# GDP Per Capita Across OECD Countries

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

12/2/2019

# 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 is related to some of the other variables in the data set, exploring relationships with trade, industry, education, health, and growth.

# 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.

# Research Questions

- What factors affect growth in GDP?
- How does the proportion of people in different professional fields vary for countries with higher and lower GDP?
- How is GDP related to health expenditures?
- What is the relationship between GDP and the distribution of financial versus non-financial assets?
- How does relate GDP to net exports?

# Growth in GDP

**Hypothesis:** There will be a negative relationship with growth in GDP and starting GDP and a positive relationship with growth in GDP and investments in education. For the relationship $growth = \beta_0 + \beta_1 * GDP + \beta_2 * investment$,

$H_0$: $\beta_0 = \beta_1 = \beta_2 = 0$
$H_1$: At least one inequality

Using stepwise selection, we find a significant relationship between growth in GDP and starting GDP as well as between growth in GDP and investments in primary education. There does not appear to be a strong relationship between growth in GDP and investments in secondary education, tertiary education, or test scores.
$log(growth) = 11.86 + 1.14e-4 * education - 1.39 * log(GDP)$

# GDP Growth Relationship

```
##
## Call:
## lm(formula = log(growth) ~ EXEDULV_T1A + log(GDP2006), data =
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9694 -0.1222  0.1546  0.2533  0.5690
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.186e+01  3.186e+00   3.722 0.000787 ***
## EXEDULV_T1A   1.141e-04  4.326e-05   2.637 0.012956 *
## log(GDP2006) -1.390e+00  3.397e-01  -4.091 0.000283 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.3965 on 31 degrees of freedom
## Multiple R-squared:  0.406,  Adjusted R-squared:  0.3677
## F-statistic: 10.59 on 2 and 31 DF,  p-value: 0.0003116
```

# Results of Growth

- Given the recent news about protests in Chile despite strong economic growth there, we wanted to understand if the variables in the data set could be used to predict whether a protest would occur in a particular country.
- To determine this, we coded each country as yes or no for having a sustained, widespread protest with a specific political goal since 2006.
- Using this data set, we used logistic regression and stepwise selection to see if we could predict the protests.
- The initial model included growth in GDP, GDP in 2006, investments in secondary education, number of immigrants, income inequality (Gini index), unemployment levels, and imports and exports. Only growth and 2006 GDP were retained in the final model.

# Predicting Protests

- The resulting model predicts the probability of a protest according to the following: $P(protest) = exp(97.215 - 16.812 * growth - 9.053 * log(GDP))/(1 + exp(97.215 - 16.812 * growth - 9.053 * log(GDP)))$
- This model correctly predicts the results of a small testing set that was withheld when fitting the model.
- However, there is reason to believe the model is unreliable. 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.

# Introduction of data

Focus on the **value added** in difference areas
- here are all the areas:

```
## # A tibble: 10 x 2
##    SUB         Subject
##    <chr>       <chr>
##  1 VALADDAC_T~ Value added in industry; including energy
##  2 VALADDAC_T~ Value added in agriculture; hunting and forestry; fishing
##  3 VALADDAC_T~ Value added in professional; scientific; technical; administrati~
##  4 VALADDAC_T~ Value added in distributive trade; repairs; transport; accommoda~
##  5 VALADDAC_T~ Value added in construction
##  6 VALADDAC_T~ Value added in financial and insurance activities
##  7 VALADDAC_T~ Value added in public administration; defence; education human h~
##  8 VALADDAC_T~ Value added in real estate activities
##  9 VALADDAC_T~ Value added in other services activities
## 10 VALADDAC_T~ Value added in Information and communication
```

# What is Value Added?

- Measured as the value of output minus the value of intermediate consumption.
- Value added reflects the value generated by producing goods and services.
- Value added also represents the income available for the contributions of labour and capital to the production process.

# Hypothesis

There will be a positive relationship in GDP per capital with the fraction of economic value being added by activities of professional; scientific; technical; administration and support services($X_3$).

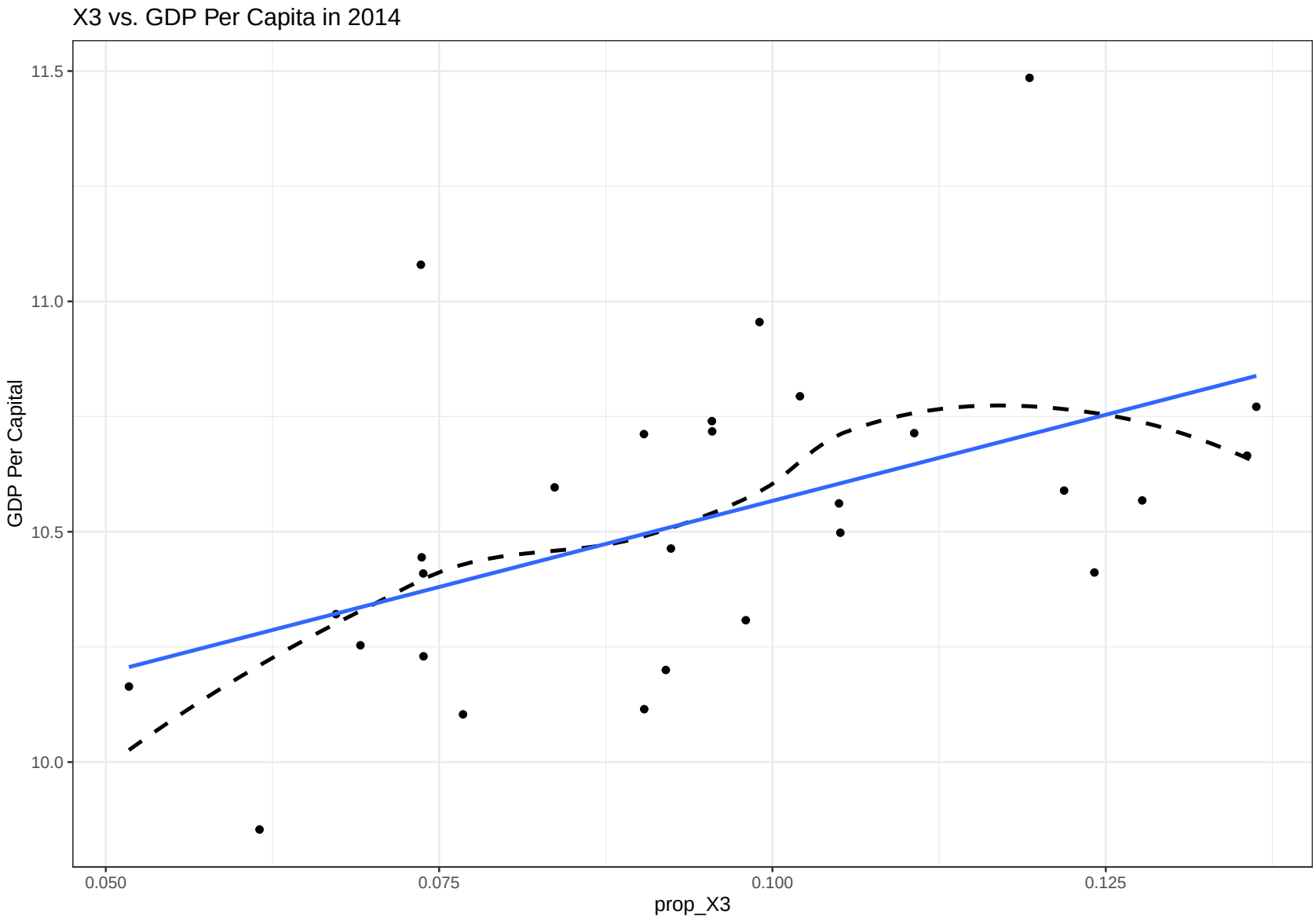
# Model(1):

Model: $log(Y) = \beta_0 + \beta_3 X_{3i} + \varepsilon_i$

$H_0$: $\beta_3 = 0$
$H_1$: $\beta_3 \neq 0$

# Scatter Diagram(1):

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



# Summary of X3 (1):

```
##
## 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 ' '
##
## 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
```

# Model(2):

Model: $log(Y) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} ... + \beta_{10} X_{10i} + \varepsilon_i$

$H_0$: $\beta_1 = \beta_2 ... \beta_{10} = 0$

$H_1$: At least one inequality

# Relationship of GDP and value added in 10 different areas(2):

- Count all of the 10 variables is the setup to get some sense how those variables worked
- Overall the slope of X3 decrease from 7.4754 to 0.2890, but p-value changed to 0.933283, which was not expected. By looking at the p-value, X6 & X7 stand out

```
##
## Call:
## lm(formula = log(Value) ~ prop_X1 + prop_X2 + prop_X3 + prop_X4 +
##     prop_X5 + prop_X6 + prop_X7 + prop_X8 + prop_X9 + prop_X10,
##     data = combine_oecd)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.198768 -0.080900 -0.008876  0.050398  0.294879
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2660     2.0818   4.451 0.000309 ***
## prop_X1       1.4310     2.3515   0.609 0.550433
## prop_X2      -5.0788     4.2773  -1.187 0.250513
## prop_X3       0.2890     3.4043   0.085 0.933283
## prop_X4      -0.7398     2.0552  -0.360 0.723049
## prop_X5       1.7180     3.3802   0.508 0.617449
## prop_X6       5.7309     2.5420   2.254 0.036864 *
## prop_X7       4.0347     2.2676   1.779 0.092088 .
## prop_X8      -0.2326     2.6260  -0.089 0.930405
## prop_X9       1.0128     3.1292   0.324 0.749915
## prop_X10          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1591 on 18 degrees of freedom
## Multiple R-squared:  0.8507, Adjusted R-squared:  0.7761
## F-statistic:  11.4 on 9 and 18 DF,  p-value: 9.012e-06
```

# Model(3):

- deletes some notices of having too many independent variable, only focuses on the areas that we are interested in:

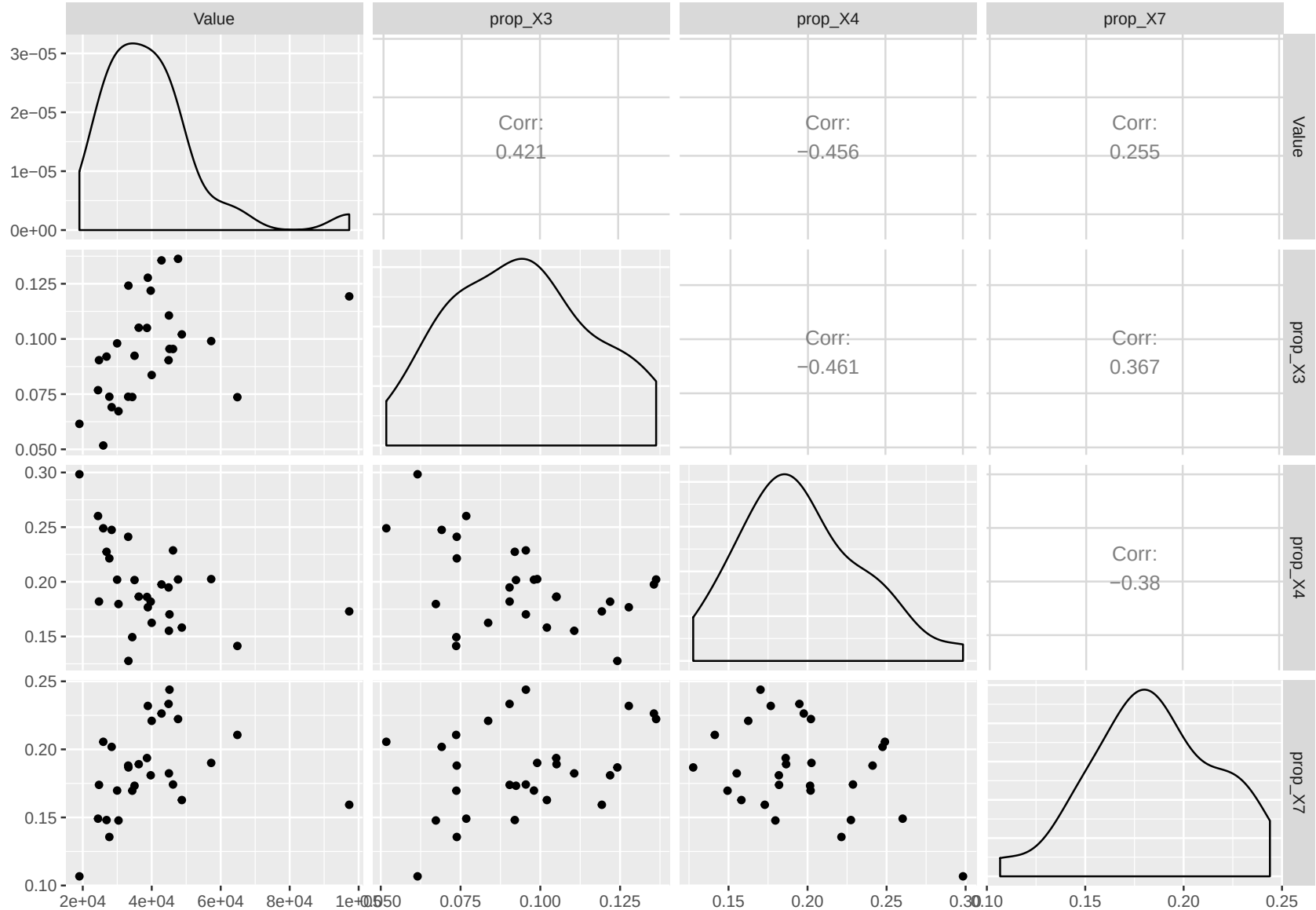
```
##
## Call:
## lm(formula = log(Value) ~ prop_X3 + prop_X4 + prop_X6 + prop_X7,
##     data = combine_oecd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29081 -0.06768 -0.00751  0.03464  0.41034
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.0842450  0.3715275  27.143  < 2e-16 ***
## prop_X3     -0.0003309  1.7996812   0.000  0.99985
## prop_X4     -2.6665414  0.9824891  -2.714  0.01238 *
## prop_X6      5.1870387  0.8425884   6.156 2.79e-06 ***
## prop_X7      3.5399485  1.1602207   3.051  0.00567 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1707 on 23 degrees of freedom
## Multiple R-squared:  0.7804, Adjusted R-squared:  0.7422
## F-statistic: 20.43 on 4 and 23 DF,  p-value: 2.676e-07
```

# Model Improvment (4):

- Focus on the non-financial parts that we are interested in
- There are some evidence of the relationship between X3 and GDP per capital. A unit change on the porportion of X3 will cause 3.9577 change in GDP per capital on average

```
##
## Call:
## lm(formula = log(Value) ~ prop_X3 + prop_X4 + prop_X7, data = combine_oecd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45817 -0.09786 -0.05752  0.04966  0.83030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.4598     0.5838  17.917 2.14e-15 ***
## prop_X3       3.9577     2.6776   1.478   0.1524
## prop_X4      -3.2388     1.5580  -2.079   0.0485 *
## prop_X7       1.7784     1.7910   0.993   0.3306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.272 on 24 degrees of freedom
## Multiple R-squared:  0.4186, Adjusted R-squared:  0.3459
## F-statistic:  5.76 on 3 and 24 DF,  p-value: 0.004089
```

# Correlation between those variables:



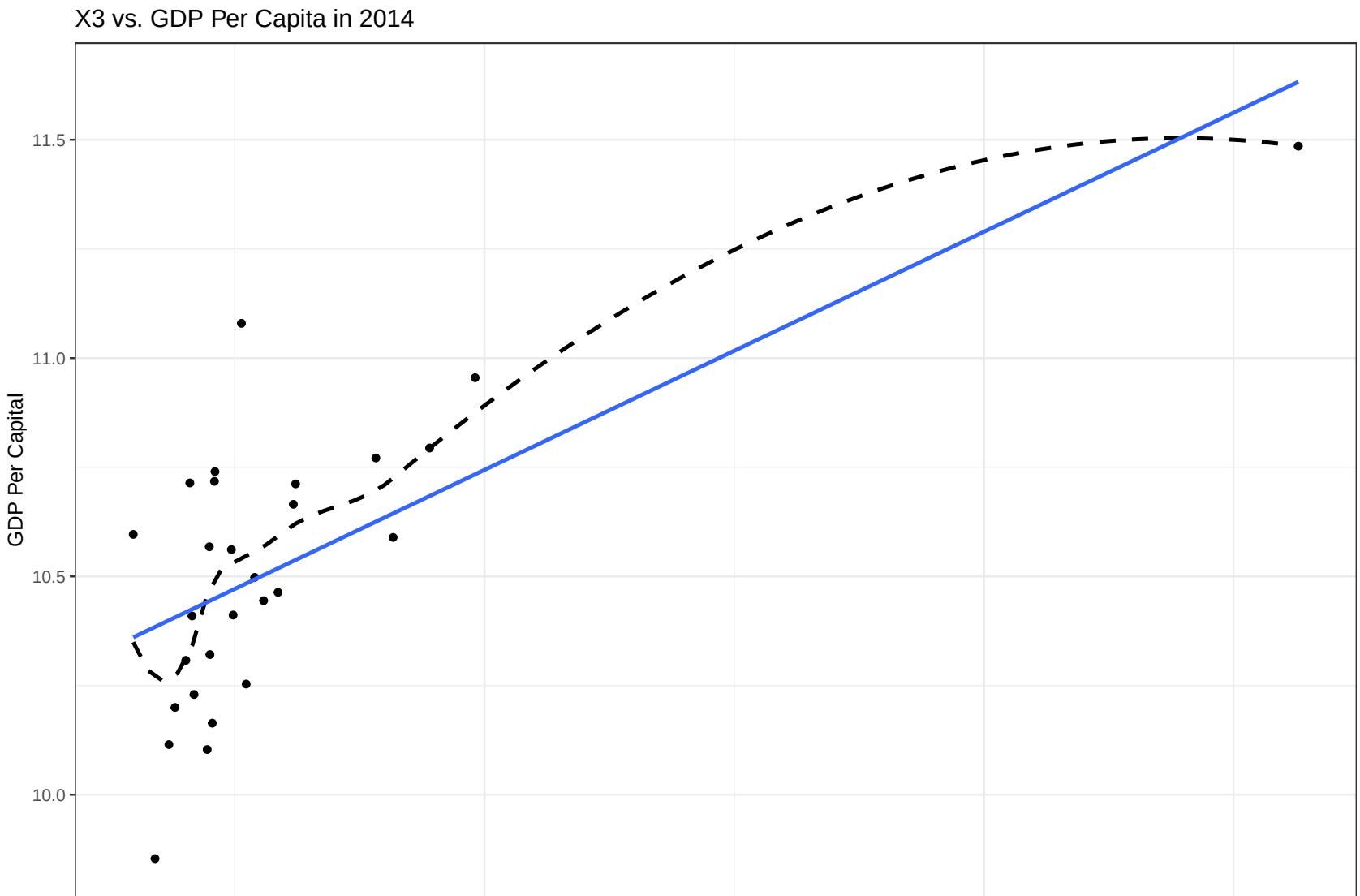

## Model (5) - Focus on the financial parts:

**##**

# Scatter Diagram(5)

- Focus on the financial parts:

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



# Topic: life expectancy

In the OECD data set:
Are there any variable(s) have effect on people's life expectancy?

# Observational units

1. GDP
2. Year/Country
3. Number of practising physicians in 1000 people
4. Population
5. Expenditure on health(Public/Private)

Response Variable:Life expectancy

$H_0$:There are some relationship between observational unit and life

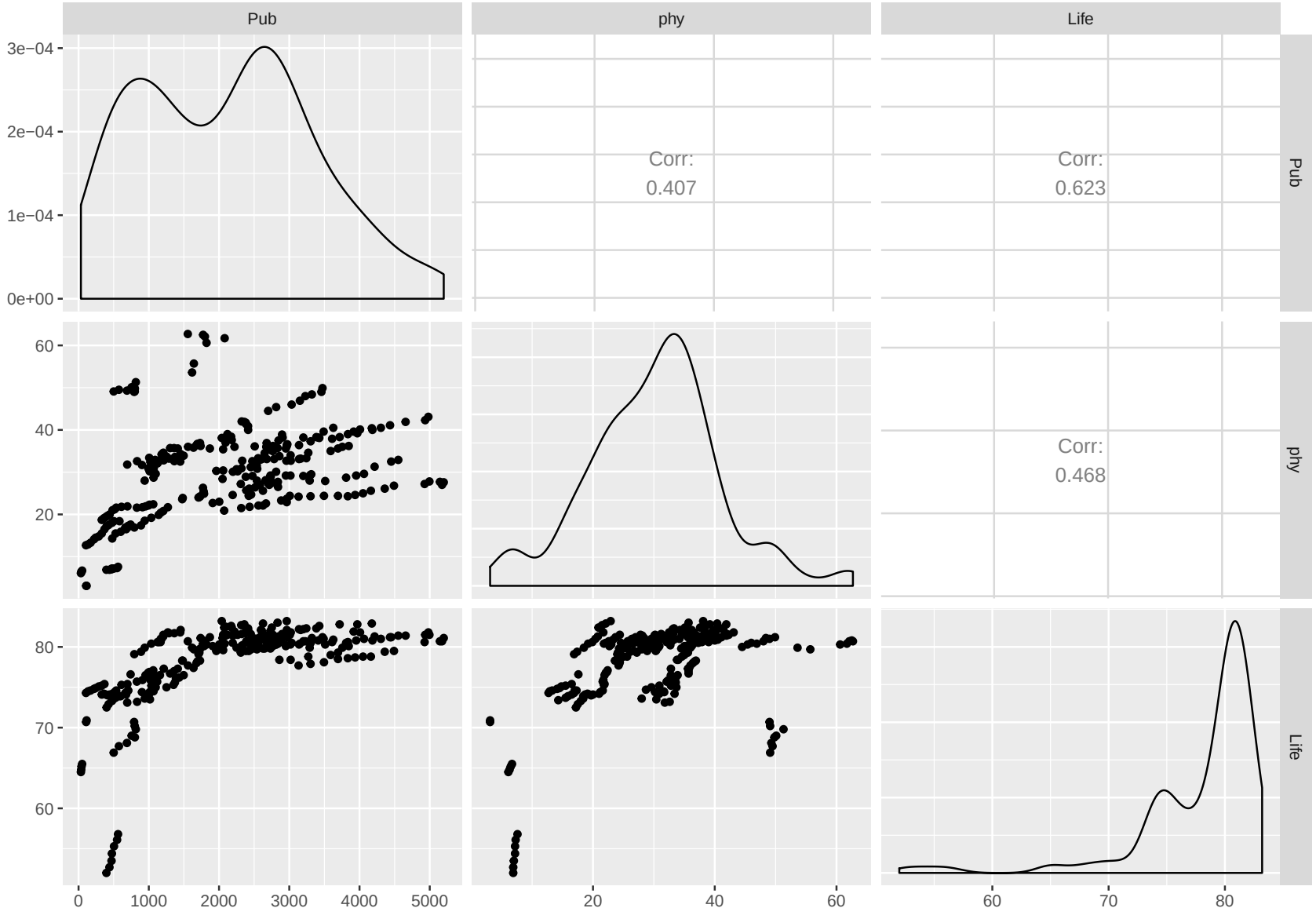
$H_1$:There is no relationship between them

# Data analyse

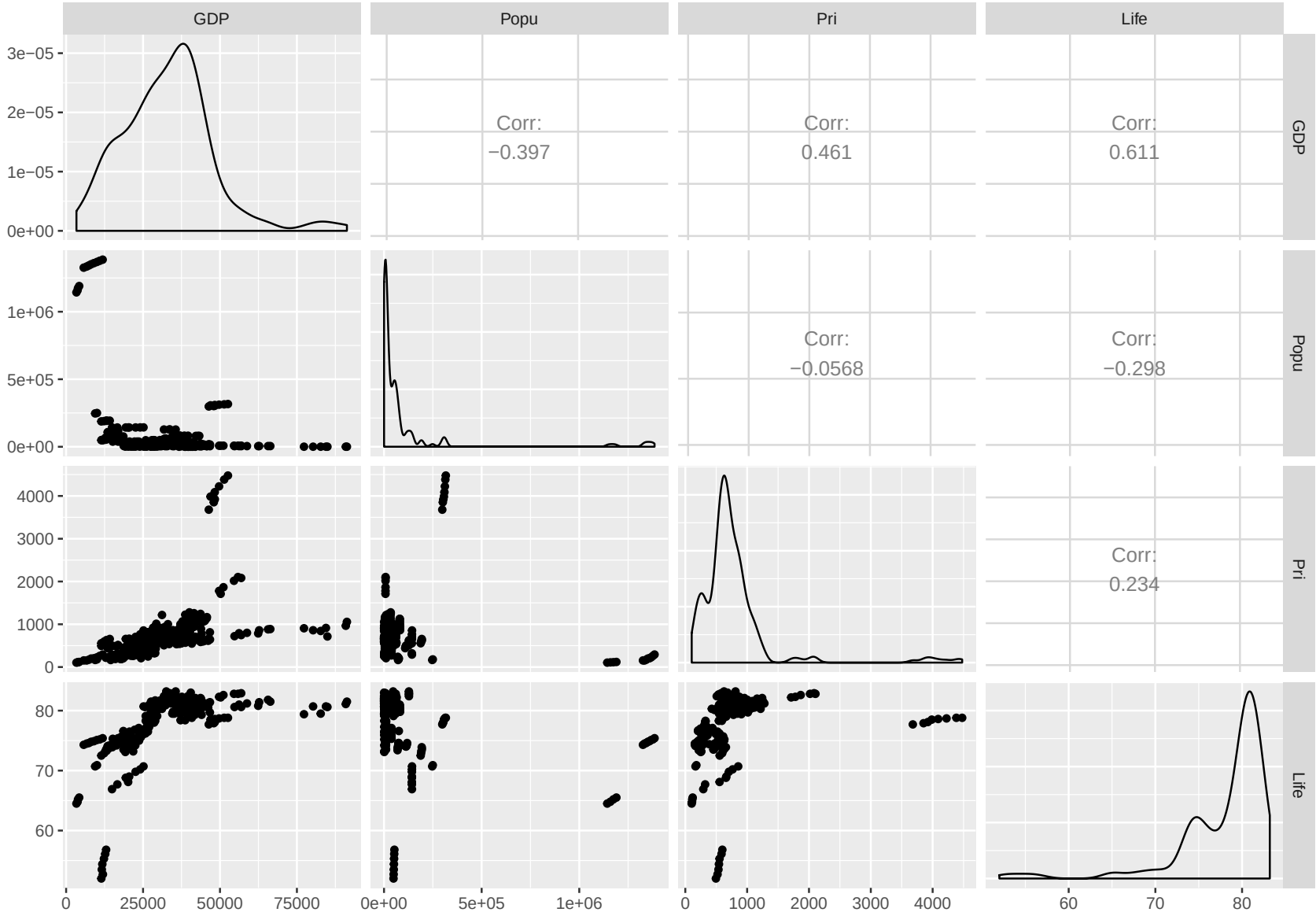
Tidy the data value s Select the columns and remove the rows containing NA

```
## # A tibble: 6 x 8
##   Loc   Year      GDP    Popu   Pri   Pub   phy  Life
##   <fct> <fct>   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 CHN   2006    5717. 1326146  154.  106.  12.7  74.3
## 2 CHN   2007    6665. 1334344  154.  136.  12.9  74.5
## 3 CHN   2008    7412. 1342733  172.  171.  13.3  74.6
## 4 CHN   2009    8118. 1351248  198.  219.  14.1  74.8
## 5 CHN   2010    9031. 1359822  205.  244.  14.5  74.9
## 6 CHN   2011   10017. 1368440  227.  288.  14.8  75.1
```

# Explore the units relationship



# Explore the units relationship



# Some possible relationships

1. Public expenditure and physicians
2. Public expenditure and life expectancy
3. Private expenditure and life expectancy
4. GDP and physicians
5. GDP and life expectancy
6. GDP and private expenditure

We still need do more research on those variables

# Linear model

$Y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} ... + \beta_7 x_{7i} + \varepsilon_1$

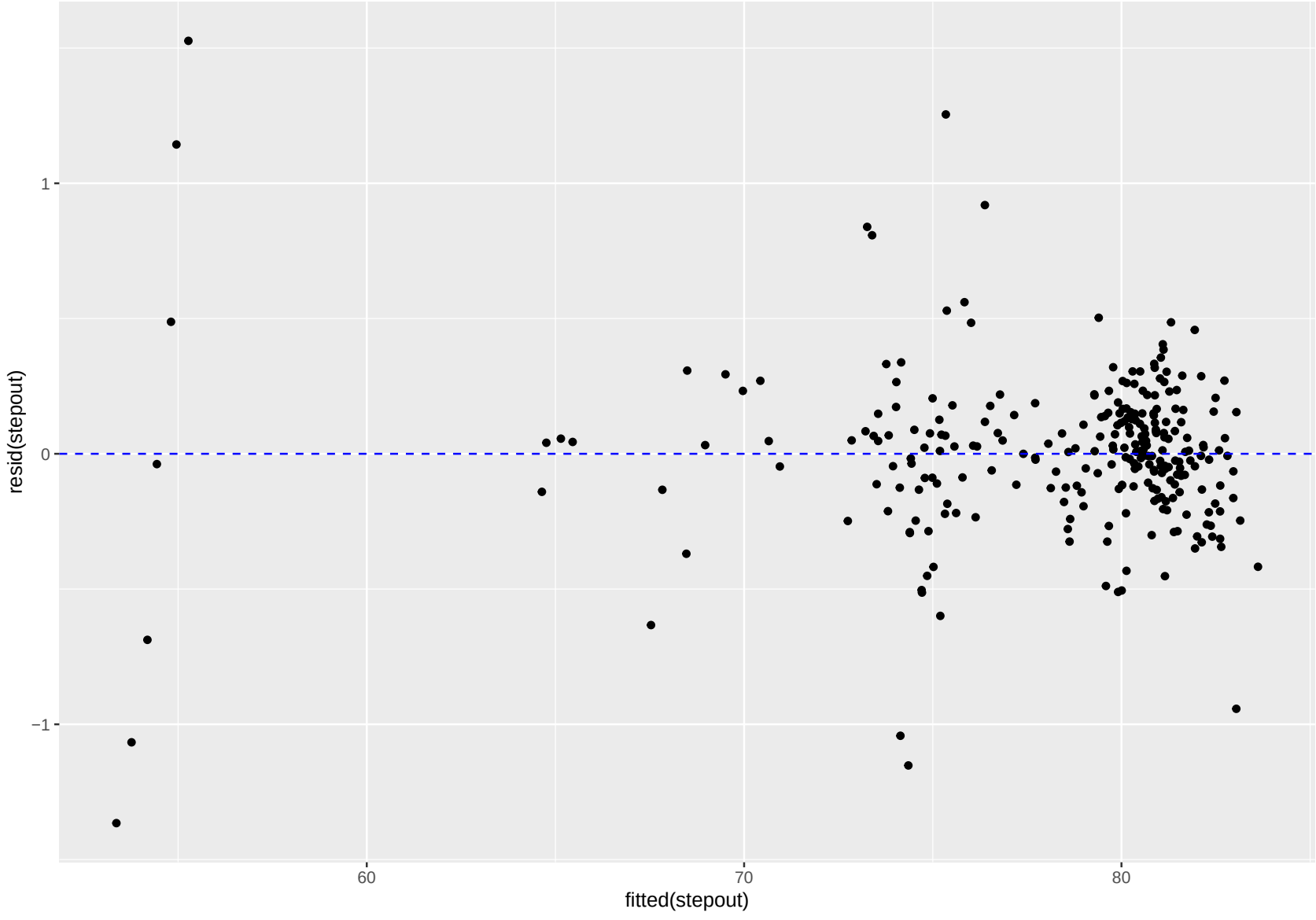
- $\beta_0$ The intercept of the mean line.
- $\beta_1$ The effect on Y when change location given the other variables in the model.
- $\beta_2$ The effect on Y when change year given the other variables in the model.
- $\beta_3$ The effect on Y when change private expenditure given the other variables in the model.
- $\beta_4$ The effect on Y when change private GDP given the other variables in the model.
- $\beta_5$ The effect on Y when change public expenditure given the other variables in the model.
- $\beta_7$ The effect on Y when change physicians given the other variables in the model.
- $\varepsilon_1$ Noise .

# Use “step” do more research

```
## # A tibble: 52 x 3
##    term         estimate   p.value
##    <chr>           <dbl>     <dbl>
##  1 (Intercept)    76.2  2.82e-198
##  2 LocRUS         -3.26 1.14e-  2
##  3 LocCZE          4.28 1.52e-  5
##  4 LocIND        -10.3  3.81e- 68
##  5 LocSVK          2.22 8.68e-  3
##  6 LocZAF        -21.0  2.47e-153
##  7 LocLUX          8.64 3.15e-  6
##  8 LocNLD          7.67 5.79e-  9
##  9 LocISL          8.93 5.01e- 13
## 10 LocAUT          8.68 5.81e- 10
## # ... with 42 more rows
```

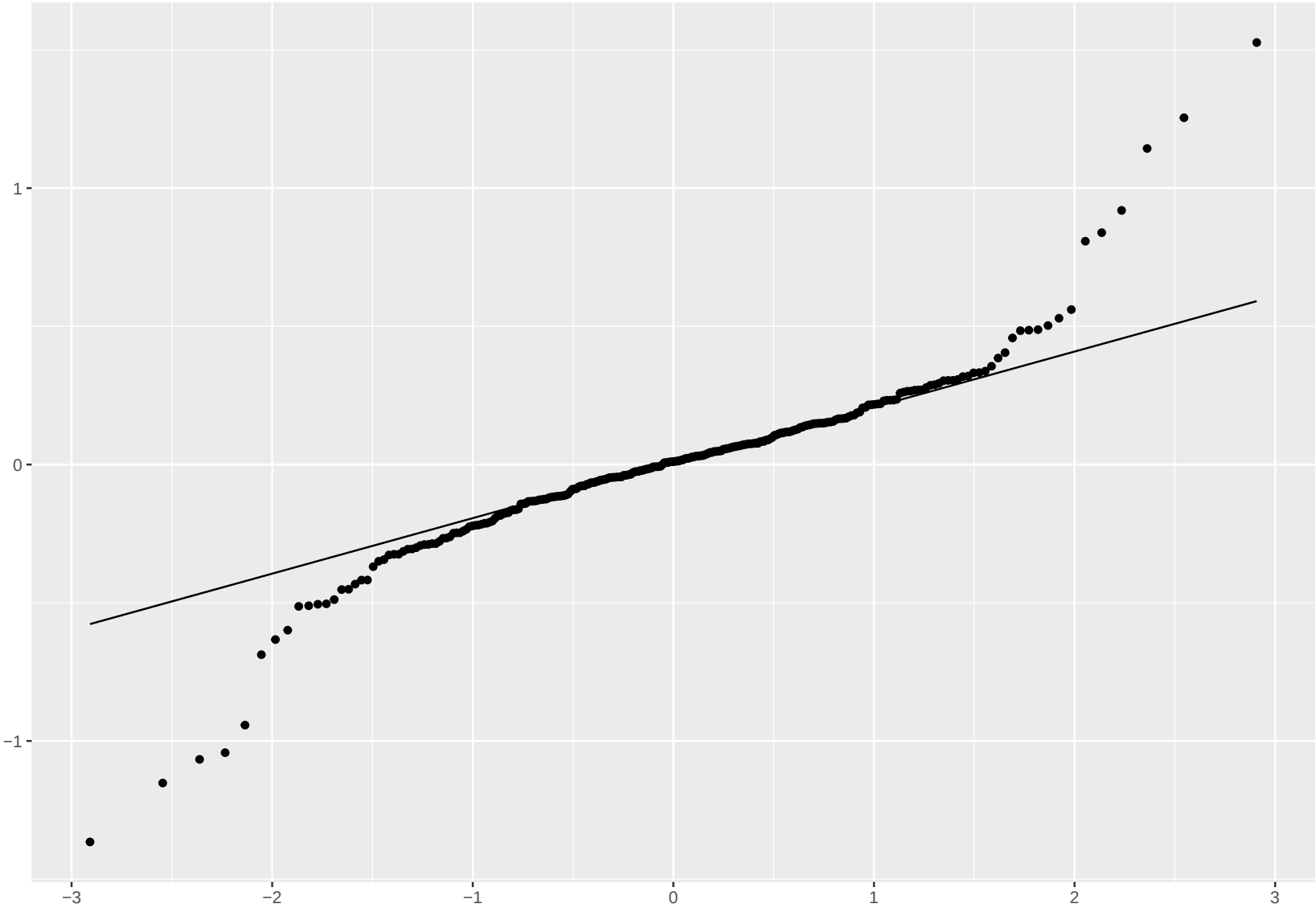
# Conclusion

We see some outline here



# Conclusion

## Heavy tails



# Conclusion

```
##
## Call:
## lm(formula = (Life) ~ Loc + Year + Pri + GDP + Pub + phy + GDP:Pub +
##     GDP:phy + Pub:phy, data = oecd_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36538 -0.12854  0.01049  0.14242  1.52661
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.620e+01  6.783e-01 112.340  < 2e-16 ***
## LocRUS      -3.258e+00  1.277e+00  -2.551 0.011398 *
## LocCZE       4.285e+00  9.688e-01   4.423 1.52e-05 ***
## LocIND      -1.032e+01  4.042e-01 -25.538  < 2e-16 ***
## LocSVK       2.215e+00  8.366e-01   2.648 0.008681 **
## LocZAF      -2.104e+01  3.020e-01 -69.674  < 2e-16 ***
## LocLUX       8.639e+00  1.807e+00   4.782 3.15e-06 ***
## LocNLD       7.672e+00  1.266e+00   6.059 5.79e-09 ***
## LocISL       8.931e+00  1.163e+00   7.679 5.01e-13 ***
## LocAUT       8.676e+00  1.339e+00   6.480 5.81e-10 ***
## LocNOR       8.522e+00  1.460e+00   5.839 1.84e-08 ***
## LocFIN       6.903e+00  1.054e+00   6.552 3.88e-10 ***
## LocSVN       5.662e+00  7.622e-01   7.429 2.31e-12 ***
## LocSWE       8.826e+00  1.225e+00   7.203 8.98e-12 ***
## LocITA       9.211e+00  1.151e+00   8.002 6.67e-14 ***
## LocKOR       6.789e+00  6.852e-01   9.908  < 2e-16 ***
## LocDEU       7.993e+00  1.185e+00   6.748 1.28e-10 ***
## LocTUR      -2.162e-01  3.062e-01  -0.706 0.480902
## LocEST       1.778e+00  8.017e-01   2.218 0.027576 *
## LocUSA       7.578e+00  1.742e+00   4.349 2.08e-05 ***
## LocCAN       7.798e+00  1.058e+00   7.369 3.31e-12 ***
## LocHUN       1.170e+00  7.254e-01   1.613 0.108155
## LocIRL       7.219e+00  1.132e+00   6.380 1.01e-09 ***
## LocIDN      -5.823e+00  5.382e-01 -10.820  < 2e-16 ***
```

# Error analyse

## Error

- Data missing
- Few variables under one factor

## Solutions

- Combine countries with similar GDP
- Research more variables like edcuation level, environment

# Conclusion

- Life $\sim$ Loc + Year + Pri + Popu + GDP + Pub + phy + GDP:Pub + GDP:phy + Pub:phy
- Multiple R-squared: 0.9967, Adjusted R-squared: 0.9959
- F-statistic: 1285 on 52 and 222 DF
- p-value: $<$ 2.2e-16

We may accept the $H_0$

# Net Export

Net exports are the difference between a country's total value of exports and total value of imports. Depending on whether a country imports more goods or exports more goods, net exports can be a positive or negative value.

# My Hypothesis

Countries with higher GDP will have lower imports of goods and higher exports of goods

- H0: We obtain negative results.
- H1: We get a positive relation of GDP and Net Exports (Exports of Goods - Import of Goods)

# Summary GDP

```
datgdpnet %>%
  select(GDP2010,GDP2014) %>%
  summary()
```

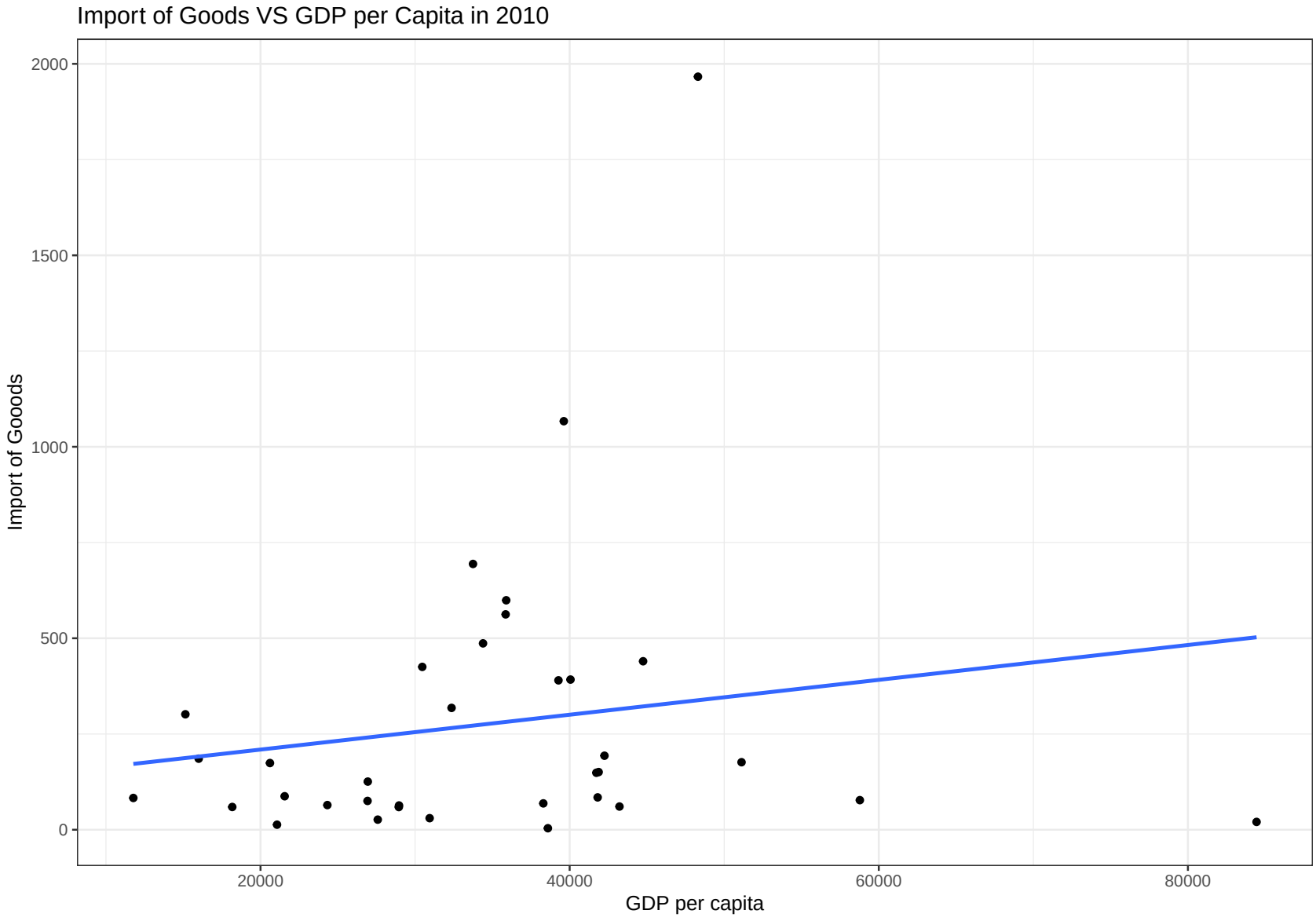
```
##     GDP2010         GDP2014
##  Min.   :11772   Min.   :13146
##  1st Qu.:26933   1st Qu.:28047
##  Median :34396   Median :36810
##  Mean   :34737   Mean   :38422
##  3rd Qu.:41770   3rd Qu.:44978
##  Max.   :84440   Max.   :97273
```

# Summary of Imports

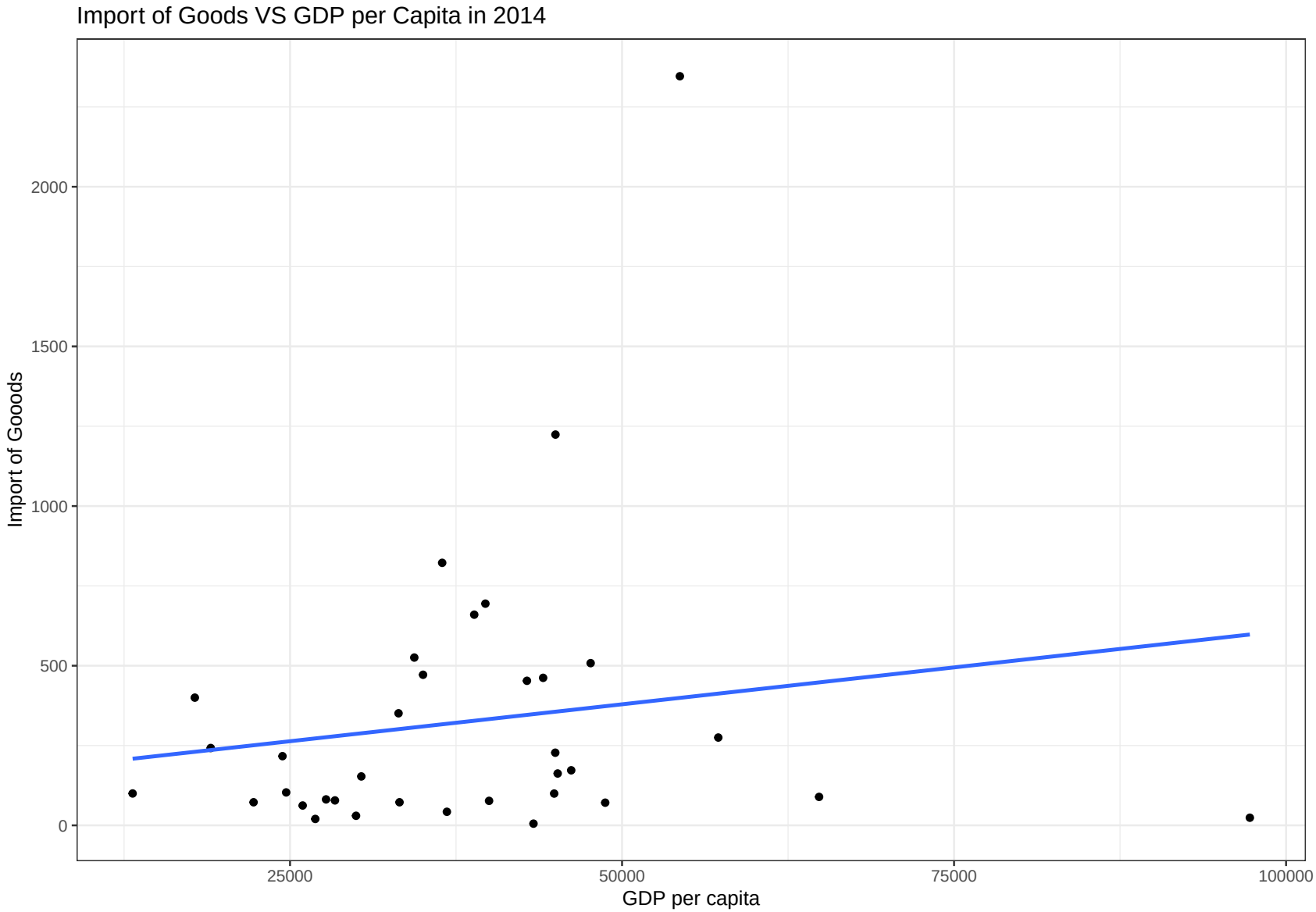
```
datgdpnet %>%
  select(IMP2010,IMP2014) %>%
  summary()
```

```
##     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
```

# Plots


Import of Goods VS GDP per Capita in 2010
Import of Gooods
2000
1500
1000
500
0
20000
40000
60000
80000
GDP per capita

# Plots

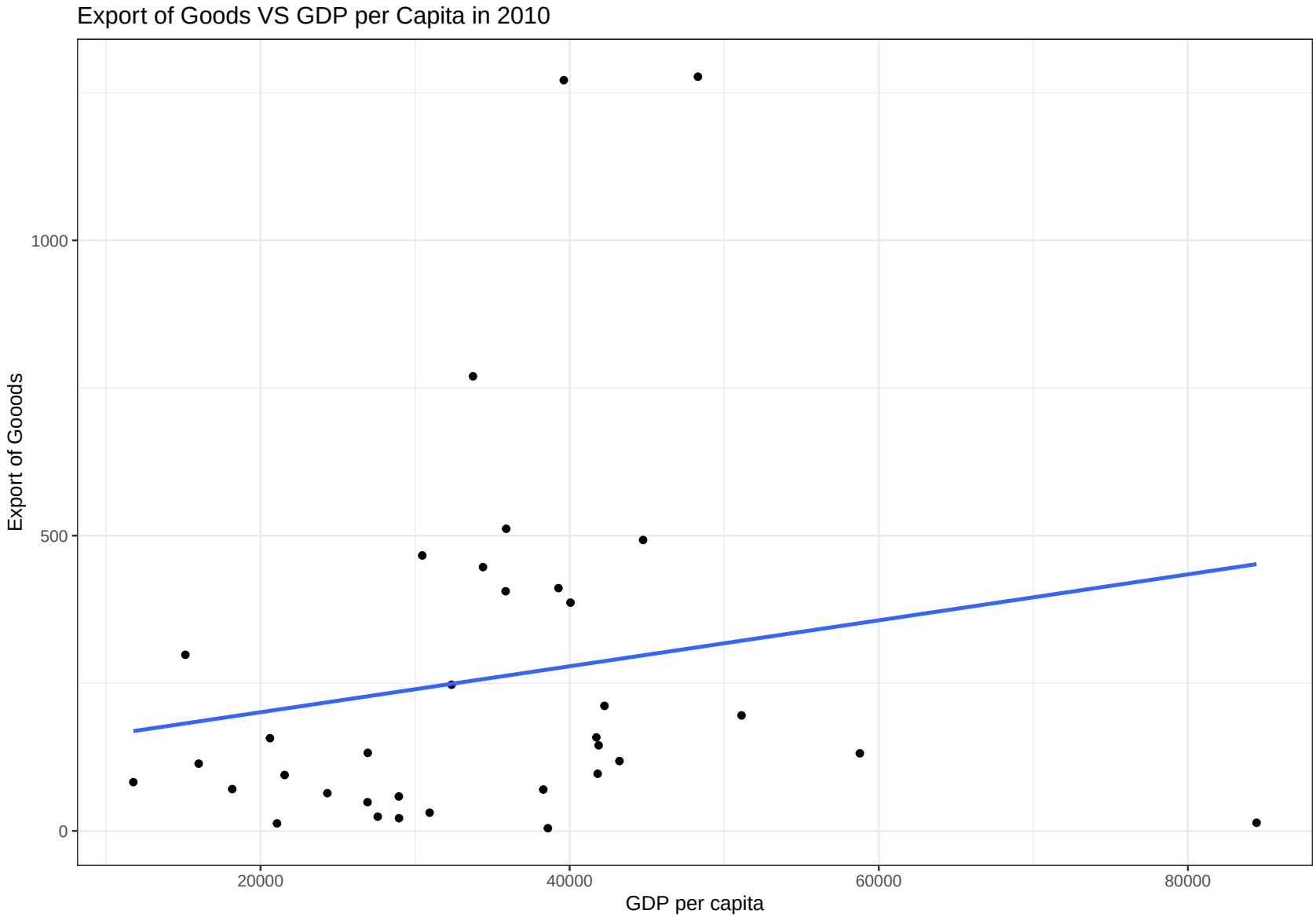

Import of Goods VS GDP per Capita in 2014
Import of Gooods
2000
1500
1000
500
0
25000
50000
75000
100000
GDP per capita

# Summary Of Exports

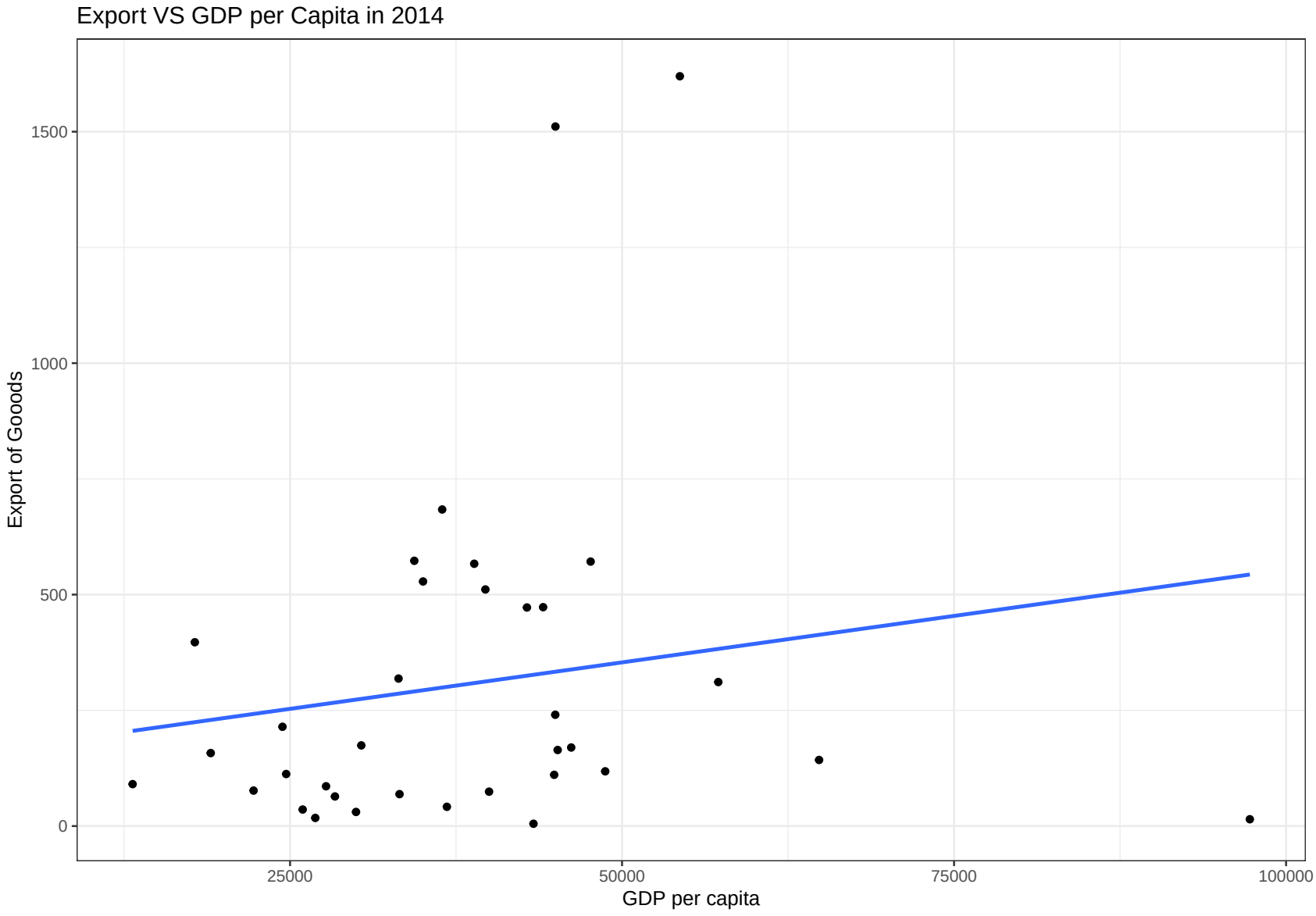
```
datgdpnet %>%
  select(EXP2010,EXP2014) %>%
  summary()
```

```
##     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
```

# Plots


Export of Goods VS GDP per Capita in 2010
Export of Gooods
1000
500
0
20000
40000
60000
80000
GDP per capita

# Plots

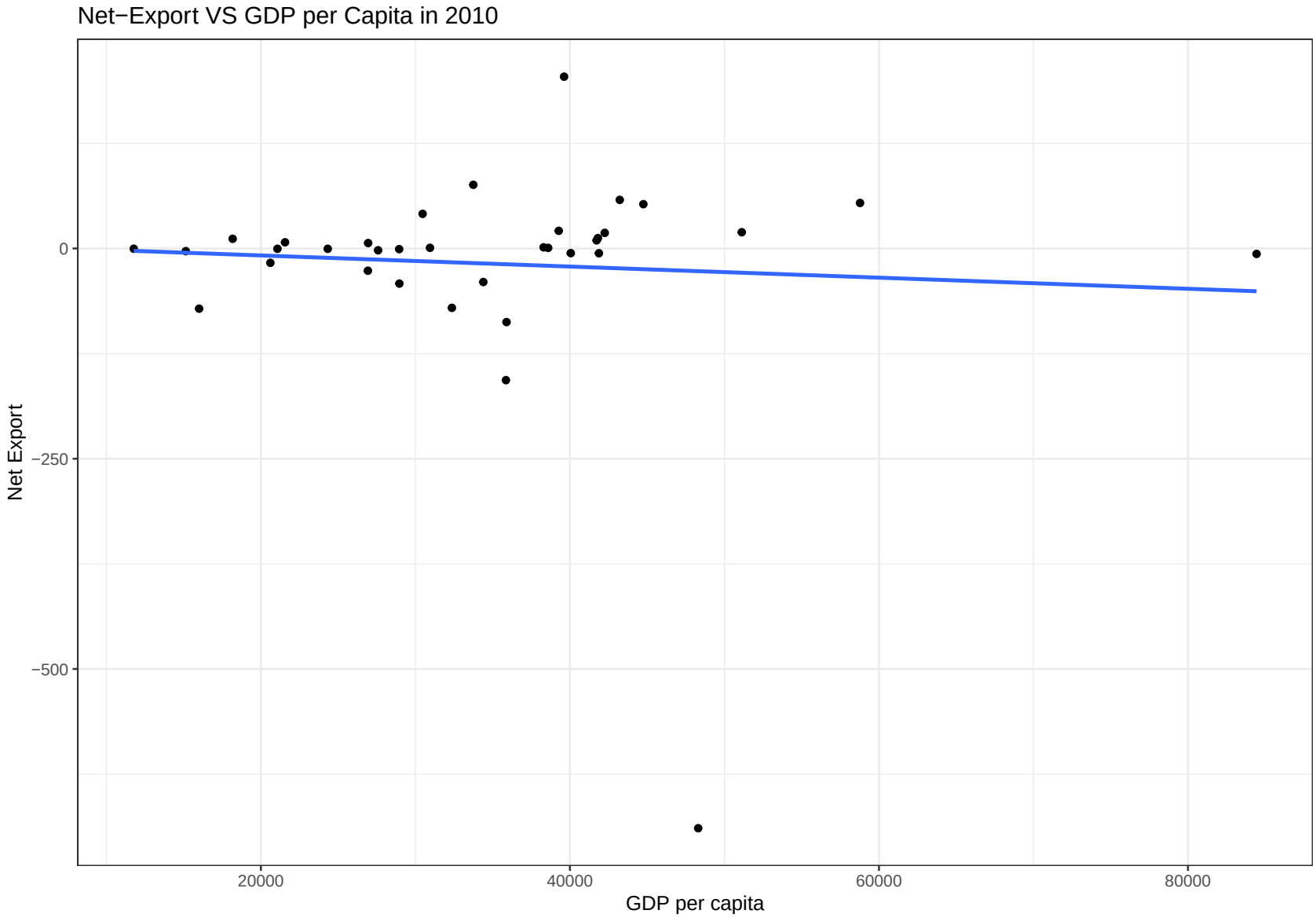

Export VS GDP per Capita in 2014
Export of Gooods
1500
1000
500
0
25000
50000
75000
100000
GDP per capita

# Summary of Net-Export

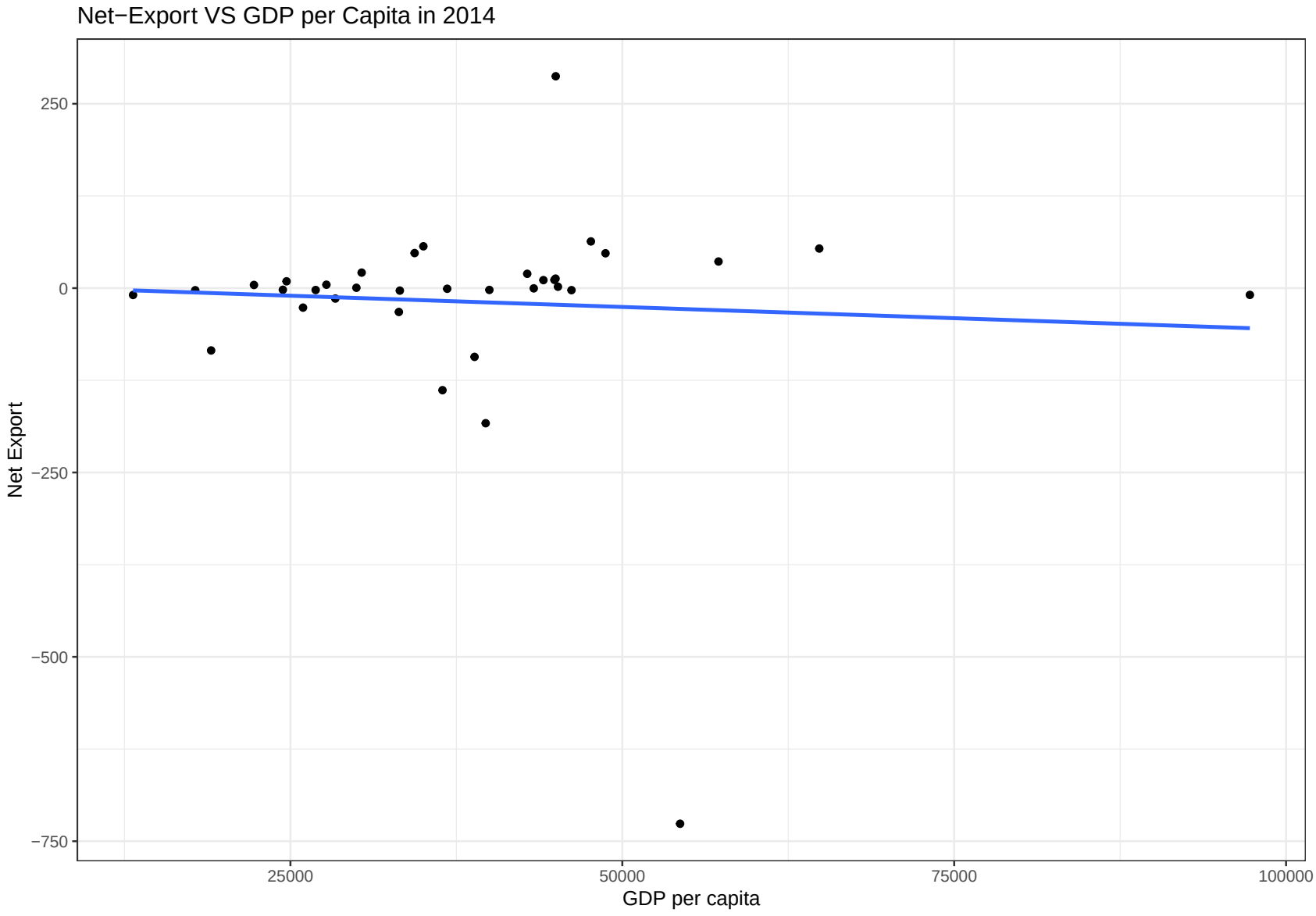
```
datgdpnet %>%
  select(NetExp2010,NetExp2014) %>%
  summary()
```

```
##    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
```

# Plots


Net−Export VS GDP per Capita in 2010
Net Export
0
−250
−500
20000
40000
60000
80000
GDP per capita

# Plots


Net−Export VS GDP per Capita in 2014
250
0
−250
−500
−750
Net Export
25000
50000
75000
100000
GDP per capita

# Test for data

```
datgdpnet %>%
  select(GDP2010,NetExp2010) %>%
  cor(use = "everything")
```

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

```
datgdpnet %>% drop_na()->
  datgdpnet
datgdpnet %>%
  select(GDP2014,NetExp2014) %>%
  cor(use = "everything" )
```

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

# Results

- for both of the years we get negative correlation Value of r.
- Value of r is close to zero so we can say that there is no correlation in them

# Cause of error/Comments

- Outlier
- Data Issue
- Old Data
- Applied to few countries

# Conclusions

- GDP is related to many other economic, health, education measures,trade,value added,expenditure power, and life expectancy but the relationships are sometimes complex.
- Analysis with this data set is limited because it includes only 40 Observations.
- Non-constant variance and departures from normality reduce the validity of some of the analyses.
- Additional analysis should seek to better address assumptions that have not been met in the analysis.
- There is also room for time series analysis to better understand how the countries have changed over time.

Questions/Concerns/Comments?