# Final Project: Analysis of Standard of Living and Economic Performance of Countries

**Author: Alden Tan, Justin Du**

**Discussants: [List people who helped you with the project and/or websites used]**

## Introduction

The aim of this project is two-fold:

1) to investigate the differences in standard of living indicators between more economically developed countries (MEDCs) and less economically developed countries (LEDCs); and
2) to build a model that predicts the gross national income (GNI) of a country.

The second objective builds on the first. After investigating the standard of living indicators in which MEDCs and LEDCs differ, we will use these indicators, together with others, to build the prediction model.

This topic is interesting because it has a multitude of potential policy applications. Policymakers in countries can make use of such a model to identify possible reasons for the level of economic performance of the country, and deduce ways to improve or further strengthen their economic situation.

The data is obtained from the World Bank. (TODO: insert link) As the World Bank is a popular source of data, other analyses may have been done using this data to investigate the economic, health and social situations of differing countries. However, we have not come across an analysis attempting to build the model aforementioned.

## Results

## Analysis 1:

In this analysis, we want to see whether a country's level of economic development is associated with the percentage of that country's population living in rural areas. We hypothesize that the variable `Rural` will be a good economic indicator because rural areas tend to be sparsely populated, have low housing density, and are far from urban centers, which may represent fewer amounts of transportation, less infrastructure, and less commerce, all signs of poorer economic development.

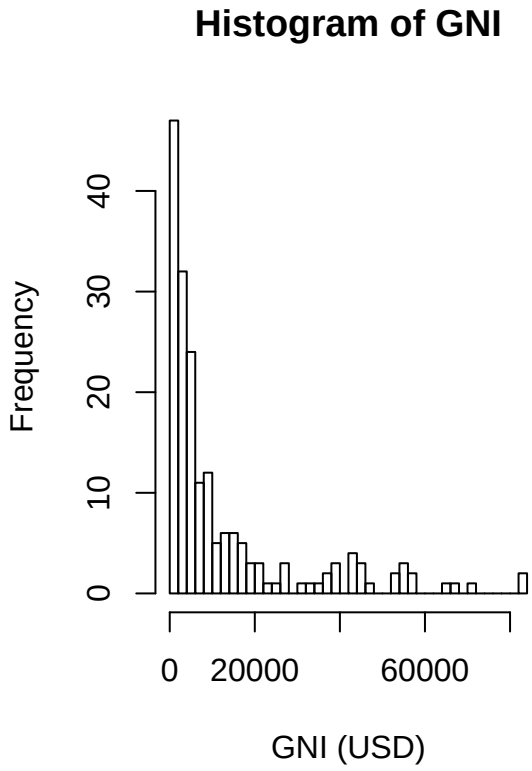
According to this 2017 report by the United States Census Bureau, urban areas make up only 3 percent of the entire land area of the United States but are home to more than 80 percent of the population. Conversely, 97 percent of the country's land mass is rural but only 19.3 percent of the population lives there.

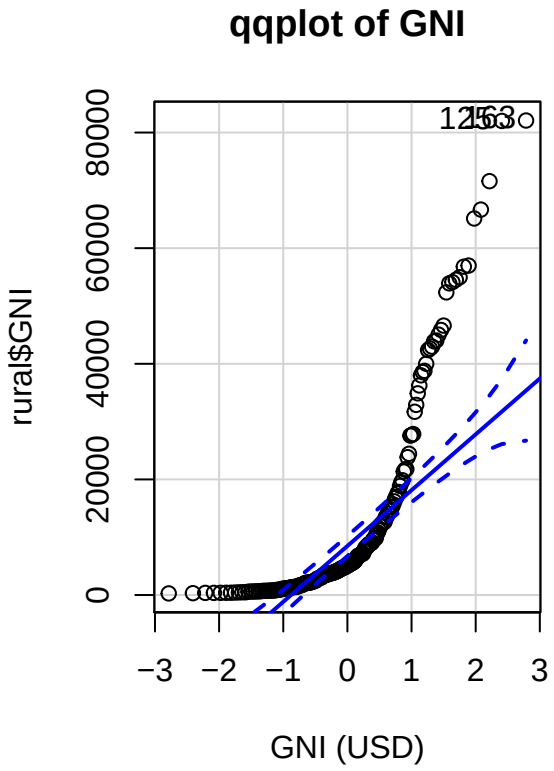
In this analysis, we run a permutation test to see whether there is a significant correlation between income levels and the percentage of a country living in rural areas.

### Data Wrangling

Before we begin, we suspect that `GNI` will be right skewed (most countries will have rather low incomes). In order to verify this, we make two diagnostic plots: a histogram and qqplot of `GNI`.

```
rural <- world_bank_2016 %>% dplyr::select(GNI, Rural) %>% na.omit()
par(mfrow = c(1, 2))
# Histogram
hist(rural$GNI, breaks = 50, main = "Histogram of GNI", xlab = "GNI (USD)")
# Quantile Plot
car::qqPlot(rural$GNI, main = "qqplot of GNI", xlab = "GNI (USD)")
```




```
## [1] 163 125
```

```
# Correlation
cor(rural$GNI, rural$Rural)
```

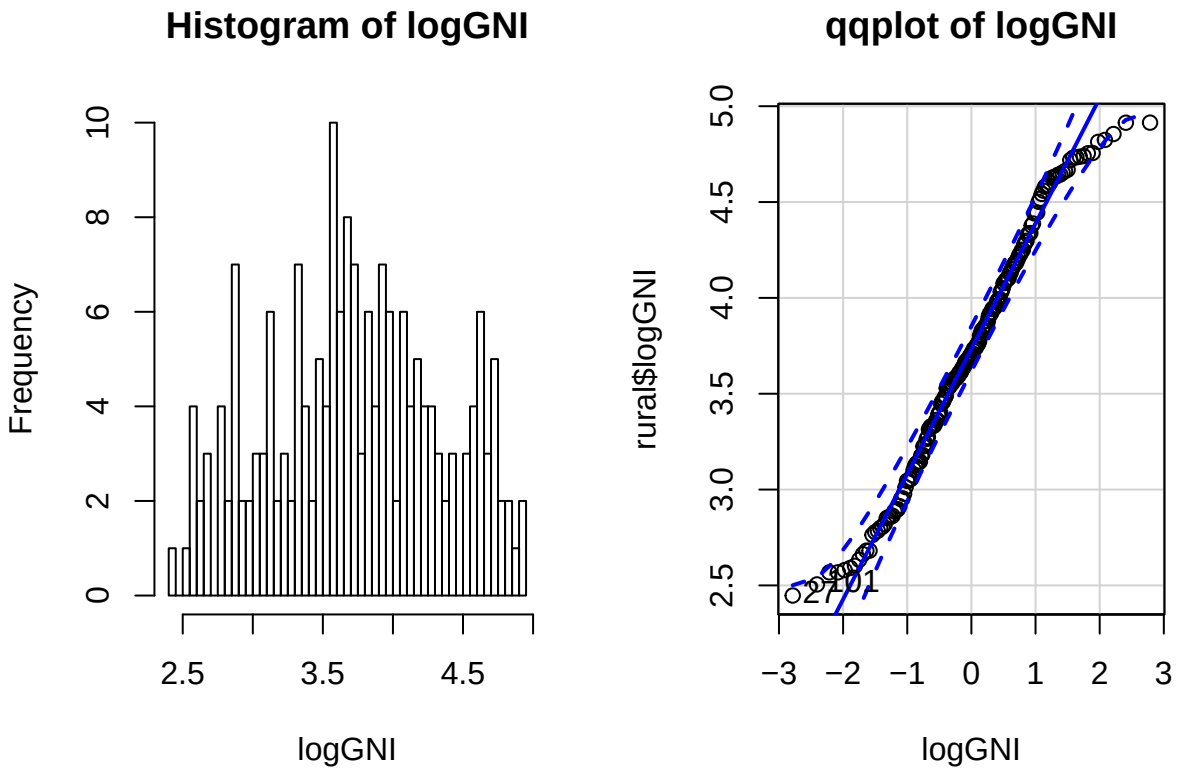
```
## [1] -0.6030187
```

As we can see from the histogram, most GNIs are clustered towards lower incomes. Furthermore, as we see the that many of the points at higher quantiles have significantly higher GNIs than expected. Both of these point to a high degree of right skewness.

Thus, in order to fix this skewness, we transform `GNI` into `logGNI` before we run our correlation test. This transformation will also help us later when we make our predictions using a linear fit.

```
rural <- rural %>% mutate(logGNI = log10(GNI)) %>% na.omit()
```

```
par(mfrow = c(1, 2))
# Histogram
hist(rural$logGNI, breaks = 50, main = "Histogram of logGNI", xlab = "logGNI")
# Quantile Plot
car::qqPlot(rural$logGNI, main = "qqplot of logGNI", xlab = "logGNI")
```



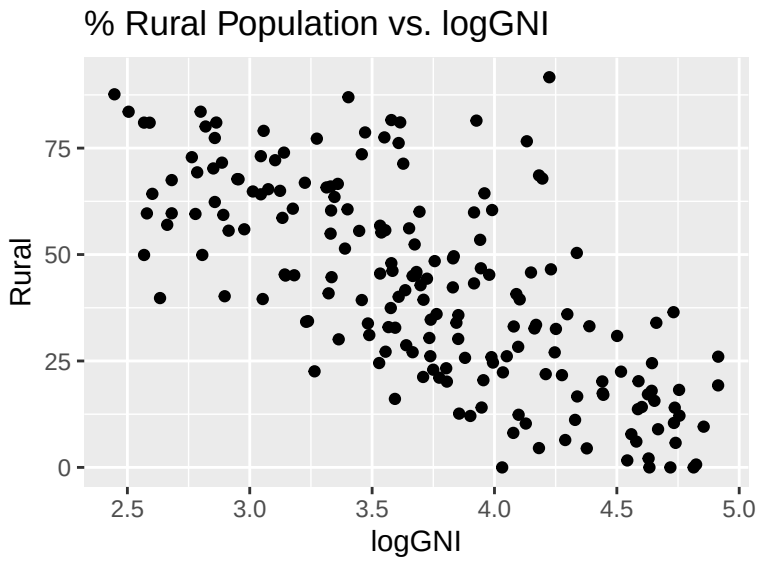

```
## [1]  27 101
```

The diagnostic plots of the histogram and the qqplot look much more normal now! Let us run the the correlation test between `logGNI` and `Rural`!

### Analysis - Permutation Test for Correlations (`logGNI` and `Rural`)

First we make a scatter plot.

```
ggplot(data = rural, aes(x = logGNI, y = Rural)) +
  geom_point() +
  ggtitle("% Rural Population vs. logGNI")
```




It appears that there is a moderately strong negative correlation between `logGNI` and `Rural`. This makes sense, as the wealthier a country is, the less likely that its citizens live in rural areas, hence a lower value for `Rural`. In regards to conditions for linear regression (for the analysis below), we see that the data seems to have equal variance throughout, fulfilling the equal variance assumption.

### 1. Hypotheses

Null Hypothesis: There is no correlation between the percentage of a country living in a rural area and the logGNI of a country. That is, the correlation equals 0.

Alternative Hypothesis: There is a correlation between the percentage of a country living in a rural area and the logGNI of a country. That is, the correlation does not equal 0.

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

### 2. Observed Statistic

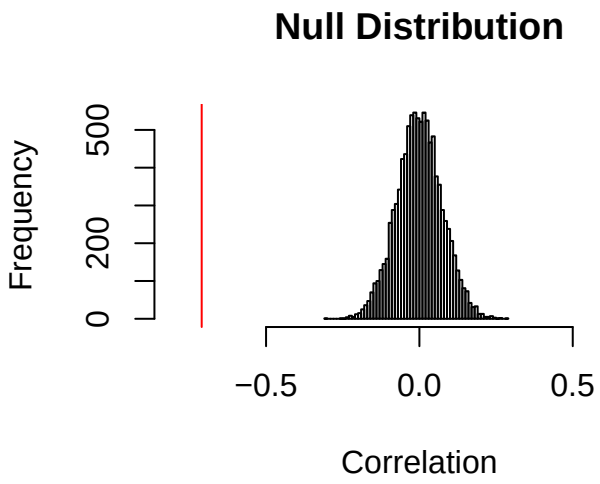
```
(obs_stat_corr_rural_logGNI <- cor(rural$Rural, rural$logGNI))
```

```
## [1] -0.7099465
```

### 3. Create null distribution

```
null_dist_corr_rural_logGNI <- rep(NA, 10000)
for (i in 1:10000){
  null_dist_corr_rural_logGNI[i] <- cor(sample(rural$Rural), rural$logGNI)
}
# Plot the null distribution with a red vertical line for obs stat
hist(null_dist_corr_rural_logGNI, main = "Null Distribution",
```

```
     xlab = "Correlation", xlim = c(-0.8, 0.8), nclass = 50)
# So far off that it doesn't even matter
abline(v = obs_stat_corr_rural_logGNI, col = "red")
```




#### 4. Calculate p-value

```
(p_value_corr_rural_logGNI <-
   sum(abs(null_dist_corr_rural_logGNI) >=
         abs(obs_stat_corr_rural_logGNI)) /
          length(null_dist_corr_rural_logGNI))
```

```
## [1] 0
```

#### 5. Conclusion

A p-value of 0 is not consistent with there being no correlation between percentage of the population living in rural areas and the logGNI. Our null is that the correlation coefficient is 0, so a p-value of 0 means that assuming the null is true, there is a 0% chance that we would observe a correlation coefficient statistic as extreme as ours (-0.7127915).

Thus, we see that logGNI and percentage of rural population are quite correlated with each other.

## Analysis 2: Relationship between income levels and fossil fuel energy consumption

In this analysis, we split the countries into low-income countries, lower-middle-income countries, higher-middle-income countries, and high-income countries according to their GNI per capita. The cutoff points used are obtained from the United Nations (https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf).

Then, we run an ANOVA to see if there is any difference in the mean fossil fuel energy consumption (as a percentage of total energy consumption) between these 4 groups.

### Data wrangling

We create a function to determine the development level of each country based on its GNI, and add a column to the data frame called `development`.

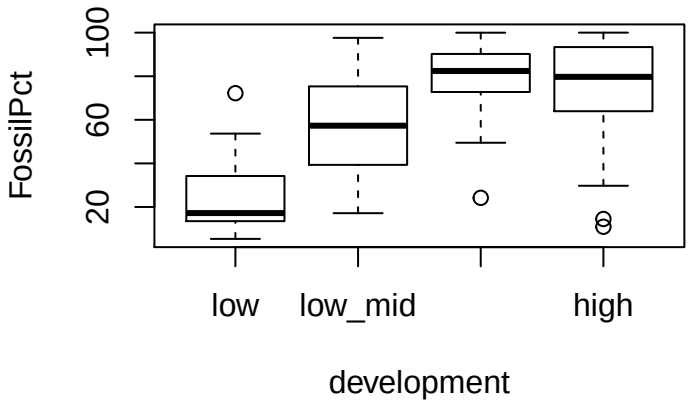
```
# function to convert GNI to development level for each country
convert_GNI_to_level <- function(income){
  if (income <= 1035){
    "low"}
  else if (income > 1035 & income <= 4085){
    "low_mid"}
  else if (income > 4085 & income <= 12615){
    "high_mid"}
  else{
    "high"}
}
# remove cases where either GNI or FossilPct is NA
many_means_data <- world_bank_2016[!is.na(world_bank_2016$GNI), ]
many_means_data <- many_means_data[!is.na(many_means_data$FossilPct), ]
# add a column called development to give development level of country
development_list <- sapply(many_means_data$GNI, convert_GNI_to_level)
many_means_data$development = development_list
```

### Data visualization

```
# boxplot to visualize the data

## for ordering of boxplot
many_means_data$development <- factor(many_means_data$development,
                                      levels=c("low", "low_mid", "high_mid", "high"))

boxplot(FossilPct ~ development, many_means_data)
```



### Analysis: ANOVA to compare 4 means

We run a hypothesis test using the ANOVA on the following hypothesis:

$H_0 : \mu_{low} = \mu_{mid-low} = \mu_{mid-high} = \mu_{high}$

$H_A : \mu_i \neq \mu_j$ for some $i, j \in \{\text{low, mid-low, mid-high, high}\}$, $i \neq j$

$\alpha = 0.05$

```
# run an ANOVA
fit <- aov(FossilPct ~ development, many_means_data)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## development   3  30523   10174   22.82 7.89e-12 ***
## Residuals   124  55294     446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value $= 0.00000000000789 < 0.05 = \alpha$, so we can conclude that there is strong evidence for a difference in mean fossil fuel energy consumption between the different development levels.

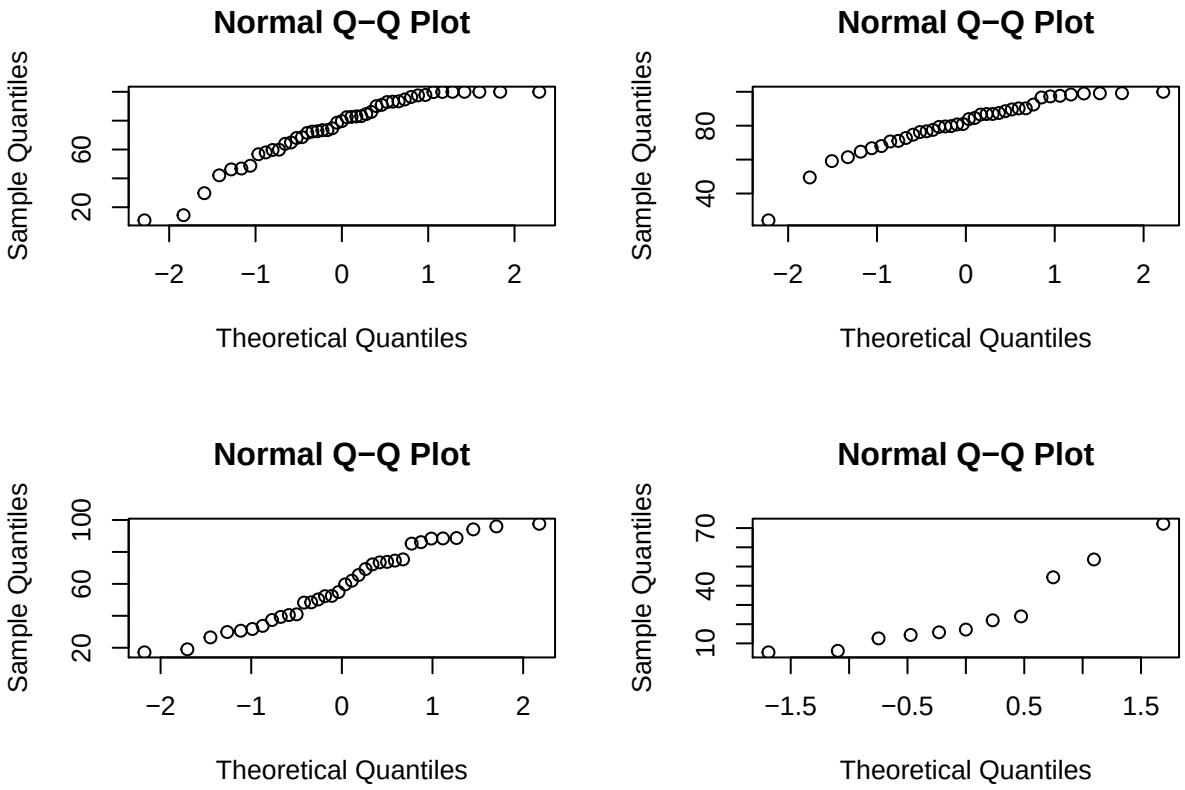
### Checking ANOVA assumptions

We note that the assumptions underlying a one-way ANOVA are:

- Data in each group come from a normal distribution
- Homoscedasticity: Each group has equal variance

To check for normal distribution, we plot qqplots for each group:

```
par(mfrow = c(2, 2))
# high development
high_ind <- many_means_data[many_means_data$development == "high", ]
qqnorm(high_ind$FossilPct)
# high-mid development
high_mid_ind <- many_means_data[many_means_data$development == "high_mid", ]
qqnorm(high_mid_ind$FossilPct)
# low-mid development
low_mid_ind <- many_means_data[many_means_data$development == "low_mid", ]
qqnorm(low_mid_ind$FossilPct)
# low development
low_ind <- many_means_data[many_means_data$development == "low", ]
qqnorm(low_ind$FossilPct)
```




We note that the qqplots are not straight lines and hence normality is not strictly fulfilled, but the deviation is still generally acceptable because the ANOVA is rather resistant to violations of the normal assumption.

To check for homoscedasticity, we see if the standard deviations in each group are similar:

```
# are standard deviations in each grp similar?
by(many_means_data$FossilPct, many_means_data$development, sd)
```

```
## many_means_data$development: low
## [1] 21.39622
## ------------------------------------------------------------
## many_means_data$development: low_mid
## [1] 23.84506
## ------------------------------------------------------------
## many_means_data$development: high_mid
## [1] 15.69547
## ------------------------------------------------------------
## many_means_data$development: high
## [1] 22.78229
```

The standard deviations are quite similar, except for the high-mid development level, and this is a limitation to this analysis.

### Pairwise comparisons

To investigate which pairs of means differ, we perform pairwise t tests with the Bonferroni correction to be conservative.

```r
pairwise.t.test(many_means_data$FossilPct, many_means_data$development, p.adj = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  many_means_data$FossilPct and many_means_data$development
##
##          low     low_mid high_mid
## low_mid  0.00010 -       -
## high_mid 5.0e-11 0.00015 -
## high     1.4e-09 0.00569 1.00000
##
## P value adjustment method: bonferroni
```

We see that every pair of mean fossil fuel energy consumption differs at a statistically significant level, except for the high-mid and high development categories. This suggests that the high-mid and high development categories may have similar fossil fuel energy consumptions.

**Data wrangling: Change the subtitle here to describe what you are plotting etc.**

**Visualize the data: Change the subtitle here to describe what you are plotting etc.**

**Analyses: Sub-title about the analyses/models you are using**

## Conclusion

### Future Directions

## Reflection

## Appendix