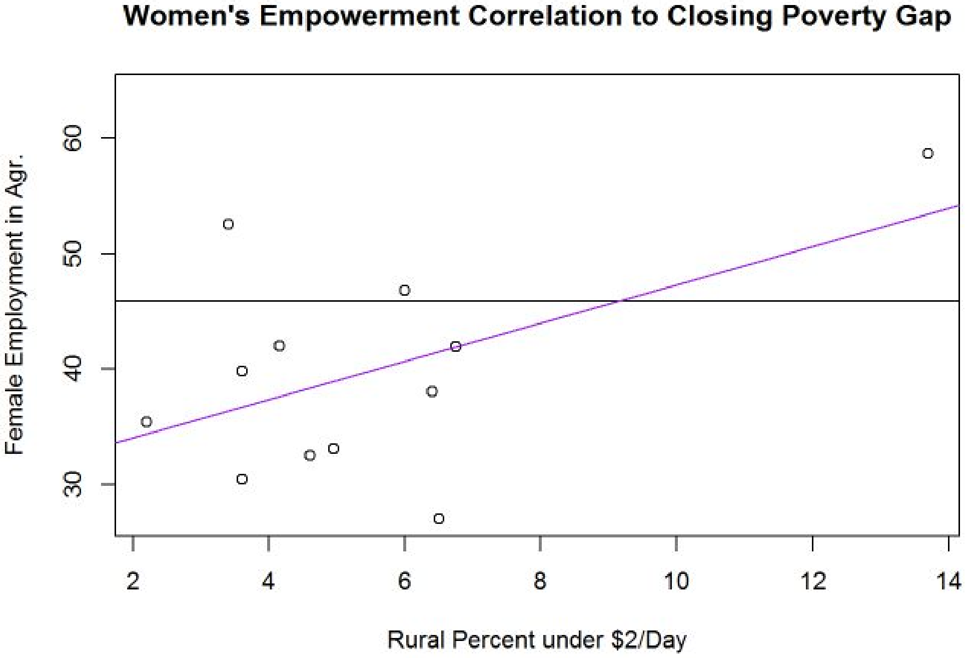


Figure 2.4

Simple Linear Regression

Approach

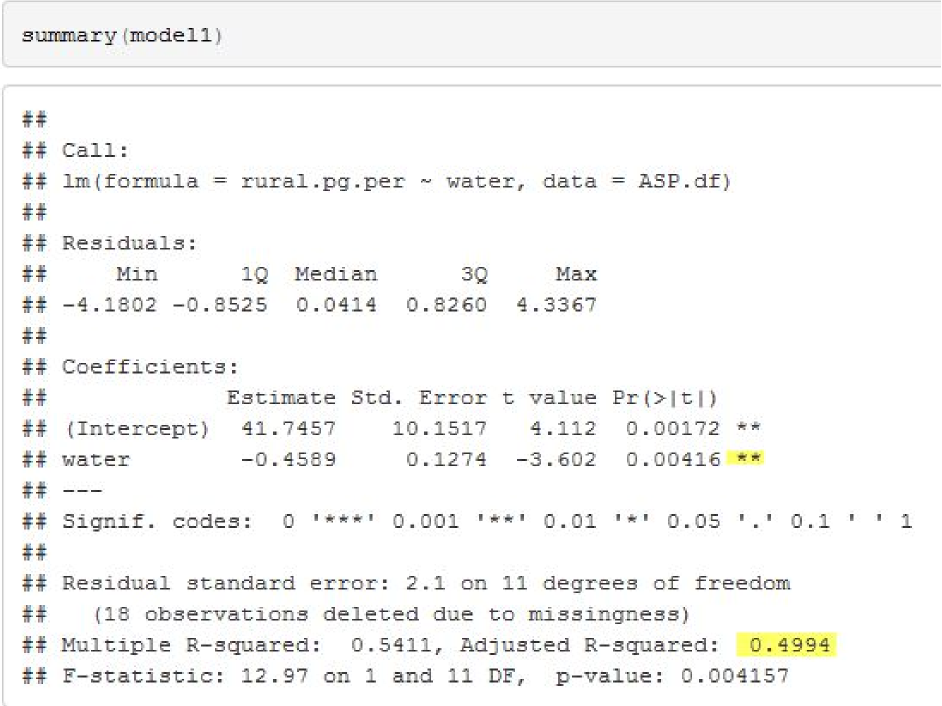
A basic linear regression model estimates the change in the dependent variable (Rural Percent under $2/Day), given a change in the community development indicator value. The model calculates a coefficient, a numerical constant placed before a multiplying variable, that holds all other explanatory variables constant. The formula for a simple linear model is :
Y (dependent) = a + bx + e, where “a” is the intercept, “b” is the coefficient times “x” multiplier, and “e” equals errors.

I ran a simple regression model and created a plot for each of the four sustainable measures in relation to the rural percentage of the population living in poverty. The important statistic to observe in this model is R^2. It will tell us the variation in y that can be explained by the variation in x. Basically, it verifies whether the sustainable measure is or is not helping significantly reduce poverty in Asia.

I called the linear regression function by using the lm() function below.
model1 <- lm(*dependent ~ independent,* data = ASP.df). I went on to inspect the variable significance indicated by one to three asterisks “***” or “.”, as well as the adjusted R^2 (R^2 adjusted for the number of data points); both are highlighted in the findings below.

Findings

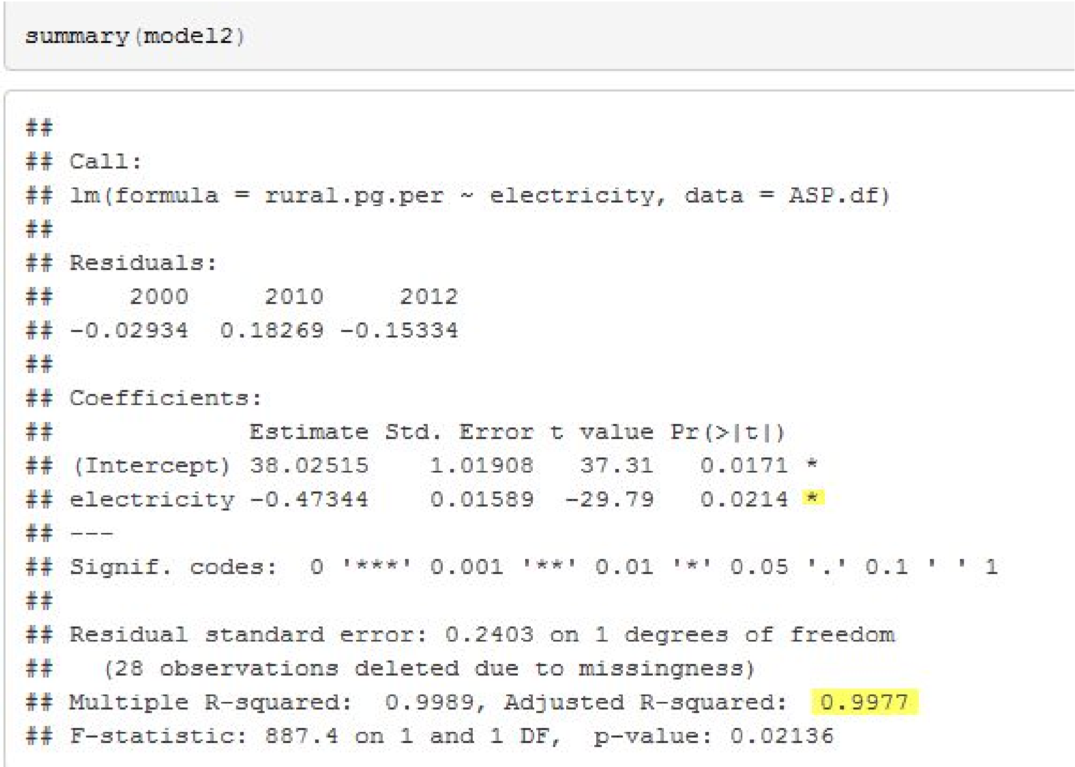
Model 1: Water



```
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
```

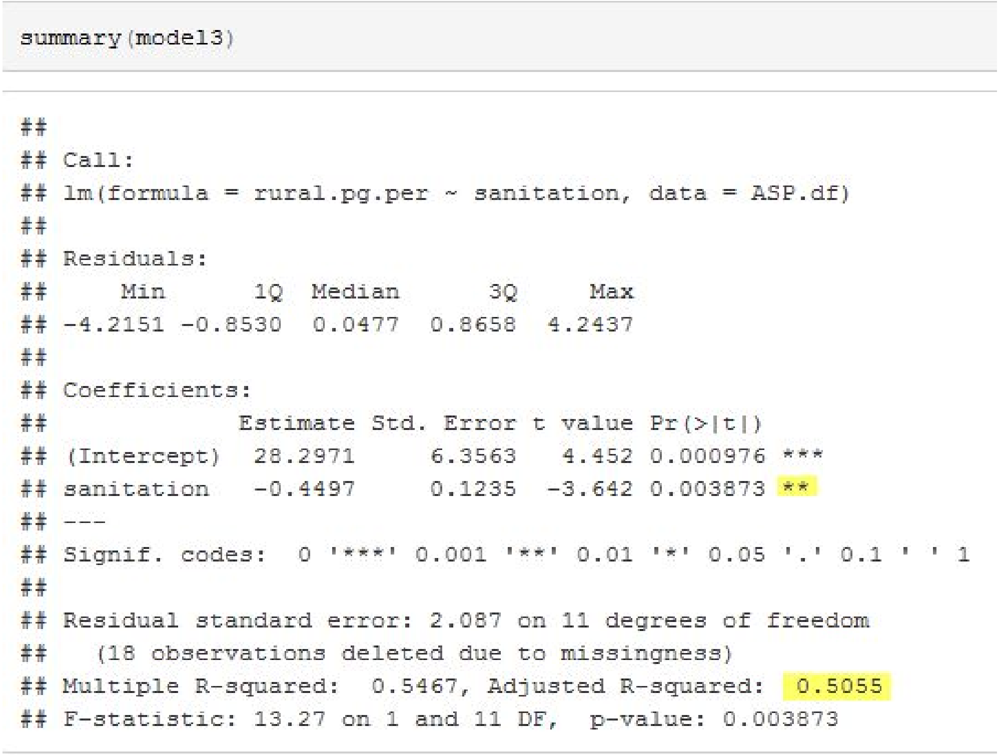
Model 2: Electricity



```
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
```

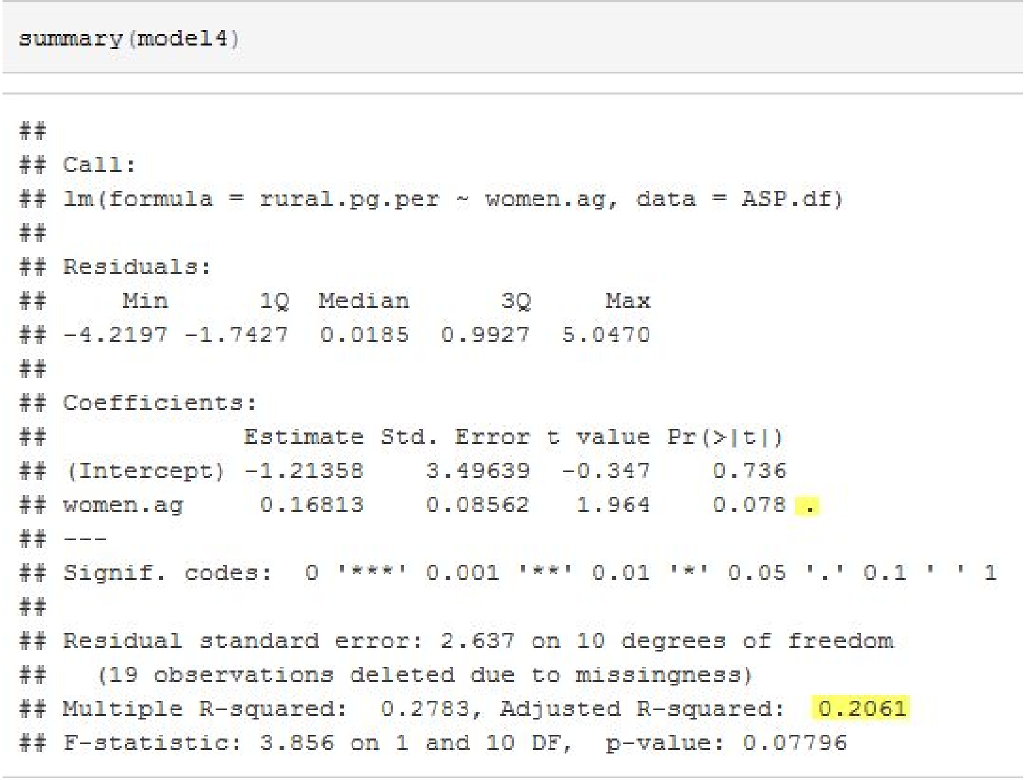
Model 3: Sanitation



```
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
```

Model 4: Women in Agriculture



```
summary(model4)
```

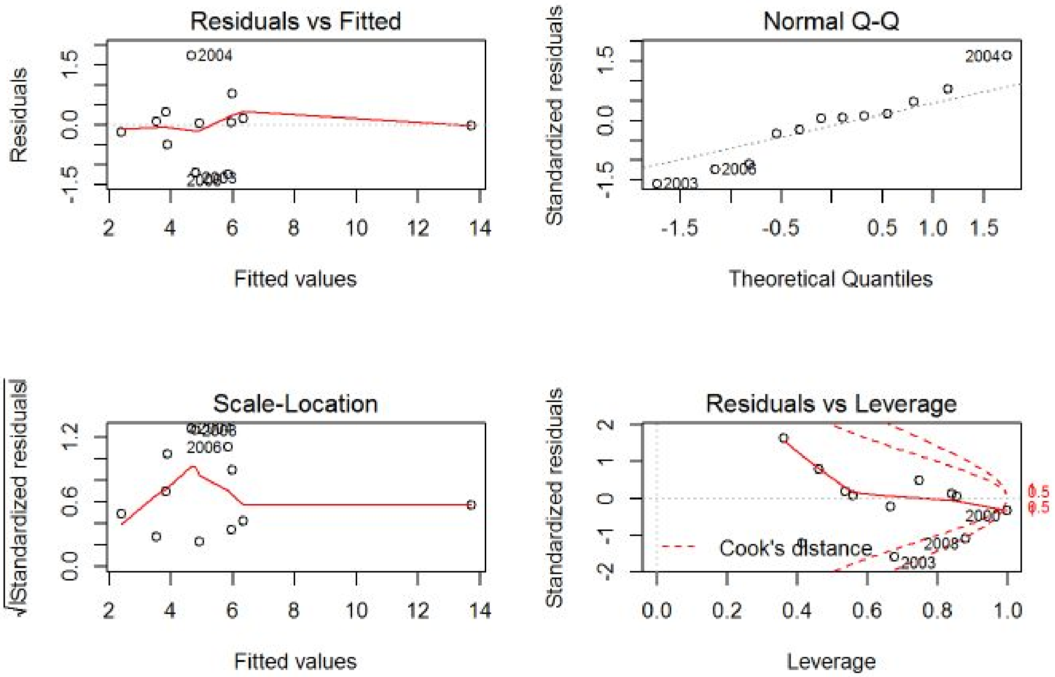
```
##
## Call:
## lm(formula = rural.pg.per ~ women.ag, data = ASP.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2197 -1.7427  0.0185  0.9927  5.0470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.21358    3.49639  -0.347    0.736
## 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
```

By analyzing the Adjusted R Squared and variable significance in tandem, I can confirm or deny more about my original hypothesis. Water was significant with two stars and a .49 adjusted R squared. Electricity had an even higher (.99) adjusted r squared, but only one star for significance. Sanitation had the strongest linear relationship, with two stars and an adjusted R squared 1 point higher than water. Finally, Women's participation in agriculture was barely significant (indicated by the ".") and had weak adjusted R squared of about .21. Although, none of the variables alone had a very strong adjusted R squared, I can continue my analysis by using Multiple Regression to see how I can manipulate or conjoin variables to get an adjusted R squared closer to 1; the gauge of a perfectly fit regression model. After the best model was chosen, I ran a residual test to calculate the Root Mean Squared Error. This measures the magnitude of errors made in the analysis.

Multiple Linear Regression

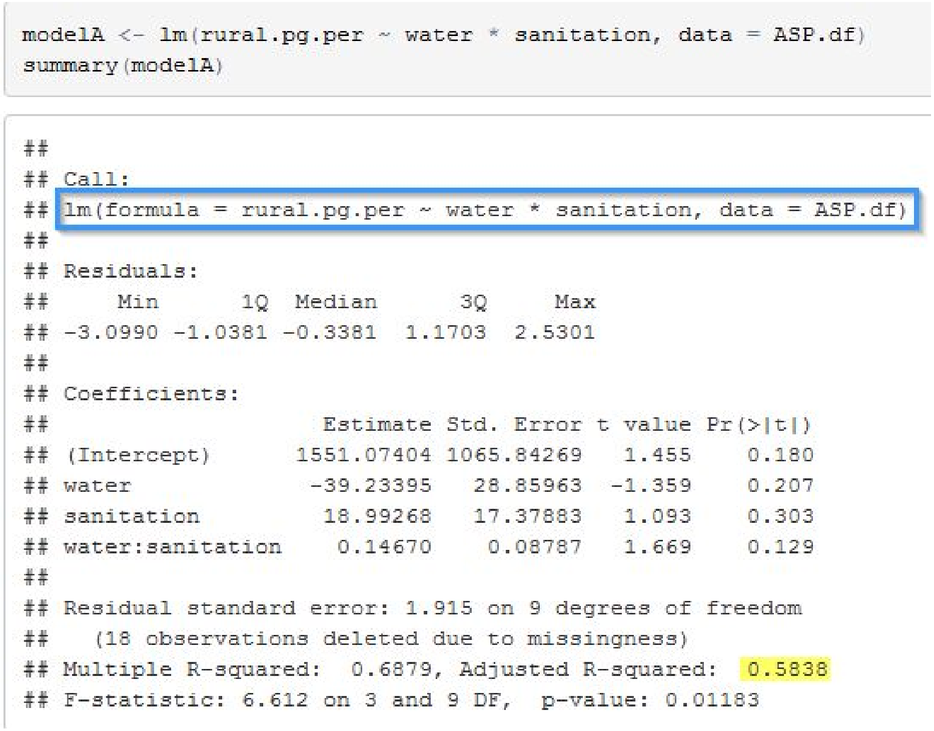
<u>Approach</u>

Performing a multiple linear regression will return a statistical analysis of the rural poverty gap's response on two or more predictive measures. The formula is essentially the same:
example.model <- lm(rural.pg.per ~ independent var. + independent var. , data = ASP.df). I was able to run this additive version of the linear regression model, but my adjusted R squared always came back lower, which signaled to me that the model was becoming less accurate the more indicators I added on. However, I did find that the adjusted R squared notably increased when I took the product of two or more of the factors. I created and ran many different models, but narrowed it down to the best fitted version of the lm() formula. The best fitting model is the one I will choose to run a predictive analysis one.



Findings

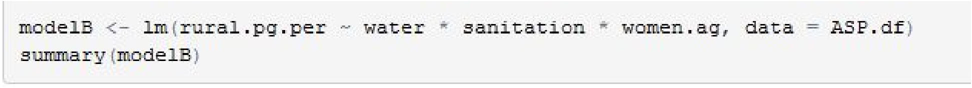
Model A: Water * Sanitation



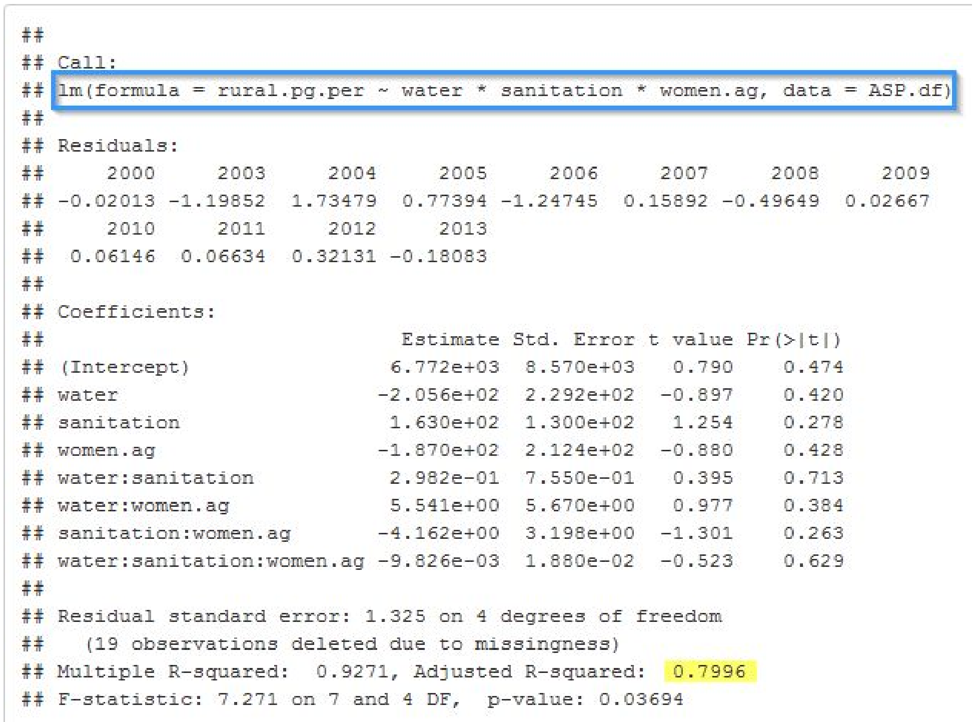
```
modelA <- lm(rural.pg.per ~ water * sanitation, data = ASP.df)
summary(modelA)
```

```
##
## 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
```

Model B: Water * Sanitation * Women’s Empowerment



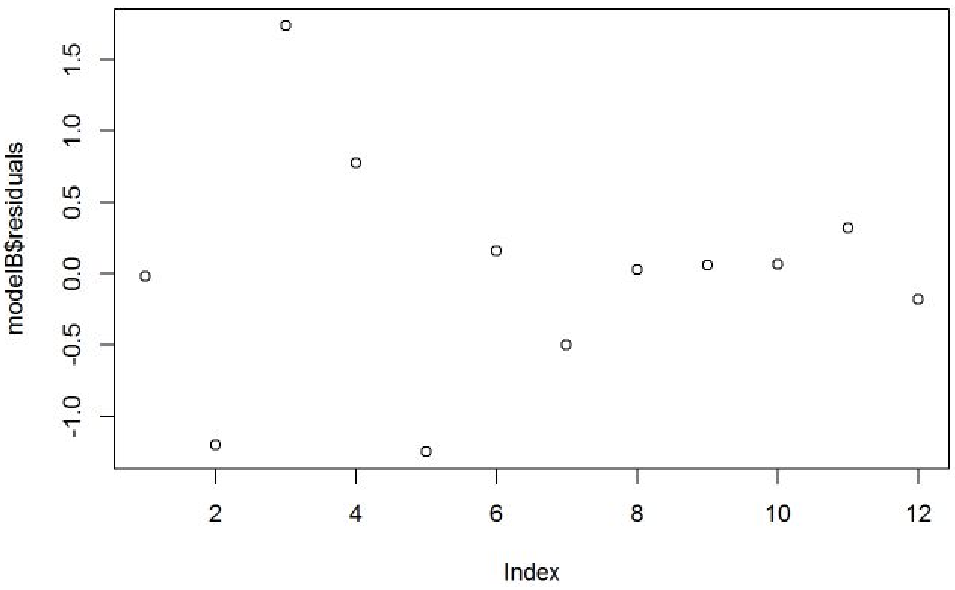
```
modelB <- lm(rural.pg.per ~ water * sanitation * women.ag, data = ASP.df)
summary(modelB)
```



```
## 
## 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
```

These results helped me deduce that the three indicators: access to water, improved sanitation, and women’s empowerment all compounded together would provide the optimal environment to reduce poverty in these Asian countries. Of every model test, Model B won out with an adjusted R squared of .799 (almost .80); which is significantly better considering our best single linear regression only had .50.

But can I really trust these results? The Root Mean Squared Error was = .47, which is excruciatingly high. The residual plot below shows standard errors that are not normally distributed, but skewed all over the plot.



Testing Predictions

Approach

Although I'm aware of errors, I'm going to keep pushing forward with how this process could predict future outcomes with Model B. I review the data around rural poverty gap percentages to assign an average, low, and high level of percentages to grade the new data test against. The first object I create is povertygap1 (the mean with NA values removed), poverty gap 2 (1st quartile average), and poverty gap 3 (3rd quartile average) from the Asia data frame. Then, I call predict (modelB).

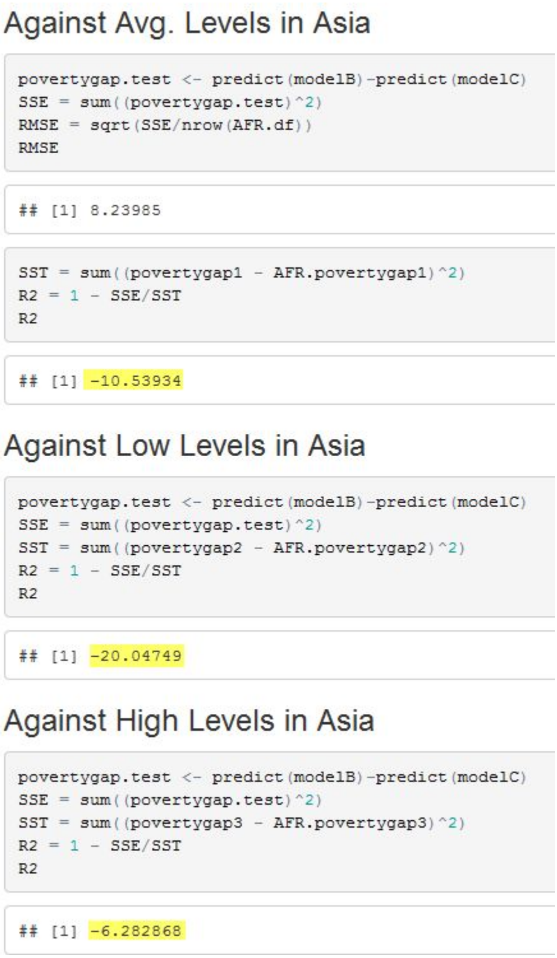
Now, I need a new data frame with the same predictors as the original model. I go back to square one, but this time, with new data from 10 developing nations in Africa. I repeat the WB_data parsing, the restructuring, renaming, and then setting up a data frame called AFR.df. I inspect the data to make sure it is similar and assign the same povertygap objects as I did with the poverty gap percentages for Asia. This time I create Model C; identical to model B, but with AFR.df data. I predict(Model C) as I did with B as well. I will used these povertygap objects and Predictive B and C models to calculate the RMSE for average, low, and high percentages of poverty in these two separate regions to the world, in order to see at which level the World Bank indicators might correlate.

Findings

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.

The final R squared analysis on the predictive poverty gap test set gave us an R squared of -10.54 for average levels, -20.05 for low levels of poverty, and -6.28 for high levels of poverty.

Since all test came back with a negative R2, we know that this model fits worse than if there were just a horizontal line of the mean. Now, all of these are poor outcomes for R squared, but I could argue that from a high level, our hypothesis should be further researched and could potentially be correct if more data was collected. Where there were high levels of extremely impoverished people groups, the sustainable development indicators were more significantly correlated with reducing the poverty and produced less errors in the data for Africa.



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
```

Conclusion and Recommendations

★ 1. Poverty Line (<\$2/day) = $\beta o + \beta 1WATER1 + \beta 1\ SANITATION1 + \beta 1WOMEN1 + u$
    - Invest more in more programs that support indicators with the highest correlation to lowering population living below the poverty line: water, sanitation and then women's agricultural practice
    - Get more research and data on electricity and re-run analysis; potentially highly predictive

2. NGO’s and Governments need to collect A LOT more data
    - Ensure standardization of data collected
    - Test other vastly developing regions in the world (South America)

3. Explore and analyze other sustainable techniques:
    - Agricultural training
    - Access to internet/technology
    - Risk management planning

4. Take other economic and sustainability metrics into account for analysis
    - Education/Literacy
    - Prominent Industries/local resources
    - Exchange/Inflation Rates
    - Openness of Government