What Factors Determine Gross National Income?

Elliott Cook, Daniel Kim, Brian Sun, Max Sun

## Introduction

One of the great economists of the 20th century, Robert Lucas, once said:

I do not see how one can look at [GDP] figures like these without seeing them as representing possibilities … The consequences for human welfare involved in questions like these are simply staggering: Once one starts to think about them, it is hard to think about anything else.

In the spirit of this quote, we will explore what factors impact GNI per capita in countries around world. Unlocking the patterns to economic growth could have staggering impacts on the development of nations. There are a countless number of potential factors: location, urbanization, openness to trade, etc. We will begin by analyzing the impacts of individual factors and eventually create a general linear model that seeks to use a multitude of categorical and quantitative variables to predict GNI per capita.

## Data Gathering and Cleaning

First, let’s get some relevant data from the World Bank (2016).

wb <- read.csv("WB.2016.csv")

Our first categorical variable will be called “Continents”, and it will contain the Continent of each country. We will scrape the data from a website that has the name of the continent and all of the nations in that continent right below it. After matching the country names with the World Bank data, we will use some for loops and while loops to assign each nation a continent, with “Other” denoting a territory that has a row in the World Bank data.

# Url for site with continents and list of countries underneath  
url <- "https://www.worldatlas.com/cntycont.htm"  
webpage <- read\_html(url)  
countries <- html\_text(html\_nodes(webpage, 'ul+ h2 , p~ ul li , p+ h2'))  
  
# Clean some names to match with World Bank  
wb$Country <- gsub("Korea, Dem.\*", "Korea, Dem\\. People Rep\\.", wb$Country)  
to\_replace <- c("Bahamas", "Cape Verde", "Brunei", "Burkina", "Congo, Democratic Republic of", "^Congo$", "Ivory Coast", "Egypt", "Gambia", "Iran", "Korea, North", "Korea, South", "Kyrgyzstan", "Laos", "Macedonia", "Micronesia", "Burma \\(Myanmar\\)", "Syria", "East Timor", "Yemen", "Venezuela", "Saint", "Slovakia")  
replace\_with <- c("Bahamas, The", "Cabo Verde", "Brunei Darussalam", "Burkina Faso", "Congo, Dem. Rep.", "Congo, Rep.", "Cote d'Ivoire", "Egypt, Arab Rep.", "Gambia, The", "Iran, Islamic Rep.", "Korea, Dem. People Rep.", "Korea, Rep.", "Kyrgyz Republic", "Lao PDR", "Macedonia, FYR", "Micronesia, Fed. Sts.", "Myanmar", "Syrian Arab Republic", "Timor-Leste", "Yemen, Rep.", "Venezuela, RB", "St.", "Slovak Republic")  
  
for (i in 1:length(to\_replace)) {  
 countries <- gsub(to\_replace[i], replace\_with[i], countries)  
}  
  
# Construct vector for continents  
continents <- rep(NA, length(wb$Country))  
for (i in 1:nrow(wb)) {  
 if (!wb$Country[i] %in% countries) {  
 continents[i] <- "Other"  
 next  
 } else {  
 index <- which(countries == wb$Country[i])[[1]]  
 while (TRUE) {  
 index <- index - 1  
 if (grepl("AFRICA", countries[index])) {  
 continents[i] <- "Africa"  
 break  
 } else if (grepl("ASIA", countries[index])) {  
 continents[i] <- "Asia"  
 break  
 } else if (grepl("EUROPE", countries[index])) {  
 continents[i] <- "Europe"  
 break  
 } else if (grepl("N\\. AMERICA", countries[index])) {  
 continents[i] <- "North America"  
 break  
 } else if (grepl("OCEANIA", countries[index])) {  
 continents[i] <- "Oceania"  
 break  
 } else if (grepl("S\\. AMERICA", countries[index])) {  
 continents[i] <- "South America"  
 break  
 }  
 }  
 }  
}  
wb$Continent <- continents

Next, we will create another categorical variable, isRural, that will denote whether or not a country is rural or urbanized. isRural will be true for a country with more than 50% of its population living in rural areas.

wb$isRural <- ifelse(wb$Rural > 50, TRUE, FALSE)

With GNI being what we seek to predict, we remove every nation that doesn’t have GNI data.

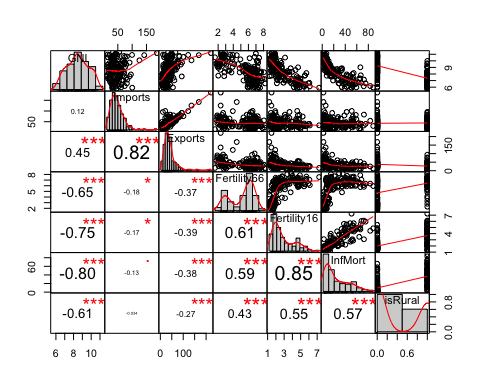
wb <- wb[!is.na(wb$GNI), ]

We first select the variables that we want to use in our analysis (we will go over the units in a little bit). We then clean the dataframe by removing all rows with NAs in any of these columns.

wb <- wb[, c("Country", "GNI", "Imports", "Exports", "Fertility66", "Fertility16", "InfMort", "Continent", "isRural")]  
wb <- na.omit(wb)

Next, we will construct a matrix scatterplot to check for non-linearity in the data (temporarily removing the categorical variables). Since we’ve seen many times that anything related to income will be close to an exponential distribution, we will set GNI per capita to the log of GNI per capita and then make the plot.

wbtemp <- wb[, !(colnames(wb) %in% c("Country", "Continent", "Surplus"))]  
wbtemp$GNI <- log(wb$GNI)  
  
# Get the 'pairsJDRS' function  
source("http://www.reuningscherer.net/s&ds230/Rfuncs/regJDRS.txt")  
pairsJDRS(wbtemp)



Based on the distributions, we will make the following transformations to normalize some of the data and linearlize some of the relationships: \* log(GNI) \* log(Imports) \* log(Exports) \* log(InfMort)

Now, we can construct our final dataset.

wb$logGNI <- log(wb$GNI)  
wb$logImports <- log(wb$Imports)  
wb$logExports <- log(wb$Exports)  
wb$logInfMort <- log(wb$InfMort)  
  
wb2 <- wb[, c("Country", "logGNI", "logImports", "logExports", "Fertility66", "Fertility16", "logInfMort", "Continent", "isRural")]

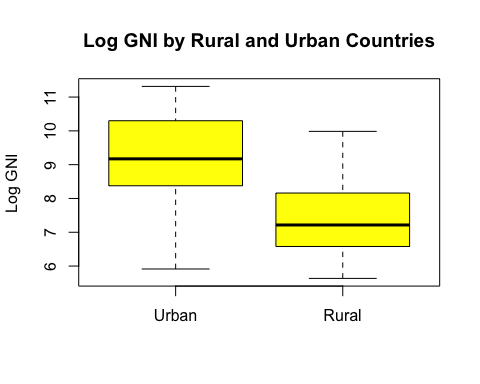
Here is the list of the final variables: \* Country - Name of the Country \* logGNI - Log of GDP per capita, measured in log(dollars per person) \* logImports - Log of imports as a percentage of GDP \* logExports - Log of exports as a percentage of GDP \* Fertility66 - Fertility rate (births per woman), 1966 \* Fertility16 - Fertility rate (births per woman), 2016 \* logInfMort - Log of infant mortaility rate per 1,000 pepole \* Continent - Continent of the country, “Other” if it’s a territory \* isRural - TRUE if >50% of the population is in rural areas, FALSE otherwise

## Urbanization and Growth

In this section, we use a t-test to determine whether or not there exists a difference in logGNI between countries that count as rural (as determind earlier) and those are are more urbanized. From a developmental standpoint, it seems like throughout history, urban centers are places where innovation occurs the most. Thus, we would expect that more urbanization would drive more growth as people move towards more productive jobs that are only available when many people are working in the same area.

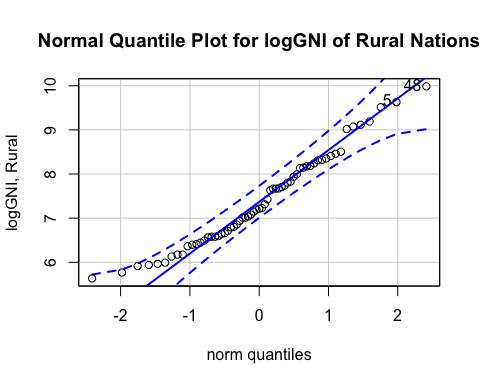
We first make a box-plot of logGNI vs isRural.

boxplot(wb2$logGNI ~ wb2$isRural, main="Log GNI by Rural and Urban Countries", names=c("Urban", "Rural"), ylab="Log GNI", col="yellow")

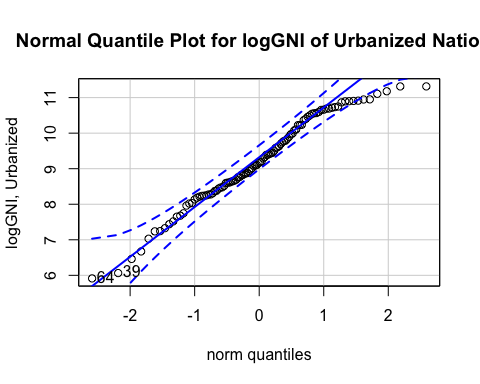


It seems like our prediction is holding true. Before we run the t-test, we check for normality by using a normal quantile plot.

gnir <- wb2[wb2$isRural, "logGNI"]  
gniu <- wb2[!wb2$isRural, "logGNI"]  
qqPlot(gnir, main="Normal Quantile Plot for logGNI of Rural Nations", ylab="logGNI, Rural")



qqPlot(gniu, main="Normal Quantile Plot for logGNI of Urbanized Nations", ylab="logGNI, Urbanized")



Although the distribtion of GNI per capita for urbanized countries might be a little left skewed, both quantile plots show that the data is sufficiently normal. We will also make the simplifying assumption that nations are free to pick and choose their own policies due to the anarchial nature of international relations, so all of our data should be independent as well. Since we meet all of the assumptions, we can conduct a one-sided t-test with

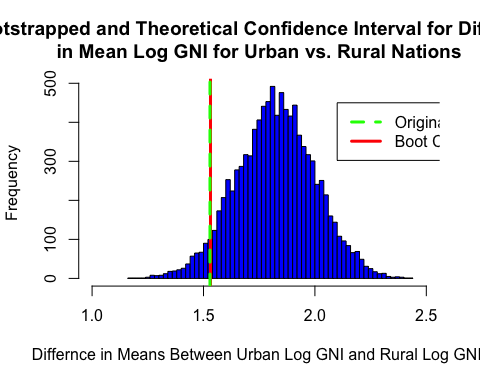
$$
H\_0 : logGNI\_U = logGNI\_R \\
H\_a : logGNI\_U > logGNI\_R
$$

t.test(gniu, gnir, "greater")

##   
## Welch Two Sample t-test  
##   
## data: gniu and gnir  
## t = 10.185, df = 147.78, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 1.529066 Inf  
## sample estimates:  
## mean of x mean of y   
## 9.216275 7.390484

We end up with a p-value of almost zero, which means that there is evidence that nations that are urbanized have higher log(GNI per capita) than countries that are not urbanized. We are also 95% confident that the true impact of urbanization on log(GNI per capita) is greater than 1.77 (the difference of the means). Instead of the t-test, we could also perform a bootstrap, which should yield similar results.

N <- 10000  
diff <- rep(NA, N)  
for (i in 1:N) {  
 data\_1 <- sample(gniu, length(gniu), replace = TRUE)  
 data\_2 <- sample(gnir, length(gnir), replace = TRUE)  
 diff[i] <- mean(data\_1) - mean(data\_2)  
}  
ci <- quantile(diff, c(0.05, 1)) # bootstrapped t-test  
  
t\_tested <- t.test(gniu, gnir, "greater")$conf.int # original t-test  
  
hist(diff,col="blue",main="Bootstrapped and Theoretical Confidence Interval for Difference\nin Mean Log GNI for Urban vs. Rural Nations",xlab="Differnce in Means Between Urban Log GNI and Rural Log GNI ",breaks=50, xlim=range(1,2.5))  
abline(v=ci[1],lwd=3,col="red")  
abline(v=t\_tested,lwd=3,col="green",lty=2)  
legend(2.1,450,c("Original CI","Boot CI"),lwd=3,col=c("green","red"),lty=c(2,1))

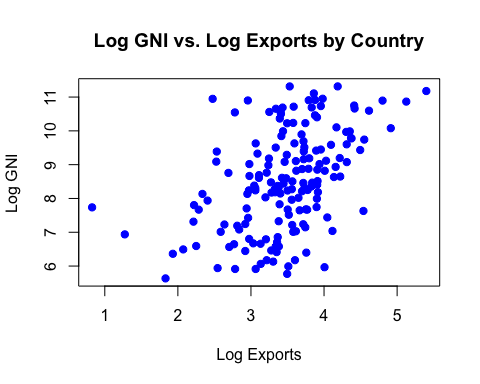


We can see that the 95% confidence interval predicted by the bootstrap is almost exactly the same as that of the t-test. There is no line for the upper bound of the confidence interval since that technically resides at positive infinity. Overall, these results hint towards the fact that having a growing urban population could help with a country’s development.

## Export-Oriented Industrialization

The miraculous growth of the “Asian Tigers” (Hong Kong, Taiwan, South Korea, Singapore) and Japan have been critical to the understanding of economic development. All of these nations followed a pattern of export-oriented industrialization, where a larger share of their economy was built around satisfying international demand and exploiting their comparative advantage. To see how a more export-based economy could affect GNI, we can do a correlation between logExports and logGNI.

plot(wb2$logExports, wb2$logGNI, main="Log GNI vs. Log Exports by Country", xlab="Log Exports", ylab="Log GNI", pch=19, col="blue")



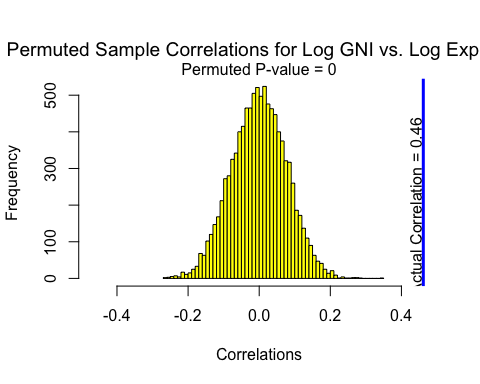
cor.test(wb2$logExports, wb2$logGNI)

##   
## Pearson's product-moment correlation  
##   
## data: wb2$logExports and wb2$logGNI  
## t = 6.683, df = 165, p-value = 3.436e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3330227 0.5732291  
## sample estimates:  
## cor   
## 0.4615442

It seems like there is a positive trend betwen logExports and logGNI. The t-test on Pearson’s correlation gave a highly significant p-value, and we are are 95% confident that the true Pearons’s correlation coefficient lies beteween .29 and .55. We can do a permutation test on correlation, which should give a similar result. We will test

$$
H\_0 : \rho = 0 \\
H\_a : \rho \ne 0
$$

n\_samp <- 10000  
corResults <- rep(NA, n\_samp)  
for(i in 1:n\_samp){  
 #get vector of rows in our fake sample  
 corResults[i] <- cor(sample(wb2$logExports), wb2$logGNI)  
}  
#Two-sided p-value for correlation  
truecor <- mean(abs(corResults) >= abs(cor(wb2$logExports, wb2$logGNI)))  
  
#Make histogram of permuted CORRELATIONS  
hist(corResults, col = "yellow", main = "", xlab = "Correlations", breaks = 50, xlim=range(-.47, .47))  
mtext("Permuted Sample Correlations for Log GNI vs. Log Exports", cex = 1.2, line = 1)  
mtext(paste("Permuted P-value =",round(truecor,4)), cex = 1, line = 0)  
abline(v = cor(wb2$logExports,wb2$logGNI), col="blue", lwd=3)  
text(cor(wb2$logExports,wb2$logGNI)-.02, 200 ,paste("Actual Correlation =", round(cor(wb2$logExports,wb2$logGNI),2)),srt = 90)

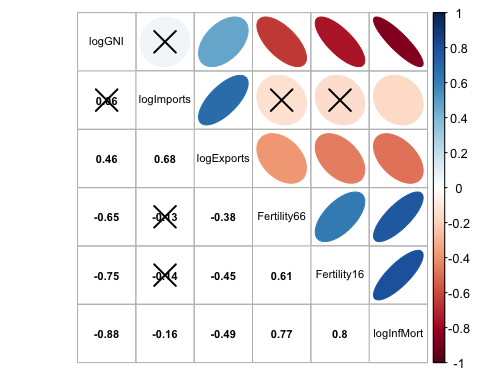


With a p-value of 0, it is clear that we have evidence for a statistically significant relationship between logExports and logGNI. Specifically, a positive correlation indicates that countries that have a higher share of GDP as exports tend to have higher GNI per capita (log scale), providing evidence for our prior beliefs.

## A General Linear Model for Predicting logGNI

In this section we will fit a general linear model to predict logGNI using the variables that we identified at the beginning of this document. First, we will create a correlation plot to check for any potential issues of multicollinearity. We will then perform a backwards stepwise reggression to elimiate factors until all are significant. Lastly, we will analyze the residual plots and conclude.

#Make correlation plot for retained variables  
wbtemp <- wb2[, !(colnames(wb2) %in% c("Country", "Continent", "isRural"))]  
  
sigcorr <- cor.mtest(wbtemp, conf.level = .95)  
corrplot.mixed(cor(wbtemp), lower.col="black", upper = "ellipse", tl.col = "black", number.cex=.7, tl.pos = "d", tl.cex=.7, p.mat = sigcorr$p, sig.level = .05)

 Since we’ve already made our transformations earlier, we won’t need to again. The correlation plot above shows that there might be some multicollinearity among Fertility66, Fertility 16, and logInfMort, as well as between logImports and logExports.

Now, we will fit the model with all of the predictors.

lm1 <- lm(wb2$logGNI ~ wb2$logImports + wb2$logExports + wb2$Fertility66 + wb2$Fertility16 + wb2$logInfMort + wb2$Continent + wb2$isRural)  
  
summary(lm1)

##   
## Call:  
## lm(formula = wb2$logGNI ~ wb2$logImports + wb2$logExports + wb2$Fertility66 +   
## wb2$Fertility16 + wb2$logInfMort + wb2$Continent + wb2$isRural)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.47708 -0.42844 -0.06494 0.42443 2.08729   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.92302 0.57161 20.859 < 2e-16 \*\*\*  
## wb2$logImports -0.62435 0.16261 -3.840 0.000179 \*\*\*  
## wb2$logExports 0.51876 0.13498 3.843 0.000177 \*\*\*  
## wb2$Fertility66 -0.02358 0.05189 -0.454 0.650209   
## wb2$Fertility16 -0.05369 0.07415 -0.724 0.470133   
## wb2$logInfMort -0.95710 0.10480 -9.133 3.45e-16 \*\*\*  
## wb2$ContinentAsia 0.14659 0.18996 0.772 0.441467   
## wb2$ContinentEurope -0.20724 0.27438 -0.755 0.451221   
## wb2$ContinentNorth America 0.27142 0.22824 1.189 0.236186   
## wb2$ContinentOceania 0.78336 0.29979 2.613 0.009857 \*\*   
## wb2$ContinentSouth America 0.12374 0.25731 0.481 0.631257   
## wb2$isRuralTRUE -0.31373 0.13314 -2.356 0.019705 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6371 on 155 degrees of freedom  
## Multiple R-squared: 0.8236, Adjusted R-squared: 0.8111   
## F-statistic: 65.8 on 11 and 155 DF, p-value: < 2.2e-16

The model accounts for approximately 82% of the variance in the data, which is pretty good. However, some of the predictors are nonsignificant, so we will perform backwards stepwise regression and remove them, starting with Fertility66.

fin <- lm(wb2$logGNI ~ wb2$logImports + wb2$logExports + wb2$logInfMort + wb2$Continent + wb2$isRural)  
summary(fin)

##   
## Call:  
## lm(formula = wb2$logGNI ~ wb2$logImports + wb2$logExports + wb2$logInfMort +   
## wb2$Continent + wb2$isRural)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.48991 -0.42681 -0.07003 0.43207 2.09843   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.71372 0.51751 22.635 < 2e-16 \*\*\*  
## wb2$logImports -0.64786 0.15943 -4.064 7.61e-05 \*\*\*  
## wb2$logExports 0.54289 0.13077 4.152 5.40e-05 \*\*\*  
## wb2$logInfMort -1.00213 0.08976 -11.165 < 2e-16 \*\*\*  
## wb2$ContinentAsia 0.20484 0.16393 1.250 0.21330   
## wb2$ContinentEurope -0.08667 0.23302 -0.372 0.71044   
## wb2$ContinentNorth America 0.35237 0.20216 1.743 0.08329 .   
## wb2$ContinentOceania 0.84019 0.29035 2.894 0.00435 \*\*   
## wb2$ContinentSouth America 0.20234 0.23518 0.860 0.39091   
## wb2$isRuralTRUE -0.32170 0.13130 -2.450 0.01538 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6346 on 157 degrees of freedom  
## Multiple R-squared: 0.8227, Adjusted R-squared: 0.8126   
## F-statistic: 80.96 on 9 and 157 DF, p-value: < 2.2e-16

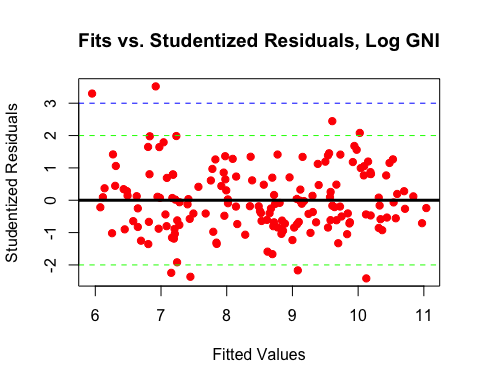
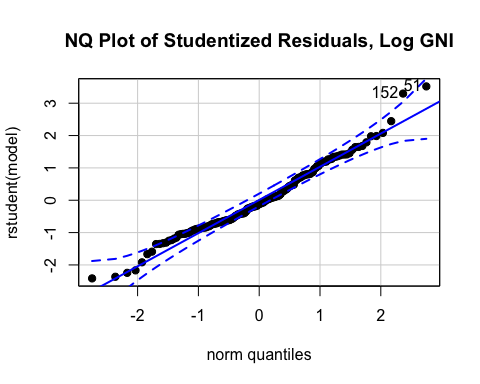
We ended up removing Fertility66 and Fertility16, which was a little surprising given the research documenting the impact of a demographic transition on economic growth. At this point, all predictors are significant at the alpha = 0.05 level (one level of Continent is, so we keep that in).

The following are associated with higher log(GNI):

* Lower share of imports in GDP
* Higher share of exports in GDP
* Lower infant mortaility rates
* Being a country from Oceania (size matters?)
* Being an urbanized country

To check our model assumptions, we will look at residual plots.

myResPlots2(fin, label="Log GNI")



Overall, the model assumptions seem satisfied:

* Normally distributed residuals (or very close to it)
* No evidence of heteroskedasticity
* Two possible outliers
* No extremely influential points
* No discernable pattern in the residual plot

## Conclusion

Determining the factors for economic growth has always been an important yet complicated question. In this analysis, for our final model, we find that higher GNI is driven by lower imports, higher exports, lower infant mortality, and urbanization. There’s also an added effect of being from Oceania. This suggests that nations that want to develop should incentivize exports, investment in health care (infant mortality could be a proxy for overall health), and urbanization. Of course, there could be many more factors including whether or not citizens have political rights or civil liberties, but what we have right now is a strong start.