GDP Per Capita Across OECD Countries

Marie Eberlein, Yabing Feng, Samriddh Gupta, and Yueyang Liu

12/9/2019

Executive Summary

Our analysis considered the relationship between GDP per capita and growth in GDP, education, the likelihood of protests, industries, life expectancy, and net exports in 40 countries. Our analyses suggest that GDP is associated with all of these factors except for net exports. Countries with higher GDP tend to see lower growth, fewer protests, higher life expectancy, and a greater proportion of economic value added from professional, scientific, technical, and administrative fields. However, our analyses were constrained by the small size of the data set, variables that were not available for all countries or in all years, and departures from normality in the data.

Introduction

The Organization for Economic Cooperation and Development (OECD) tracks detailed economic and demographic metrics for its member countries, as well as more limited information on non-member countries that are major contributors to the world economy. One of these metrics is Gross Domestic Product per capita, one of the predominant variables used to evaluate the economic health of a country. We used the data set to better understand how GDP per capita is related to some of the other variables in the data set, exploring relationships with growth, education, industry, life expectancy, and trade. We also created a separate data set showing whether a protest occurred in the countries and looked at whether the probability of a protest occurring is related to GDP.

About the Data

The data set includes information for 40 countries plus aggregate numbers for all OECD countries, the European Union, and the Euro area. Data are presented for 2006 to 2014, but not all metrics are available for each year. There are 184 variables in the data set.

Literature Review

We used the results of existing studies on GDP to inform our analyses. Barro (2001) explored the association between growth in GDP and human capital accumulation, including educational attainment and quality of education, for 100 countries from 1960 to 1990. They found a small positive association between GDP per capita and average number of years of secondary school for men and a larger association between GDP per capita and test scores (used as a proxy for education quality). News coverage about protests in Chile has attributed it to high levels of income inequality and suggested that more broad-spread growth in GDP would prevent protests (Bunyan, 2019). Turečková, K. and S. Martinát (2015) found that a greater proportion of the economies of developed countries are in highly skilled fields requiring technology. Leung and Wang (2010) found that income disparities and illiteracy rates were negatively associated with life expectancy. GDP per capita was positively associated with life expectancy. Gorman (2003) found a positive relationship for countries between net exports and higher GDP.

Initial Hypotheses

1. Growth in GDP per capita will be negatively associated with starting GDP and positively associated with investments in education and test scores.
2. Protests will be more likely in countries with countries with faster growth in GDP and lower income inequality.
3. Countries with higher GDP per capita will have a larger proportion of value added from professional, scientific, technical, and administrative fields.
4. Life expectancy will be higher in countries with higher GDP per capita, more physicians per 1000 people, and higher health expenditures.
5. Countries with higher GDP will have lower imports of goods and higher exports of goods.

Exploratory Data Analysis

Growth in GDP

To understand what factors affect growth in GDP per capita, we calculated the percent change in GDP from 2006 to 2014. This growth variable was used as the response variable in a multiple linear regression model. Initially, the model was fit using starting GDP (in 2006), investments in primary education, investments

in secondary education, investments in tertiary education, girls' PISA reading scores, boys' PISA reading scores, girls' PISA math scores, and boys' PISA math scores. Because there is a large spread of GDP values, we used a log transform of GDP in the model.

The linear model tests the hypothesis that the coefficients for all the response variables are 0. For the linear model $Y = \beta_0 + \beta_1 * X_1 + ... + \beta_8 * X_8$:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

H_1: At least one inequality

The initial fit of the model that only when all of the variables are included, there is very strong evidence that the log of GDP is negatively associated with growth in GDP (p-value is very small). Investments in primary education (coded in the data as EXEDULV_T1A) has a p-value of 0.0571, so the observed relationship may not be due to the natural variation of the data, but the evidence is not as strong. All other variables have higher p-values. (The summary of the initial regression model is in Appendix B.)

To better understand the relationship, we removed the least significant variables. We also checked the regression diagnostics and took the log transform of growth to address non-constant variance. The final model includes only the starting GDP and investments in primary education.

```
##
## Call:
## lm(formula = log(growth) ~ EXEDULV_T1A + log(GDP2006), data = growth_by_ed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9650 -0.1183  0.1524  0.2514  0.5737
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.192e+01  3.132e+00   3.807 0.000601 ***
## EXEDULV_T1A   1.144e-04  4.263e-05   2.684 0.011435 *
## log(GDP2006) -1.397e+00  3.339e-01  -4.182 0.000209 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3908 on 32 degrees of freedom
```

```
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4109, Adjusted R-squared:  0.374
## F-statistic: 11.16 on 2 and 32 DF,  p-value: 0.0002106
```

Predicting Protests

To understand what factors are related to protests occurring in a particular country, we first created a data set indicating whether there had been a protest in each country since 2006. For the purpose of this analysis, we looked for sustained, wide-spread protests with a specific political goal. Of the 40 countries in the data set, protests occurred in 23.

We used logistic regression to model the probability that a protest would occur. To check the model, we kept a few observations out of the training data set. Logistic regression fits coefficients for the following model, where β is a vector of coefficients and X is a matrix of predictors.

$$P(protest) = e^{\beta * X}/(1 + e^{\beta * X})$$

We test $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

The model is initially fit using growth in GDP, GDP in 2006, investments in secondary education, number of immigrants, income inequality (Gini index), unemployment levels, and imports and exports. We used step-wise selection to identify which predictors should remain in the model. Only growth and 2006 GDP were retained in the final model.

The final model is:

$$P(protest) = e^{97.215 - 16.812 * growth - 9.053 * log(GDP2006)}/(1 + e^{97.215 - 16.812 * growth - 9.053 * log(GDP2006)})$$

We see there is a negative relationship between GDP and growth and the likelihood that a protest will occur. A summary of the model can be found in Appendix B.

Assuming that a protest did not occur in a country where the predicted probability is less than 50%, when we use the final model to predict whether there were protests in the testing data that was withheld, it correctly predicts all of the countries.

```
mean(yhat==protest_predictors$Protest[Z])
```

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

Value Added

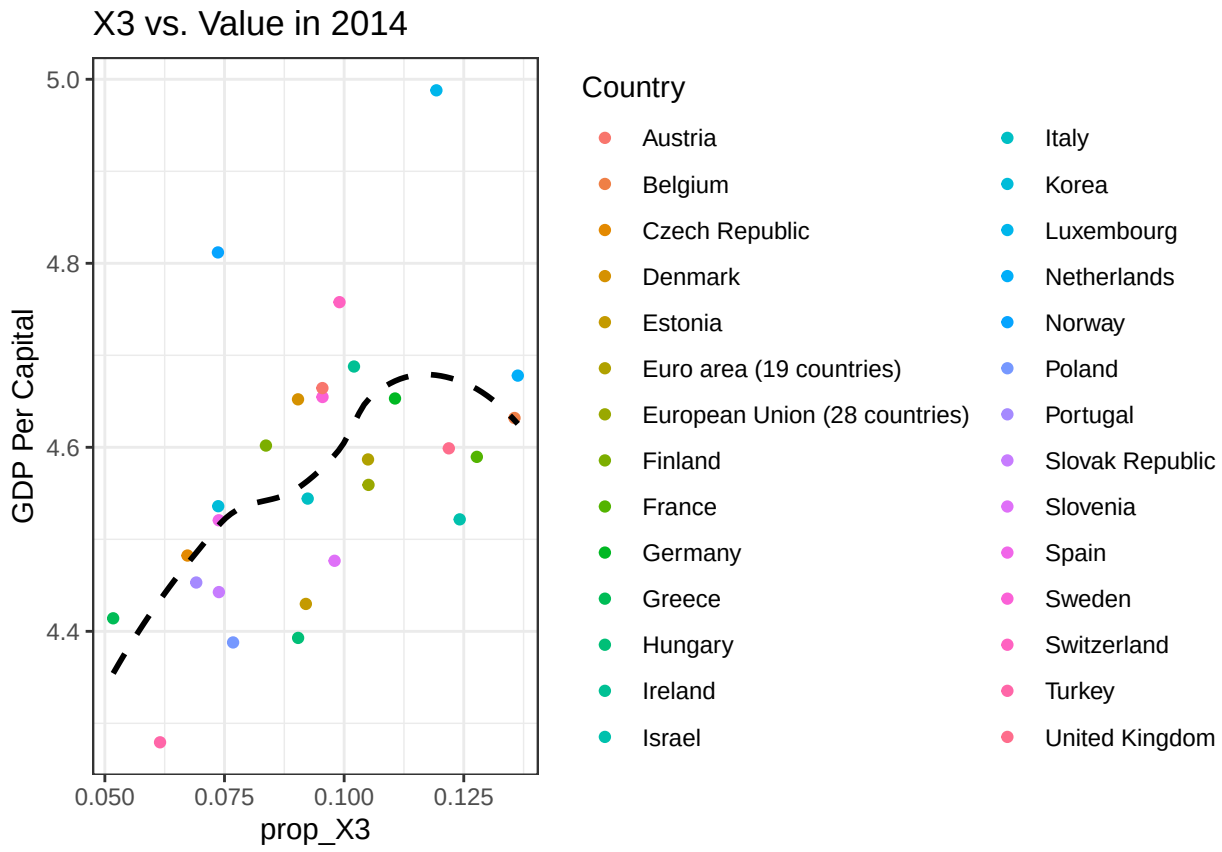
From 2006 to 2014, GDP per capita grew in each country, except Greece whose data was distorted by the "Greek government-debt crisis". We wanted to test the hypothesis that there is a positive relation between GDP per capita and the proportion of high technical activities in the economy. By observing the data, technical terms are shown in the "Value Added" factor and described as "Value added in professional; scientific; technical; administration and support services activities(x3)". And we will use the data in 2014 to avoid the effect of the financial crisis happened in 2008. There are 10 different terms in the category of value added. The response variable will be GDP per capita (log(Value)), and the proportion of the high technical activities, prop_X3 (calculated from X3 divided by total value added) would be independent variable.
Value added in the data measured as the value of output minus the value of intermediate consumption. It reflects the value generated by producing goods and services. Also represents the income available for the contributions of labour and capital to the production process.
Here is the model: $log(Y) = \beta_0 + \beta_3 X_{3i} + \varepsilon_i$
Here is the hypothesis: H_0: $\beta_3 = 0$ vs H_1: $\beta_3 \neq 0$
A summary of the regression model can be found in Appendix B. The model would be: $\hat{Y} = 9.8195 + 7.4754 X_3$
By looking at the result, there is statistical significance with p value equal to 0.0063, which has strong evidence to say there is a correlation between technical activities and GDP per capita. But the R^2 is 0.2573, suggesting that technical activities are not the only thing that can influence GDP per capita of a country, and the correlation of 0.4214 also supports this point.

```
cor(combine_oecd[,c(3,6)])
```

```
##             Value   prop_X3
## Value   1.0000000 0.4214163
## prop_X3 0.4214163 1.0000000
```

A scatter diagram of the data is below. (Note that we used the smooth line instead of using the straight line to give more story of the data behind.) There are 28 countries in this data. One blue dots on the top: Luxembourg (log(97272.57) = 4.99). This country has a highest GDP per capita in 2014 that is likely to drag the analysis into a biased result. Thus we will exclude Luxembourg in the regression model.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



After removing Luxembourg, the result we had followed our expectation. The new regression line:$\hat{Y} = 9.9316 + 5.9706X_3$, the significant level decrease to 0.0125, R^2 decrease to 0.2246, and the correlation decrease to 0.3989. However, there is still evidence supporting our hypothesis.

Life expectancy

To figure out the relationship between GDP per capita and life expectancy, we first select some the variables may be related to life expectancy for 46 countries from 2006 to 2014. These are:

1.GDP

2.Year/Country

3.Number of practising physicians in 1000 people

4.Population

5.Expenditure on health (Public/Private)

At first we used a linear model to fit the relationship between the GDP and life expectancy and update the model based on the residual plot. The updated model looks like:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1$$

- β_0 The intercept of the mean line.

- β_1 The effect on Y when change GDP given the other variables in the model.

- β_2 The effect on Y when change GDP^2 given the other variables in the model.

- x_1 GDP value

- x_2 GDP^2 value

```
## # A tibble: 3 x 3
##   term        estimate   p.value
##   <chr>          <dbl>     <dbl>
## 1 (Intercept)  6.34e+1 1.19e-185
## 2 GDP          6.73e-4 2.05e- 40
## 3 GDP2        -5.78e-9 1.02e- 24
```

we also use the step function to fit the model with all variables,the final model is: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_1$

- β_1 The effect on Y when change private expenditure given the other variables in the model.

- β_2 The effect on Y when change public expenditure given the other variables in the model.

- β_3 The effect on Y when change GDP given the other variables in the model.

- β_4 The effect on Y when change GDP^2 given the other variables in the model.

- x_1 private expenditur

- x_2 public expenditure

- x_3 GDP value

- x_4 GDP^2 value

```
## # A tibble: 5 x 3
##   term        estimate   p.value
##   <chr>          <dbl>     <dbl>
## 1 (Intercept)  6.21e+1 3.43e-163
## 2 Pri         -1.43e-3 9.66e-  5
## 3 Pub         -8.12e-4 1.18e-  1
## 4 GDP          8.31e-4 4.87e- 24
```

```
## 5 GDP2        -6.62e-9 6.40e- 26
```

We still see heavy tails in qqplot but the residual plot looks better. We believe the errors come from data missing since we had remove all the lines that contain "NA".

Also, there are too many factors in the data set which, contributes to the overfitting. To improve this we could combine different countries with similar GDP. We could also include additional variables like education level and environmental conditions.

Net Exports

Imports and exports are an important part of every country's revenue. Not every area in the world has all the stuff available and the money exchange in these exports can sometimes be the only revenue that a country has. For this project we focus on how GDP is related to the imports, exports, and net exports of the country. Countries' net exports can be used to calculate the expenditure power of that country. It can also be used to calculate whether a country is developed or not. For this project, we assumed that the countries with a higher GDP are developed countries.

To see if higher net exports were associated with higher GDP, we tested the following:

H_0: We obtain negative results

H_1: We get a positive relation of GDP and Net Exports (Exports of Goods - Import of Goods)

Summary statistics for GDP, imports, and exports can be seen in Appendix B.

The data we expected would show that imports decrease as the GDP increases, but that was not the trend that we found. In this particular data set, the imports of goods increase as the GDP increases. For the exports of goods, the data shows that as the GDP increases the export of goods also increase, which we expected. Since both the imports and exports increase as the GDP grows and both of them have a similar mean, the net exports came out to be close to zero. The following shows a summary of the data.

```
##    NetExp2010          NetExp2014
##  Min.   :-689.3876   Min.   :-726.2977
##  1st Qu.: -11.7712   1st Qu.:  -9.2347
##  Median :  -0.3232   Median :  -0.3206
##  Mean   : -18.0299   Mean   : -18.4733
##  3rd Qu.:  15.4468   3rd Qu.:  16.1646
##  Max.   : 204.2796   Max.   : 287.2999
```

Plotting the graph to get the better understanding of the relationship between the net exports and the GDP gives results as shown in the graph.As we can see from the graph, there was a negative trend in the net exports VS GDP. The trend is similar for both years. We also see a huge outlier in both cases. Both outliers are the United States of America. Germany has the highest net exports for both years. For the final analysis test, I used the correlation between the two variables (net exports and GDP for both years) to verify the result and to get the conclusion. The result of the correlation is stated below.

```
##                GDP2010  NetExp2010
## GDP2010     1.00000000 -0.07032793
## NetExp2010 -0.07032793  1.00000000

##                GDP2014  NetExp2014
## GDP2014     1.00000000 -0.06614101
## NetExp2014 -0.06614101  1.00000000
```

Data-Driven Hypotheses

Based on the results of our analysis, we developed the following revised hypotheses:

1. Growth in GDP per capita is negatively associated with starting GDP and positively associated with investments in primary education.
2. Countries with higher GDP per capita and higher growth in GDP are less likely to have protests.
3. Countries with higher GDP per capita will have a larger proportion of value added from professional, scientific, technical, and administrative fields. (Unchanged)
4. GDP per capita some other variables are associated with life expectancy.
5. There is not a significant relationship between net exports and GDP.

Discussion

We found a highly significant relationship between growth in GDP, starting GDP, and investments in primary education. However, checking the regression diagnostics, there appear to be departures from normality, which we confirmed with a Shapiro-Wilk test. This raises some concerns about the validity of our conclusions from the model. The low p-value of 0.00076 (seen in Appendix B) means we reject the null hypothesis that the residuals are normally distributed.

There were also concerns about the reliability of the model predicting protests. We know from the linear model that GDP per capita and growth in GDP are collinear, which would make the model less stable. The model fitting process is also quite sensitive to changes in the initial variables that are included and the training data set that is selected. Small changes to either will prevent the model from converging when fitting the coefficients.

Looking at how GDP per capita is related to the proportion of value added, from the data we observe that the proportion of the high technical activities ranges from 0.0517 (Greece, log(25950.39) = 4.41) to 0.1363 (Netherlands, log(47634.76) = 4.68). 0.1 is the cut-off of the proportion of the high technical activities. Since there are 10 categories, any country with a value smaller than 0.1 is considered to have a lower proportion. Most of the countries have lower proportions. There are 10 countries had higher proportion: Ireland (0.1021), Euro area (19 countries, 0.1050), European Union (28 countries, 0.1051), Germany (0.1106), Luxembourg (0.1193), United Kingdom (0.1219), Israel (0.1242), France(0.1277), Belgium (0.1356), and the Netherlands (0.1363). In the graph above, we saw that the dashed line started to decrease from its peak, around 0.113, which is slightly larger than the cut off (0.1). The increasing trend slows down at the proportion range of 0.075 to 0.0875 and recovers after 0.0875, which continues until 0.113. The slope becomes smaller at the proportion range of 0.075 to 0.0875 and starts to increase after and stop around 0.113.

Overall, the hypothesis, there is a positive relation between GDP per capita and the proportion of high technical activities in the economy is reasonable within a certain range. However, when the proportion of the technical activities exceeds 0.113, the GDP per capita will depict a negative track.

In the ggpair plot we can see a positive relationship between GDP per capita and life expectancy. We also tried include more related variables. The final liner model's R-squared value is 0.6025 and p-value is 2.2e-16, so we can say GDP per capita and some other variables are associated with life expectancy.

For the relationship between GDP and net exports, we see that the correlation value is negative and the value of r is close to zero. Based on this, we concluded that there is no relation between the net exports and GDP in this data set. One of the reasons for this relationship may be that the US was a huge outlier. For future analysis, we could scale the net exports by population to see if that affects the relationship.

Our analyses suggest that there is reason for the continued prominence of GDP in evaluating the economic health of a country. While there are challenges with the data that raise questions about the validity of some analyses, we do see strong relationships between GDP per capita and health, education, types of jobs, and living conditions. The analyses could be further improved by seeing if these trends also hold true in non-OECD countries that were not included in the data set, as well as using time series analysis to see if the relationships change over time.

Appendix A: References

Barro, R.J. (2001). Education and Economic Growth. *OECD.* http://search.oecd.org/education/innovation-education/1825455.pdf

Bunyan, R. (2019). 18 Killed as Hundreds of Thousands of Protestors Take to the Streets in Chile. Here's What to Know. *Time.* https://time.com/5710268/chile-protests/

Gorman, T. (2003). The Complete Idiot's Guide to Economics. https://www.infoplease.com/homework-help/social-studies/gdp-and-players-three-imports-and-exports

Leung, M. C., & Wang, Y. (2010). Endogenous health care, life expectancy and economic growth. *Pacific Economic Review*, *15*(1), 11-31.

Turečková, K. and S. Martinát. (2015). Quaternary sector and extended sectoral structure of the economy in the selected European countries. *Working Paper in Interdisciplinary Economics and Business Research no. 10. Silesian University in Opava, School of Business Administration in Karviná.* http://www.iivopf.cz/images/Working_papers/WPIEBS_10_Tureckova_Martinat.pdf

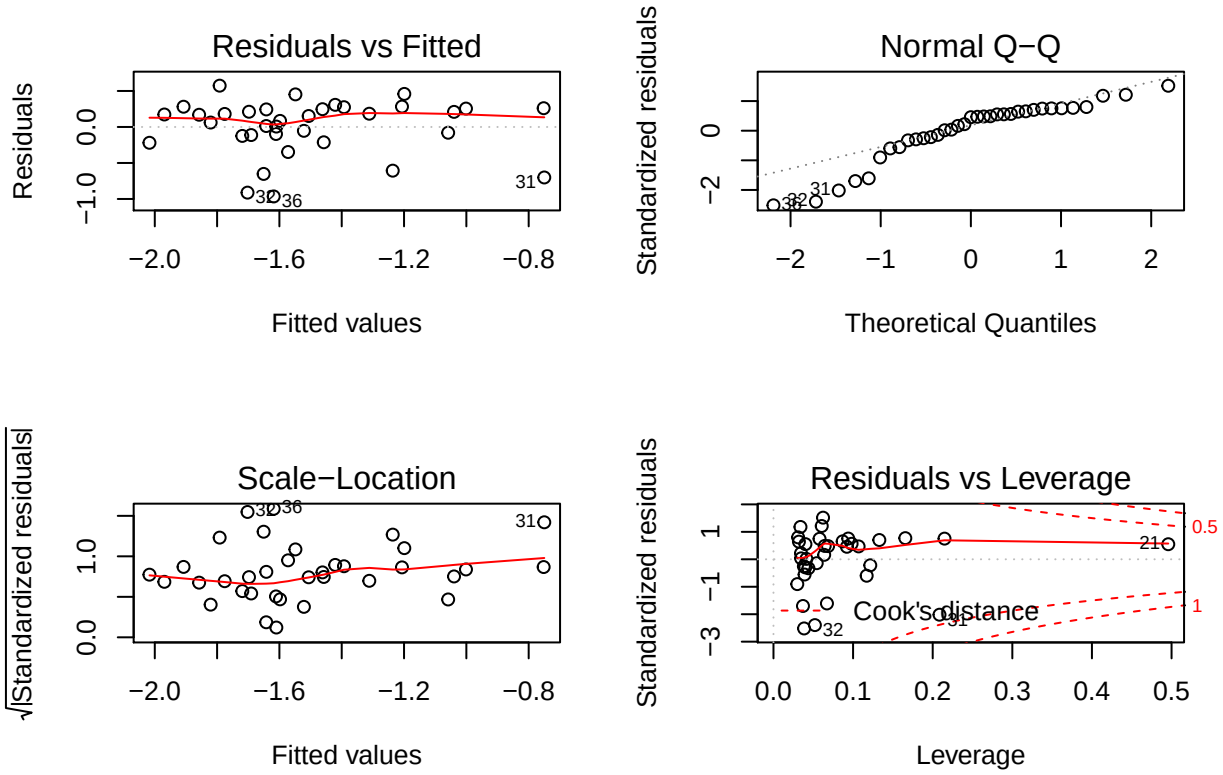
Appendix B: Additional R Code

In the initial model predicting growth in GDP, most of the variables in the model were not significant.

```
##
## Call:
## lm(formula = growth ~ EDUTEREXPND_T1C + EXEDULV_T1A + EXEDULV_T1D +
##     log(GDP2006) + INCINEQUAL_T1A + PISA_T1K + PISA_T1E + PISA_T1G +
##     PISA_T1I, data = growth_by_ed)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.144435 -0.036137  0.005163  0.050038  0.123290
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.280e+00  1.245e+00   5.045 4.73e-05 ***
## EDUTEREXPND_T1C  3.107e-06  5.730e-06   0.542   0.5931
```

```
## EXEDULV_T1A      2.564e-05  1.277e-05   2.008   0.0571 .
## EXEDULV_T1D      1.965e-05  1.544e-05   1.273   0.2162
## log(GDP2006)    -6.308e-01  1.103e-01  -5.718 9.44e-06 ***
## INCINEQUAL_T1A  -5.134e-01  5.329e-01  -0.964   0.3457
## PISA_T1K         7.985e-04  2.556e-03   0.312   0.7577
## PISA_T1E        -9.762e-05  3.483e-03  -0.028   0.9779
## PISA_T1G        -1.595e-03  3.434e-03  -0.464   0.6470
## PISA_T1I         1.266e-03  2.967e-03   0.427   0.6738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08101 on 22 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.707,  Adjusted R-squared:  0.5872
## F-statistic: 5.899 on 9 and 22 DF,  p-value: 0.000331
```

Diagnostic plots for the growth linear regression model and a test for non-constant variance. After taking the log transform of growth, there is not sufficient evidence to believe the variance is not constant. However, a Shapiro-Wilk test shows there is evidence the residuals are not normally distributed.




```
## Warning in log(growth): NaNs produced
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1643909, Df = 1, p = 0.68515
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(reg_prim)
## W = 0.87191, p-value = 0.0007552
```

Summary of the logistic regression model predicting the probability of a protest occurring:

```
##
## Call:
```

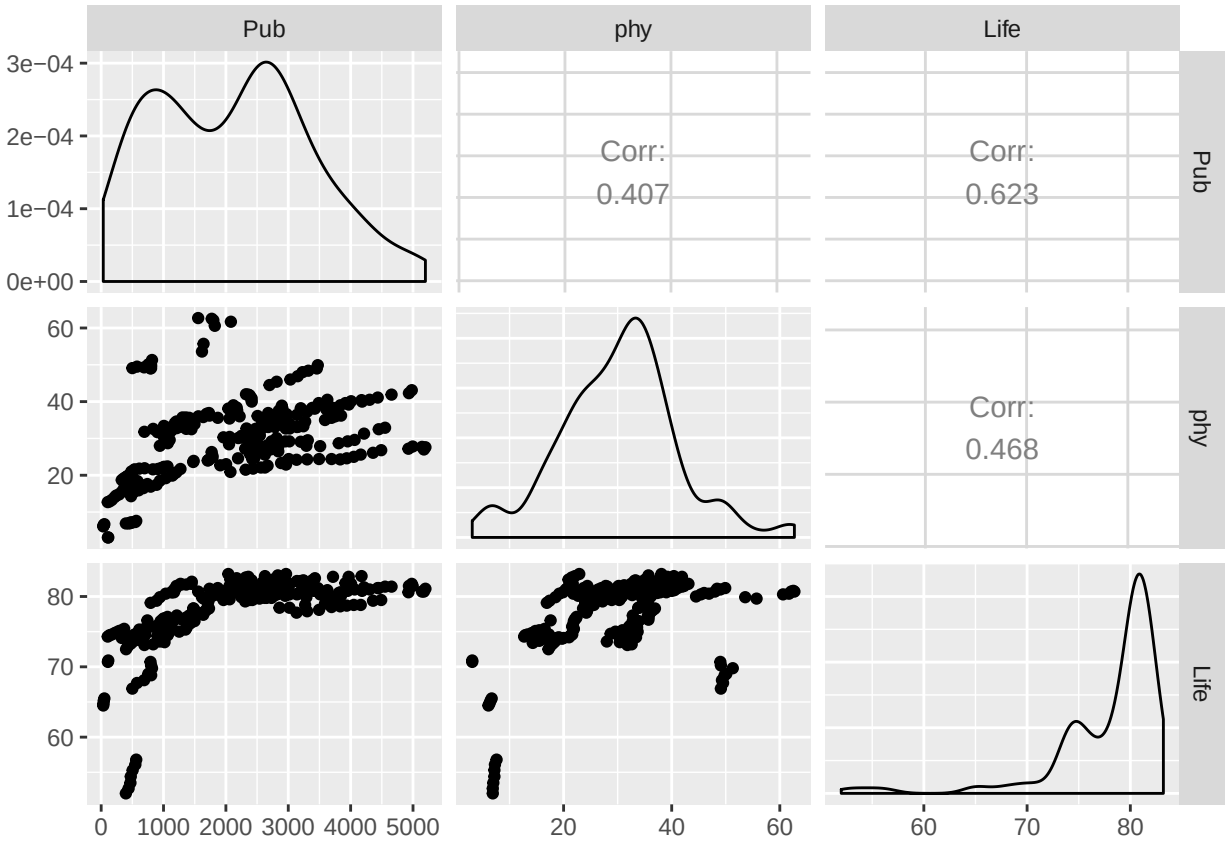
```
## glm(formula = Protest ~ growth + log(GDP2006), family = "binomial",
##     data = protest_predictors[-Z, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9207  -0.6305  -0.2690   0.4639   1.7907
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    97.215     48.430   2.007   0.0447 *
## growth        -16.812     11.501  -1.462   0.1438
## log(GDP2006)   -9.053      4.526  -2.000   0.0455 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21.930  on 15  degrees of freedom
## Residual deviance: 12.984  on 13  degrees of freedom
## AIC: 18.984
##
## Number of Fisher Scoring iterations: 5
```

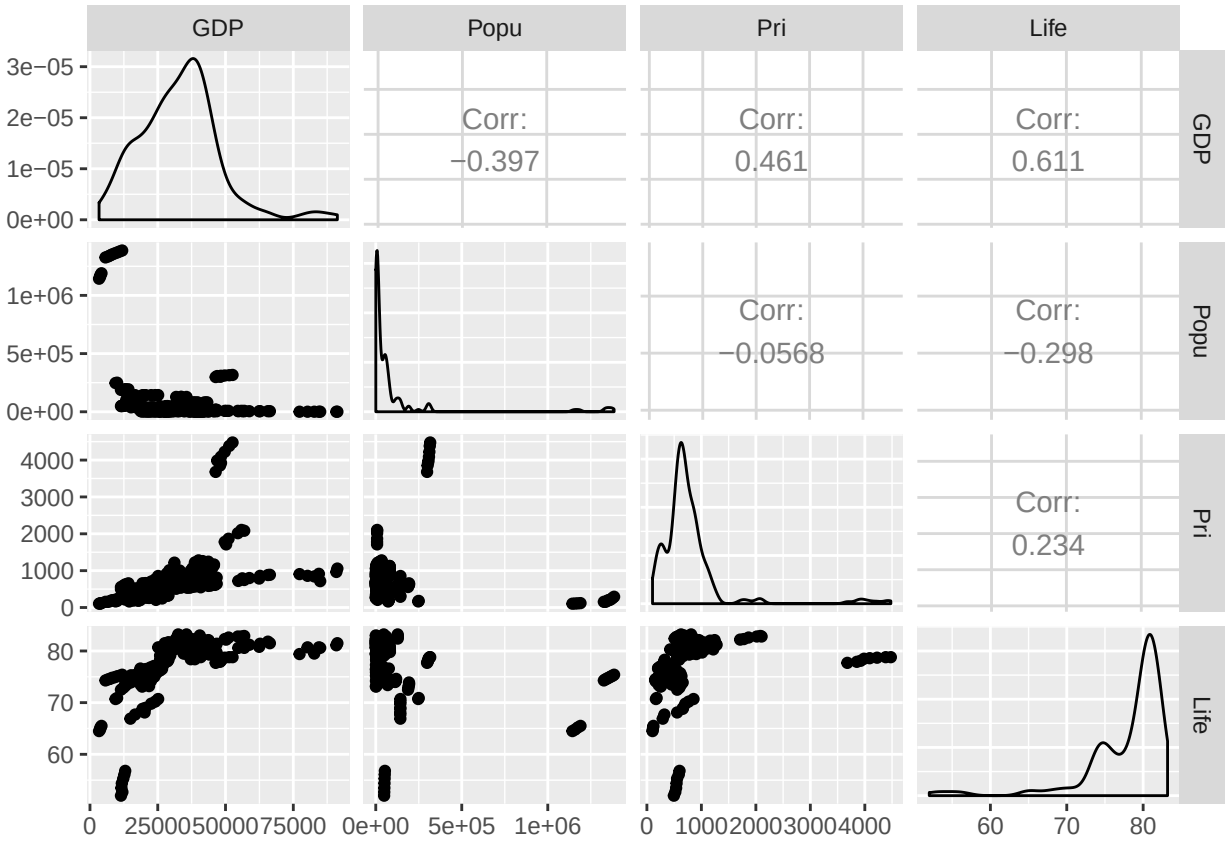
Summary of the linear regression model comparing GDP per capita and the proportion of value added in professional, scientific, technical, and adminstrative fields.

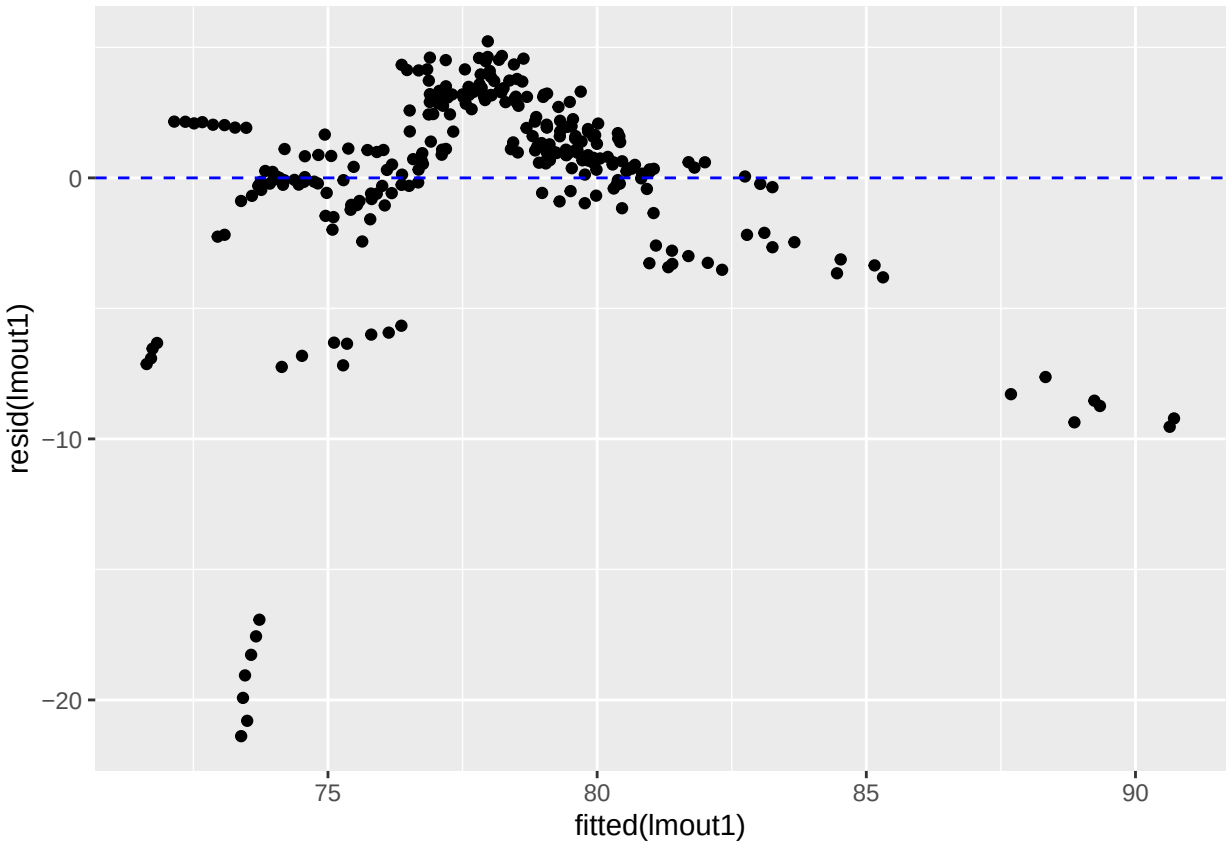
```
##
## Call:
## lm(formula = log(Value) ~ prop_X3, data = combine_oecd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42582 -0.17758 -0.04471  0.15970  0.77409
```

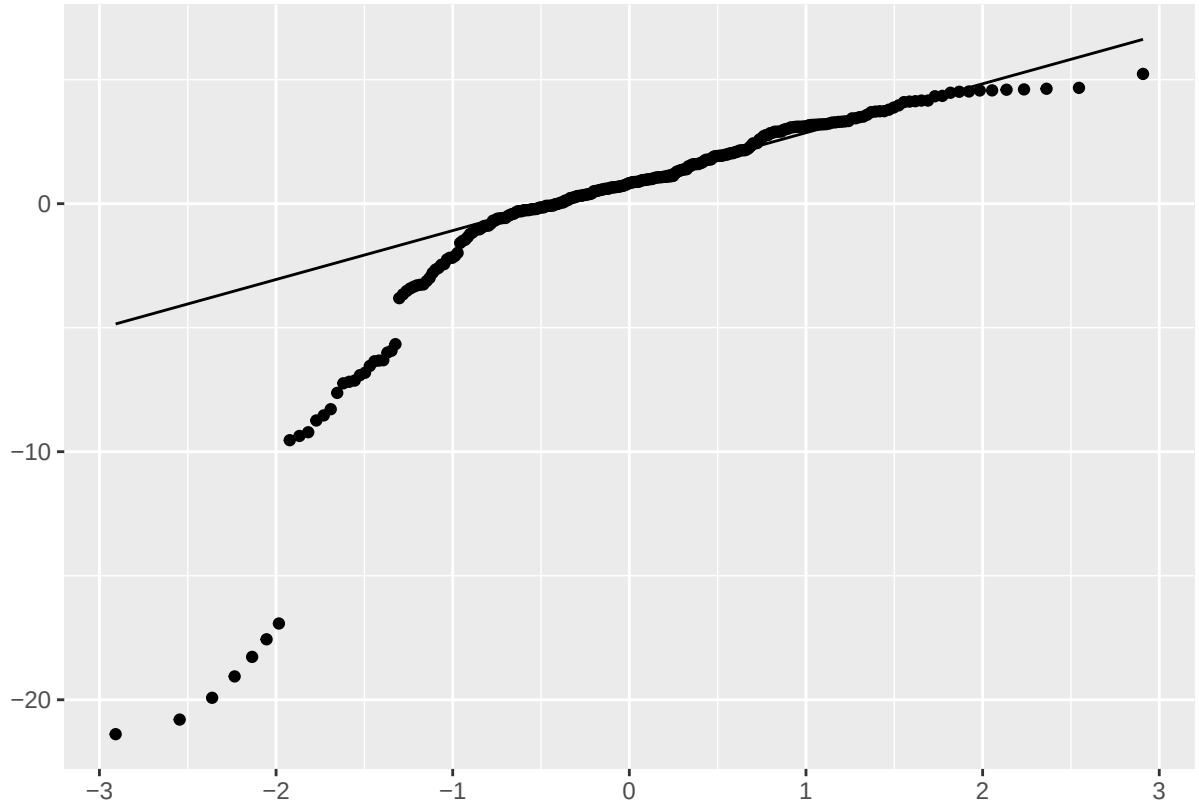
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8195     0.2441  40.224  < 2e-16 ***
## prop_X3       7.4754     2.5147   2.973  0.00629 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2961 on 26 degrees of freedom
## Multiple R-squared:  0.2537, Adjusted R-squared:  0.225
## F-statistic: 8.837 on 1 and 26 DF,  p-value: 0.006288
```

Prior to exploring the relationship between life expectancy and GDP, we use the GGally package to draw plots and make initial explorations of the relationship between variables.




GDP
Popu
Pri
Life
Corr:
−0.397
Corr:
0.461
Corr:
0.611
Corr:
−0.0568
Corr:
−0.298
Corr:
0.234
3e−05
2e−05
1e−05
0e+00
1e+06
5e+05
0e+00
4000
3000
2000
1000
0
80
70
60
0
25000
50000
75000
0e+00
5e+05
1e+06
0
1000
2000
3000
4000
60
70
80


0
−10
−20
resid(lmout1)
75
80
85
90
fitted(lmout1)



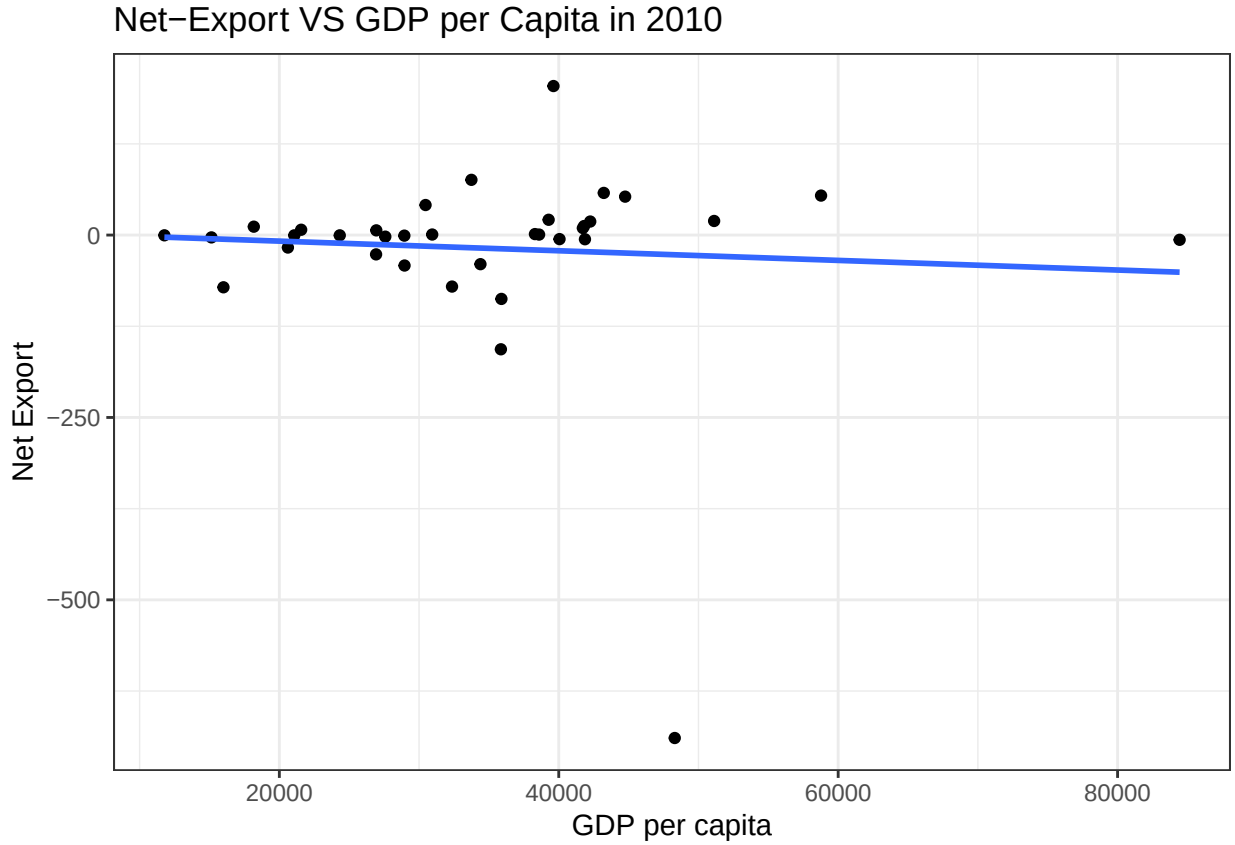
We see a curve in the residual plot, to fit the assumption well, we add a squared GDP value

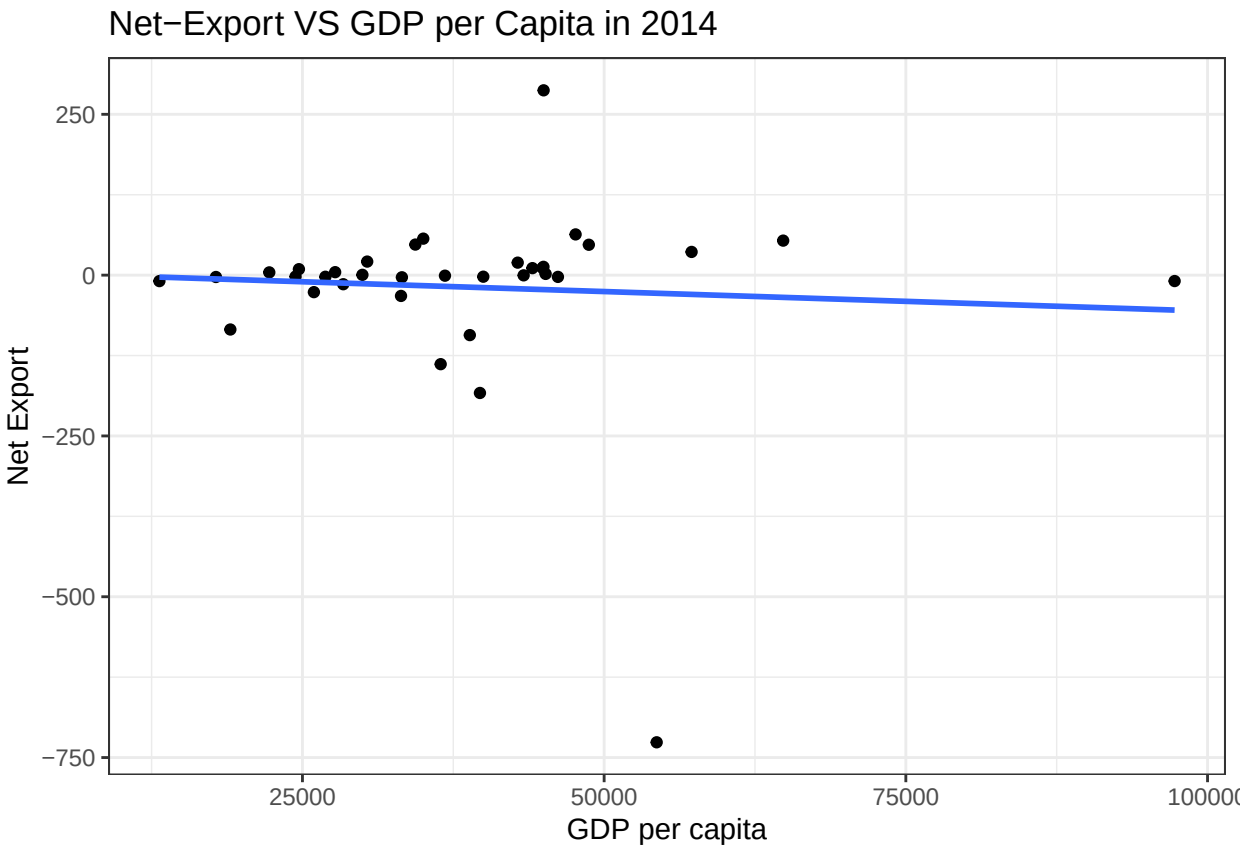
Summary statistics of GDP, imports, and exports:

```
##     GDP2010         GDP2014
##  Min.   :11772   Min.   :13146
##  1st Qu.:26933   1st Qu.:28047
##  Median :34396   Median :36810
##  Mean   :34737   Mean   :38422
##  3rd Qu.:41770   3rd Qu.:44978
##  Max.   :84440   Max.   :97273

##     IMP2010            IMP2014
##  Min.   :   3.914   Min.   :   5.372
##  1st Qu.:  63.855   1st Qu.:  74.559
##  Median : 148.788   Median : 162.452
##  Mean   : 276.396   Mean   : 325.569
```

```
##  3rd Qu.: 391.100   3rd Qu.: 457.386
##  Max.   :1966.497   Max.   :2346.041

##     EXP2010            EXP2014
##  Min.   :   4.603   Min.   :   5.051
##  1st Qu.:  67.061   1st Qu.:  75.489
##  Median : 132.142   Median : 164.344
##  Mean   : 258.366   Mean   : 307.096
##  3rd Qu.: 396.205   3rd Qu.: 472.534
##  Max.   :1277.109   Max.   :1619.743
```

Comparing GDP per capita and net exports, we see a negative relationship both in 2010 and 2014.




Net−Export VS GDP per Capita in 2014
Net Export
250
0
−250
−500
−750
25000
50000
75000
GDP per capita