# Advanced Statistical Methods - Project.

Pawel Chilinski

January 19, 2014

## Contents

# List of Figures

# List of Tables



# 1 Introduction

This project performs exercise of correlational(observational) studies. We have data (predictors) which was gathered by the The World Bank [2] and the predicted variable (Corruption.Index) Failed States Index computed by the United States think-tank Fund for Peace [4]. Because of the nature of the data we cannot make strong causal conclusions on how predictors influence the predicted variable (because of possible lurking variables). I assume that we deal with the simple random sample. The data consists of 31 predictor variables which gives $2^{31}$ =2 147 483 648 different models that can be fitted.

# 2 Data

Description of columns and data types:

| Variable | Type | Name | Description |
|---|---|---|---|
| country | nominal | country | Country for which row contains various variables |
| AG.LND.AGRI.K2 | ratio | Agricultural land (sq. km) | Agricultural land refers to the share of land area that is arable, under permanent crops, and under permanent pastures. |
| AG.LND.ARBL.HA.PC | ratio | Arable land (hectares per person) | Arable land (hectares per person) includes land defined by the FAO as land under temporary crops ,temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded. |
| AG.LND.ARBL.ZS | counted fraction | Arable land (% of land area) | % of land area which is Arable land |
| AG.LND.CROP.ZS | counted fraction | Permanent cropland (% of land area) | A permanent crop is one produced from plants which last for many seasons, rather than being replanted after each harvest. |
| AG.LND.TOTL.K2 | ratio | Land area (sq. km) | Land area is a country's total area. |
| AG.PRD.CROP.XD | ratio | Crop production index (2004-2006 = 100) | Crop production index shows agricultural production for each year relative to the base period 2004-2006. It includes all crops except fodder crops. |
| AG.PRD.FOOD.XD | ratio | Food production index (2004-2006 = 100) | Food production index covers food crops that are considered edible and that contain nutrients. Coffee and tea are excluded because, although edible, they have no nutritive value. |
| AG.PRD.LVSK.XD | ratio | Livestock production index (2004-2006 = 100) | Livestock production index includes meat and milk from all sources, dairy products such as cheese, and eggs, honey, raw silk, wool, and hides and skins. |
| AG.SRF.TOTL.K2 | ratio | Surface area (sq. km) | Surface area is a country's total area, including areas under inland bodies of water and some coastal waterways. |
| AG.YLD.CREL.KG | ratio | Cereal yield (kg per hectare) | Cereal yield measured as kilograms per hectare of harvested land, includes wheat, rice, maize, barley, oats, rye, millet, sorghum, buckwheat, and mixed grains. |
| BM.GSR.INSF.ZS | counted fraction | Insurance and financial services (% of service imports, % BoP) | Insurance and financial services cover various types of insurance provided to nonresidents by resident insurance enterprises and vice versa, and financial intermediary and auxiliary services (except those of insurance enterprises and pension funds) exchanged between residents and nonresidents. |
| BM.GSR.TRVL.ZS | counted fraction | Travel services (% of service imports, BoP) | Travel covers goods and services acquired from an economy by travelers for their own use during visits of less than one year in that economy for either business or personal purposes. |
| BX.GSR.CMCP.ZS | counted fraction | Communications, computer, etc. (% of service exports, % BoP) | Communications, computer, information, and other services cover international telecommunications; computer data; news-related service transactions between residents and nonresidents; construction services; royalties and license fees; miscellaneous business, professional, and technical services; personal, cultural, and recreational services; manufacturing services on physical inputs owned by others; and maintenance and repair services and government services not included elsewhere. |
| BX.KLT.DINV.WD.GD.ZS | counted fraction | Foreign direct investment, net inflows (% of GDP) | Foreign direct investment are the net inflows of investment to acquire a lasting management interest (10 percent or more of voting stock) in an enterprise operating in an economy other than that of the investor. |
| EG.GDP.PUSE.KO.PP | ratio | GDP per unit of energy use (PPP $ per kg of oil equivalent) | GDP per unit of energy use is the PPP GDP per kilogram of oil equivalent of energy use. |
| EG.GDP.PUSE.KO.PP.KD | ratio | GDP per unit of energy use (constant 2005 PPP $ per kg of oil equivalent) | |
| EG.USE.COMM.KT.OE | ratio | Energy use (kt of oil equivalent) | Energy use refers to use of primary energy before transformation to other end-use fuels, which is equal to indigenous production plus imports and stock changes, minus exports and fuels supplied to ships and aircraft engaged in international transport. |
| EG.USE.COMM.GD.PP.KD | ratio | Energy use (kg of oil equivalent) per $1,000 GDP (constant 2005 PPP) | Energy use per PPP GDP is the kilogram of oil equivalent of energy use per constant PPP GDP. |
| EG.USE.ELEC.KH.PC | ratio | Electric power consumption (kWh per capita) | Electric power consumption measures the production of power plants and combined heat and power plants less transmission, distribution, and transformation losses and own use by heat and power plants. |
| EN.ATM.CO2E.KD.GD | ratio | CO2 emissions (kg per 2005 US$ of GDP) | |
| EN.ATM.CO2E.PC | ratio | CO2 emissions (metric tons per capita) | Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring. |
| EN.ATM.PM10.MC.M3 | ratio | PM10, country level (micrograms per cubic meter) | Particulate matter concentrations refer to fine suspended particulates less than 10 microns in diameter (PM10) that are capable of penetrating deep into the respiratory tract and causing significant health damage. |
| ER.H2O.INTR.K3 | ratio | Renewable internal freshwater resources, total (billion cubic meters) | Renewable internal freshwater resources flows refer to internal renewable resources (internal river flows and groundwater from rainfall) in the country. |
| ER.H2O.INTR.PC | ratio | Renewable internal freshwater resources per capita (cubic meters) | Renewable internal freshwater resources flows refer to internal renewable resources (internal river flows and groundwater from rainfall) in the country. Renewable internal freshwater resources per capita are calculated using the World Bank's population estimates. |
| FM.LBL.MQMY.GD.ZS | counted fraction | Money and quasi money (M2) as % of GDP | Money and quasi money comprise the sum of currency outside banks, demand deposits other than those of the central government, and the time, savings, and foreign currency deposits of resident sectors other than the central government. This definition of money supply is frequently called M2; it corresponds to lines 34 and 35 in the International Monetary Fund's (IMF) International Financial Statistics (IFS). |
| FS.AST.PRVT.GD.ZS | counted fraction | Domestic credit to private sector (% of GDP) | Domestic credit to private sector refers to financial resources provided to the private sector, such as through loans, purchases of nonequity securities, and trade credits and other accounts receivable, that establish a claim for repayment. For some countries these claims include credit to public enterprises. |
| IC.CRD.PRVT.ZS | counted fraction | Private credit bureau coverage (% of adults) | Private credit bureau coverage reports the number of individuals or firms listed by a private credit bureau with current information on repayment history, unpaid debts, or credit outstanding. The number is expressed as a percentage of the adult population. |
| IC.EXP.DURS | ratio | Time to export (days) | Time is recorded in calendar days. The time calculation for a procedure starts from the moment it is initiated and runs until it is completed. |
| IC.LGL.CRED.XQ | ordinal | Strength of legal rights index (0=weak to 10=strong) | Strength of legal rights index measures the degree to which collateral and bankruptcy laws protect the rights of borrowers and lenders and thus facilitate lending. The index ranges from 0 to 10, with higher scores indicating that these laws are better designed to expand access to credit. |
| NE.RSB.GNFS.ZS | counted fraction | External balance on goods and services (% of GDP) | External balance on goods and services (formerly resource balance) equals exports of goods and services minus imports of goods and services (previously nonfactor services). |
| NE.TRD.GNFS.ZS | counted fraction | Trade (% of GDP) | Trade is the sum of exports and imports of goods and services measured as a share of gross domestic product. |
| Corruption.Index | ratio | Failed States Index (World Corruption Index) | Definition |

## 2.1 Validating the data.

The total land should be bigger the agricultural land (checking if some county has agricultural bigger that total area):

```
> as.character(kaggle.data$country[kaggle.data$AG.LND.AGRI.K2 > kaggle.data$AG.LND.TOTL.K2])
```

```
[1] "Macedonia"
```

Total area of country (including water area) should be bigger than its land area:

```
> all(kaggle.data$AG.SRF.TOTL.K2 >= kaggle.data$AG.LND.TOTL.K2)
```

```
[1] TRUE
```

Percent of financial services should be $\in (0, 100)$, show countries and values not fulfilling this:

```
> kaggle.data[kaggle.data$BM.GSR.INSF.ZS<0 | kaggle.data$BM.GSR.INSF.ZS>100,c("country","BM.GSR.INSF.ZS")]
```

```
       country BM.GSR.INSF.ZS
1  Afghanistan      -2.552965
30     Eritrea      -2.552965
52        Laos      -2.552965
68       Qatar      -2.552965
```

The rest of the data seems semantically correct.
Fixing the data:

```
> #so we can remove data for Macedonia
> kaggle.data<-kaggle.data[kaggle.data$AG.LND.AGRI.K2 <= kaggle.data$AG.LND.TOTL.K2,]
> #for Afganistan we can check on data.worldbank.org that in 2008 it had BM.GSR.INSF.ZS = 5.8457031725
> kaggle.data[kaggle.data$country=="Afghanistan","BM.GSR.INSF.ZS"]<-5.8457031725
> #because Eritrea and Laos has incorrect value for BM.GSR.INSF.ZS and
> #we cannot fill it from the website we remove observation for this country
> kaggle.data<-kaggle.data[kaggle.data$country!="Eritrea" & kaggle.data$country!="Laos",]
> #It looks that data for Qatar contains incorrect values for few attributes, also
> #without cleaning the data for this country becomes outlier and influential
> #observation so removing it
> kaggle.data<-kaggle.data[kaggle.data$country!="Qatar",]
```

Describing the data, we can see that all values are reasonable and output variable can be considered normally distributed for the requirements of the linear regression model (the kurtosis and skewness are within limits):

| | var | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| country* | 1 | 83 | 43.73 | 25.67 | 43.00 | 43.67 | 32.62 | 1.00 | 87.00 | 86.00 | 0.02 | -1.28 | 2.82 |
| AG.LND.AGRI.K2 | 2 | 83 | 397402.52 | 970121.41 | 47930.00 | 137680.34 | 70779.32 | 50.00 | 5344172.00 | 5344122.00 | 3.52 | 12.58 | 106484.66 |
| AG.LND.ARBL.HA.PC | 3 | 83 | 0.26 | 0.34 | 0.17 | 0.19 | 0.16 | 0.00 | 2.57 | 2.56 | 4.28 | 23.61 | 0.04 |
| AG.LND.ARBL.ZS | 4 | 83 | 15.39 | 13.82 | 11.98 | 13.38 | 13.22 | 0.06 | 56.17 | 56.12 | 1.18 | 0.81 | 1.52 |
| AG.LND.CROP.ZS | 5 | 83 | 5.03 | 8.25 | 1.53 | 3.18 | 1.86 | 0.02 | 48.96 | 48.94 | 2.89 | 9.82 | 0.91 |
| AG.LND.TOTL.K2 | 6 | 83 | 1084101.95 | 2692013.81 | 183780.00 | 360613.61 | 257631.40 | 200.00 | 16381390.00 | 16381190.00 | 3.60 | 13.67 | 295486.90 |
| AG.PRD.CROP.XD | 7 | 83 | 123.82 | 19.71 | 120.00 | 121.55 | 14.83 | 90.00 | 195.00 | 105.00 | 1.13 | 1.37 | 2.16 |
| AG.PRD.FOOD.XD | 8 | 83 | 123.84 | 18.61 | 122.00 | 122.07 | 17.79 | 98.00 | 191.00 | 93.00 | 1.03 | 1.43 | 2.04 |
| AG.PRD.LVSK.XD | 9 | 83 | 125.90 | 19.91 | 127.00 | 125.13 | 25.20 | 95.00 | 177.00 | 82.00 | 0.27 | -0.98 | 2.19 |
| AG.SRF.TOTL.K2 | 10 | 83 | 1128515.06 | 2815532.57 | 185180.00 | 370524.18 | 258254.09 | 200.00 | 17098240.00 | 17098040.00 | 3.61 | 13.69 | 309044.85 |
| AG.YLD.CREL.KG | 11 | 83 | 3244.41 | 1781.31 | 2807.70 | 3023.22 | 1675.63 | 812.60 | 9631.90 | 8819.30 | 1.10 | 1.08 | 195.52 |
| BM.GSR.INSF.ZS | 12 | 83 | 11.68 | 9.54 | 9.30 | 10.12 | 5.52 | 0.75 | 61.87 | 61.11 | 2.75 | 10.14 | 1.05 |
| BM.GSR.TRVL.ZS | 13 | 83 | 27.43 | 12.77 | 25.54 | 26.68 | 11.83 | 2.51 | 71.02 | 68.51 | 0.67 | 0.84 | 1.40 |
| BX.GSR.CMCP.ZS | 14 | 83 | 39.49 | 21.58 | 35.46 | 37.90 | 21.25 | 4.23 | 100.00 | 95.77 | 0.66 | -0.28 | 2.37 |
| BX.KLT.DINV.WD.GD.ZS | 15 | 83 | 17.28 | 58.08 | 6.13 | 8.50 | 4.69 | 0.11 | 524.88 | 524.77 | 8.07 | 67.34 | 6.37 |
| EG.GDP.PUSE.KO.PP | 16 | 83 | 7.83 | 4.07 | 7.02 | 7.49 | 4.20 | 1.44 | 18.48 | 17.04 | 0.70 | -0.25 | 0.45 |
| EG.GDP.PUSE.KO.PP.KD | 17 | 83 | 7.13 | 3.65 | 6.24 | 6.78 | 3.27 | 1.33 | 19.10 | 17.77 | 0.91 | 0.42 | 0.40 |
| EG.USE.COMM.KT.OE | 18 | 83 | 115820.23 | 356607.03 | 13578.00 | 38497.00 | 19866.84 | 42.00 | 2336546.00 | 2336504.00 | 5.12 | 27.21 | 39142.71 |
| EG.USE.COMM.GD.PP.KD | 19 | 83 | 233.44 | 197.17 | 185.99 | 194.42 | 98.74 | 61.29 | 1219.64 | 1158.35 | 3.18 | 12.23 | 21.64 |
| EG.USE.ELEC.KH.PC | 20 | 83 | 4105.63 | 6815.31 | 2017.49 | 2675.50 | 1786.29 | 49.15 | 50067.10 | 50017.95 | 4.33 | 23.67 | 748.08 |
| EN.ATM.CO2E.KD.GD | 21 | 83 | 1.61 | 1.99 | 0.83 | 1.14 | 0.56 | 0.20 | 11.33 | 11.13 | 2.84 | 8.62 | 0.22 |
| EN.ATM.CO2E.PC | 22 | 83 | 5.19 | 5.34 | 3.64 | 4.27 | 4.28 | 0.02 | 24.33 | 24.31 | 1.46 | 1.81 | 0.59 |
| EN.ATM.PM10.MC.M3 | 23 | 83 | 52.87 | 38.71 | 42.56 | 46.52 | 24.94 | 7.44 | 212.39 | 204.95 | 2.02 | 4.76 | 4.25 |
| ER.H2O.INTR.K3 | 24 | 83 | 377.48 | 938.57 | 37.20 | 121.09 | 53.67 | 0.02 | 5418.00 | 5417.98 | 3.48 | 12.63 | 103.02 |
| ER.H2O.INTR.PC | 25 | 83 | 28200.77 | 78417.45 | 2574.03 | 10885.75 | 3274.11 | 25.65 | 590277.78 | 590252.13 | 5.24 | 31.61 | 8607.43 |
| FM.LBL.MQMY.GD.ZS | 26 | 83 | 76.90 | 77.86 | 57.03 | 63.97 | 38.51 | 14.52 | 636.51 | 621.99 | 4.77 | 29.98 | 8.55 |
| FS.AST.PRVT.GD.ZS | 27 | 83 | 66.33 | 58.91 | 44.78 | 56.47 | 37.29 | 1.90 | 319.47 | 317.57 | 1.88 | 4.26 | 6.47 |
| IC.CRD.PRVT.ZS | 28 | 83 | 26.18 | 36.07 | 3.30 | 20.49 | 4.89 | 0.00 | 100.00 | 100.00 | 1.06 | -0.47 | 3.96 |
| IC.EXP.DURS | 29 | 83 | 26.14 | 17.48 | 21.00 | 23.30 | 7.41 | 5.00 | 102.00 | 97.00 | 2.02 | 4.80 | 1.92 |
| IC.LGL.CRED.XQ | 30 | 83 | 5.31 | 2.38 | 5.00 | 5.30 | 2.97 | 1.00 | 10.00 | 9.00 | 0.03 | -1.19 | 0.26 |
| NE.RSB.GNFS.ZS | 31 | 83 | 2.76 | 14.71 | 1.37 | 1.60 | 11.59 | -31.04 | 46.61 | 77.65 | 0.74 | 0.96 | 1.62 |
| NE.TRD.GNFS.ZS | 32 | 83 | 101.23 | 54.54 | 86.49 | 94.50 | 42.13 | 28.97 | 324.33 | 295.35 | 1.37 | 2.48 | 5.99 |
| Corruption.Index | 33 | 83 | 68.91 | 22.20 | 74.50 | 70.00 | 14.38 | 19.70 | 113.40 | 93.70 | -0.50 | -0.46 | 2.44 |

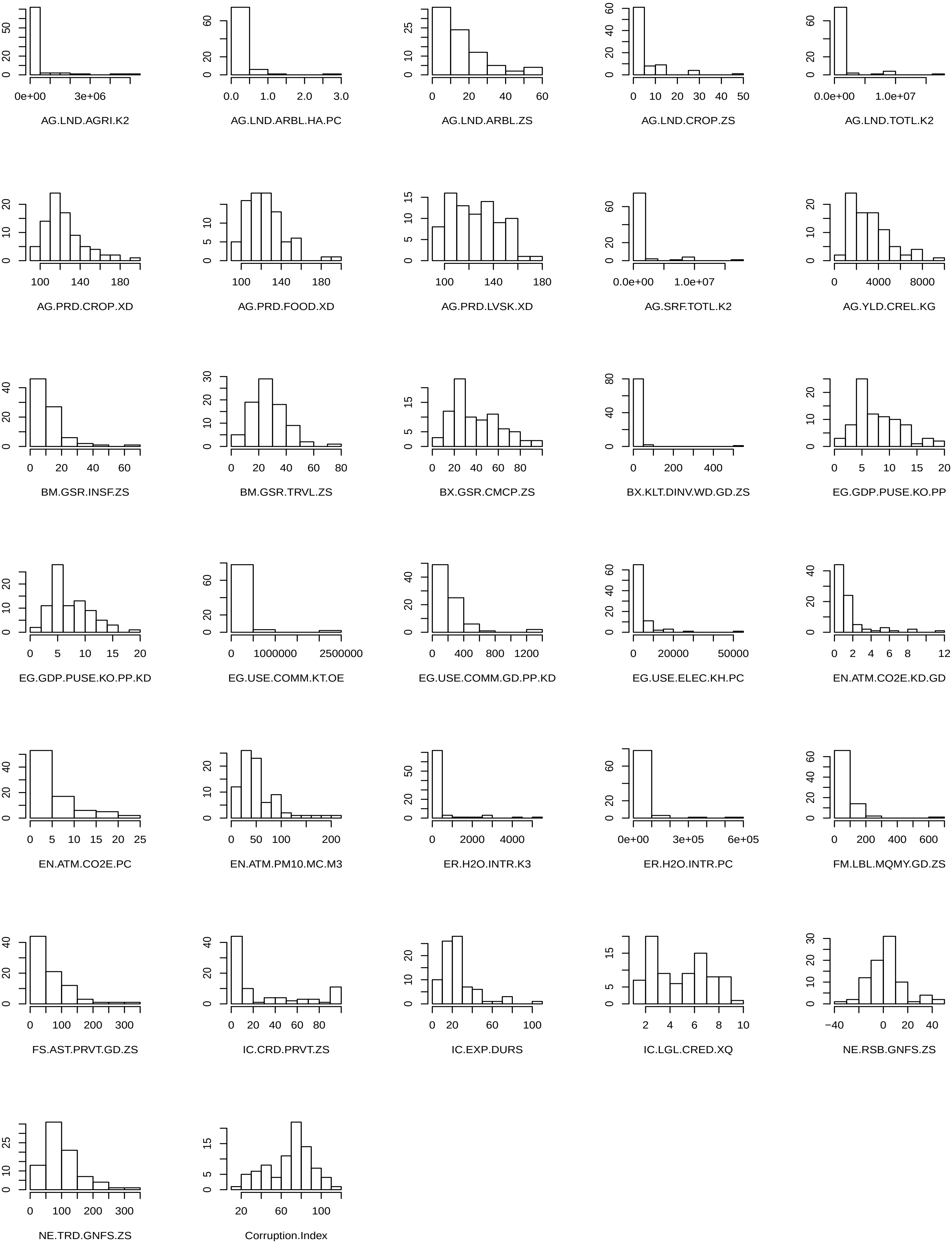


Figure 1: Histograms of variables

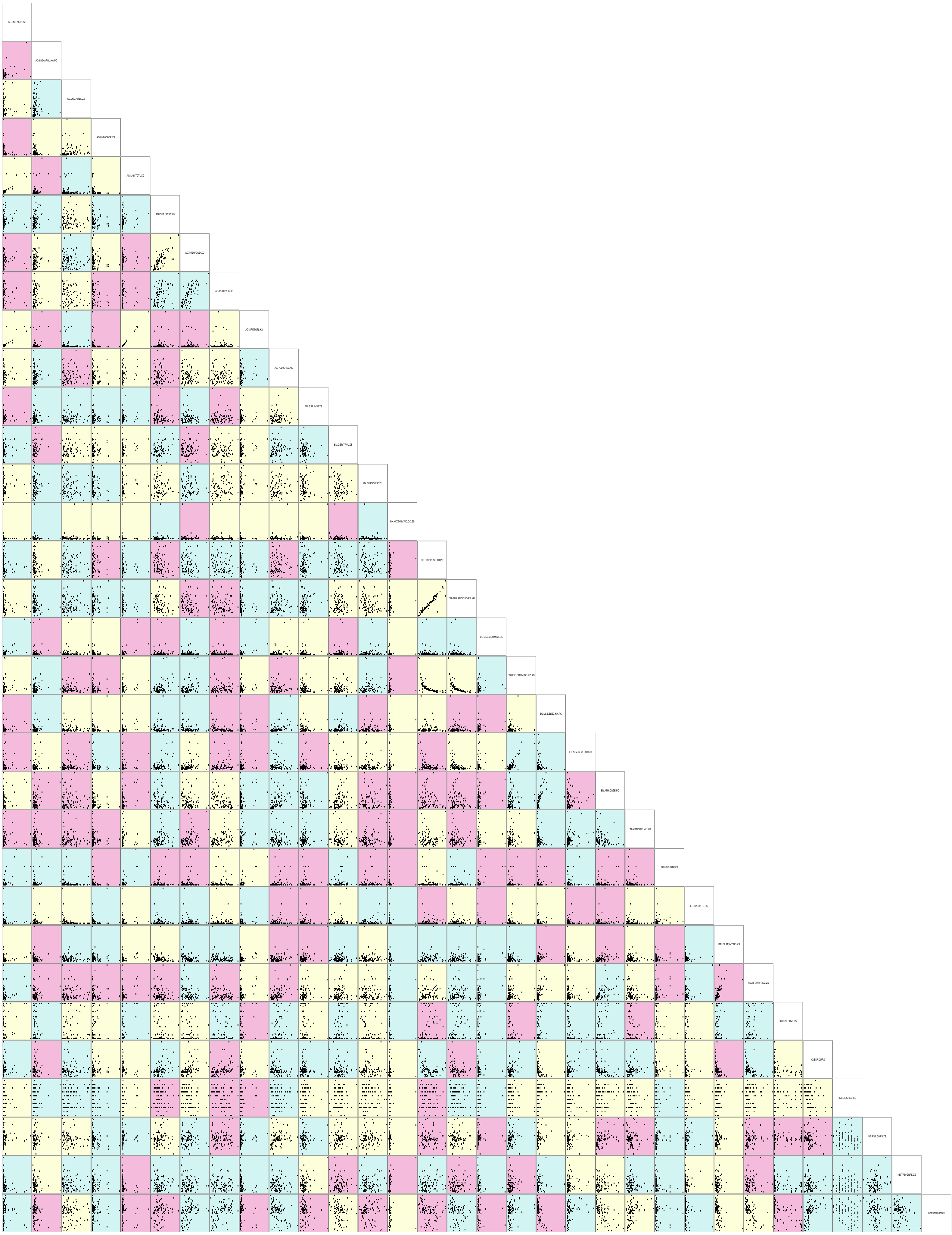

Figure 2: Scatterplots between variables

# 3 Full model

After fitting linear model with all predictors we obtain following significant cooefcients and their 95% confidence intervals (the rest are insignificant so not showing them):

| | Estimate | Std. Error | t value | Pr(>\|t\|) | 2.5 % | 97.5 % |
|---|---|---|---|---|---|---|
| (Intercept) | 135.810155516 | 2.282406e+01 | 5.950307 | 2.453536e-07 | 89.988944394 | 1.816314e+02 |
| AG.LND.ARBL.HA.PC | -26.930898892 | 1.108220e+01 | -2.430103 | 1.865271e-02 | -49.179351954 | -4.682446e+00 |
| AG.YLD.CREL.KG | -0.002641129 | 1.289932e-03 | -2.047495 | 4.577688e-02 | -0.005230776 | -5.148258e-05 |
| BX.KLT.DINV.WD.GD.ZS | 0.237463260 | 8.487408e-02 | 2.797830 | 7.241420e-03 | 0.067071436 | 4.078551e-01 |
| EG.USE.ELEC.KH.PC | -0.001328373 | 6.232705e-04 | -2.131294 | 3.790587e-02 | -0.002579640 | -7.710501e-05 |
| NE.TRD.GNFS.ZS | -0.115311034 | 4.515055e-02 | -2.553923 | 1.368273e-02 | -0.205954544 | -2.466752e-02 |

The entire model (tested by $H_0$ all cooefcients are 0 against some of them are not zero) is significant with F statistic p-value 2.58853727341801e-08 and Adjusted R-squared: 0.639273084259923.

# 4 Veryfying model assumptions

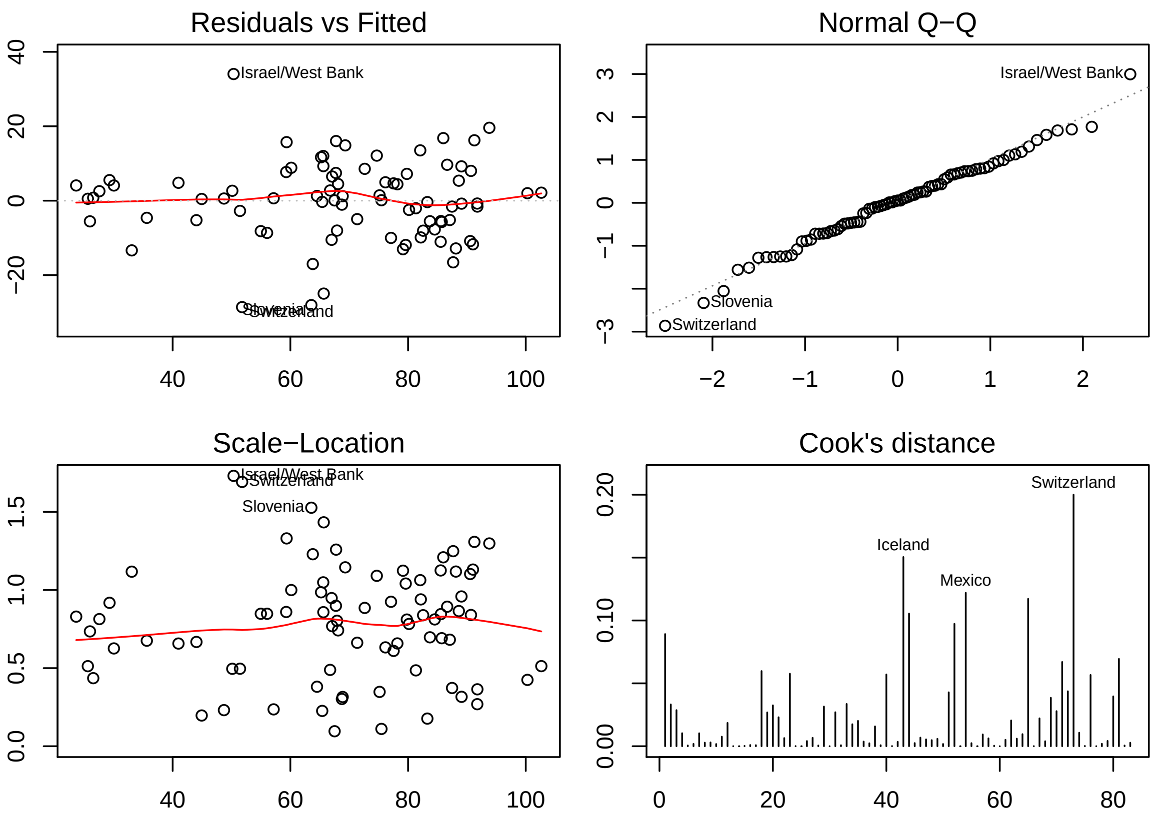


Figure 3: Diagnostic plots for Corruption.Index∼ .

Assumptions of multiple linear regression model:

## 4.1 Quantitative response variable (Corruption.Index)

The response variable is quantitative.

## 4.2 p-1 explanatory independent quantitative variables

From the pairs Figure-2 and correlation Table-1 we can see that some of the predictors are linearly dependant and they will be removed later by different selection algorithms.

## 4.3 Values of the predictors are deterministic

We can safely assume that the values of predictors are deterministic because they were collected by respected research organisation.

Table 1: Pearson correlations between variables (describing linear relationship) as first number in a cell, Spearman's rank correlations between variables as second number in a cell (used to find monotonic relationships) and permutation test as third number in a cell (percent of permutations the has bigger/smaller r value), red font is used for significant values. Because predictors and explained variable distributions depart from the normal distributions (|skew|>3, |kurtosis|>10, examining histograms) I am using Spearman's correlation test and coefficient additionally to the Pearson's one.

| | AG.LND.AGRI.K2 | AG.LND.ARBL.HA.PC | AG.LND.ARBL.ZS | AG.LND.CROP.ZS | AG.LND.TOTL.K2 | AG.PRD.CROP.XD | AG.PRD.FOOD.XD | AG.PRD.LVSK.XD | AG.SRF.TOTL.K2 | AG.YLD.CREL.KG |
|---|---|---|---|---|---|---|---|---|---|---|
| AG.LND.AGRI.K2 | | | | | | | | | | |
| AG.LND.ARBL.HA.PC | +0.51+0.58+0.00 | | | | | | | | | |
| AG.LND.ARBL.ZS | -0.04+0.04+0.41 | +0.02+0.22+0.37 | | | | | | | | |
| AG.LND.CROP.ZS | -0.19-0.38+0.00 | -0.24-0.42+0.00 | +0.10+0.39+0.16 | | | | | | | |
| AG.LND.TOTL.K2 | +0.80+0.92+0.00 | +0.55+0.51+0.00 | -0.11-0.17+0.14 | -0.20-0.56+0.00 | | | | | | |
| AG.PRD.CROP.XD | +0.06+0.32+0.27 | +0.09+0.28+0.19 | +0.01-0.04+0.43 | -0.17-0.25+0.04 | +0.13+0.29+0.14 | | | | | |
| AG.PRD.FOOD.XD | +0.04+0.34+0.34 | -0.02+0.16+0.46 | -0.12-0.12+0.13 | -0.19-0.15+0.03 | +0.07+0.29+0.24 | +0.79+0.78+0.00 | | | | |
| AG.PRD.LVSK.XD | -0.06+0.17+0.28 | -0.15+0.02+0.08 | -0.14-0.14+0.10 | -0.14-0.04+0.09 | -0.06+0.15+0.29 | +0.37+0.36+0.00 | +0.75+0.75+0.00 | | | |
| AG.SRF.TOTL.K2 | +0.79+0.92+0.00 | +0.55+0.51+0.00 | -0.11-0.18+0.17 | -0.20-0.56+0.00 | +1.00+1.00+0.00 | +0.13+0.29+0.16 | +0.07+0.29+0.26 | -0.07+0.15+0.28 | | |
| AG.YLD.CREL.KG | +0.18+0.16+0.06 | -0.05-0.01+0.37 | +0.04+0.05+0.34 | -0.15-0.19+0.07 | +0.13+0.27+0.12 | -0.07-0.01+0.26 | -0.22-0.18+0.03 | -0.23-0.17+0.01 | +0.13+0.27+0.10 | |
| BM.GSR.INSF.ZS | +0.00+0.05+0.39 | -0.07+0.00+0.23 | +0.06+0.05+0.27 | -0.17-0.04+0.03 | -0.03+0.03+0.46 | -0.17-0.17+0.04 | -0.17-0.08+0.06 | -0.01+0.12+0.45 | -0.03+0.03+0.48 | +0.06+0.01+0.28 |
| BM.GSR.TRVL.ZS | +0.16+0.18+0.08 | +0.27+0.29+0.01 | +0.11+0.11+0.18 | -0.13-0.08+0.12 | +0.20+0.15+0.05 | +0.32+0.24+0.00 | +0.04+0.02+0.37 | -0.09-0.13+0.21 | +0.19+0.15+0.04 | +0.15+0.27+0.10 |
| BX.GSR.CMCP.ZS | +0.14+0.24+0.12 | +0.01+0.14+0.42 | -0.06-0.06+0.30 | -0.31-0.33+0.00 | +0.13+0.33+0.12 | -0.08-0.02+0.23 | +0.01+0.06+0.47 | +0.14+0.14+0.11 | +0.13+0.33+0.12 | +0.11+0.10+0.14 |
| BX.KLT.DINV.WD.GD.ZS | -0.09-0.37+0.00 | -0.06-0.17+0.20 | +0.05-0.16+0.26 | -0.02+0.10+0.66 | -0.08-0.40+0.00 | -0.01+0.05+0.54 | -0.13+0.00+0.11 | -0.17-0.07+0.03 | -0.08-0.40+0.00 | +0.19-0.02+0.08 |
| EG.GDP.PUSE.KO.PP | -0.18-0.37+0.05 | -0.18-0.25+0.05 | -0.07-0.06+0.27 | +0.10+0.19+0.16 | -0.17-0.35+0.06 | -0.22-0.28+0.02 | -0.10-0.23+0.20 | -0.15-0.19+0.09 | -0.17-0.34+0.05 | -0.10-0.02+0.19 |
| EG.GDP.PUSE.KO.PP.KD | -0.19-0.40+0.03 | -0.20-0.27+0.01 | -0.04-0.03+0.38 | +0.17+0.22+0.07 | -0.20-0.38+0.02 | -0.23-0.30+0.02 | -0.15-0.26+0.07 | -0.16-0.18+0.08 | -0.20-0.37+0.02 | -0.11-0.02+0.15 |
| EG.USE.COMM.KT.OE | +0.79+0.75+0.00 | +0.13+0.32+0.08 | +0.05+0.12+0.27 | -0.14-0.37+0.00 | +0.66+0.75+0.00 | +0.00+0.26+0.47 | -0.02+0.16+0.48 | -0.10-0.09+0.19 | +0.66+0.75+0.00 | +0.32+0.38+0.01 |
| EG.USE.COMM.GD.PP.KD | +0.07+0.35+0.19 | +0.12+0.24+0.11 | -0.03-0.03+0.41 | -0.12-0.26+0.10 | +0.09+0.36+0.16 | +0.42+0.36+0.00 | +0.32+0.33+0.01 | +0.11+0.19+0.15 | +0.09+0.35+0.17 | +0.05+0.06+0.33 |
| EG.USE.ELEC.KH.PC | +0.10+0.01+0.15 | +0.15+0.00+0.06 | -0.15-0.05+0.06 | +0.05-0.32+0.23 | +0.15+0.11+0.09 | -0.03-0.04+0.43 | -0.27-0.32+0.00 | -0.35-0.48+0.00 | +0.16+0.11+0.09 | +0.46+0.51+0.00 |
| EN.ATM.CO2E.KD.GD | +0.07+0.24+0.23 | +0.13+0.24+0.12 | +0.08-0.01+0.24 | -0.22-0.18+0.00 | +0.11+0.23+0.15 | +0.44+0.35+0.00 | +0.32+0.43+0.00 | +0.13+0.29+0.12 | +0.11+0.23+0.19 | -0.06-0.09+0.28 |
| EN.ATM.CO2E.PC | +0.30-0.02+0.01 | +0.36+0.03+0.00 | -0.06-0.00+0.30 | -0.25-0.29+0.00 | +0.33+0.06+0.01 | -0.12-0.11+0.14 | -0.24-0.30+0.01 | -0.34-0.35+0.00 | +0.33+0.05+0.01 | +0.42+0.51+0.00 |
| EN.ATM.PM10.MC.M3 | +0.03+0.17+0.33 | -0.12-0.07+0.12 | -0.05+0.02+0.34 | -0.15-0.02+0.06 | -0.05+0.13+0.36 | +0.31+0.30+0.00 | +0.47+0.43+0.00 | +0.41+0.41+0.00 | -0.06+0.13+0.34 | -0.12-0.16+0.16 |
| ER.H2O.INTR.K3 | +0.63+0.62+0.00 | +0.27+0.36+0.03 | -0.15-0.18+0.06 | -0.18-0.30+0.00 | +0.86+0.73+0.00 | +0.14-0.00+0.11 | +0.11-0.01+0.16 | +0.04+0.04+0.36 | +0.85+0.73+0.00 | +0.18+0.29+0.05 |
| ER.H2O.INTR.PC | -0.07+0.11+0.28 | +0.03+0.28+0.27 | -0.32-0.38+0.00 | +0.13-0.22+0.11 | -0.00+0.25+0.68 | -0.06-0.08+0.30 | -0.16-0.12+0.07 | -0.02+0.03+0.47 | -0.00+0.26+0.67 | +0.26+0.15+0.03 |
| FM.LBL.MQMY.GD.ZS | +0.04-0.27+0.23 | -0.04-0.29+0.42 | +0.03-0.04+0.34 | -0.07+0.01+0.29 | +0.03-0.19+0.26 | -0.23-0.37+0.01 | -0.34-0.47+0.00 | -0.32-0.31+0.00 | +0.03-0.19+0.26 | +0.34+0.29+0.00 |
| FS.AST.PRVT.GD.ZS | +0.20-0.13+0.05 | +0.12-0.14+0.14 | -0.04+0.05+0.38 | +0.06-0.06+0.26 | +0.19-0.08+0.06 | -0.21-0.24+0.02 | -0.36-0.39+0.00 | -0.40-0.33+0.00 | +0.20-0.08+0.05 | +0.51+0.53+0.00 |
| IC.CRD.PRVT.ZS | +0.24+0.34+0.03 | +0.33+0.22+0.00 | +0.02+0.10+0.38 | -0.08-0.08+0.27 | +0.24+0.35+0.02 | -0.16-0.11+0.07 | -0.21-0.14+0.03 | -0.18-0.10+0.06 | +0.24+0.35+0.02 | +0.34+0.46+0.00 |
| IC.EXP.DURS | -0.11+0.18+0.15 | -0.07+0.03+0.25 | +0.02-0.04+0.38 | -0.02+0.08+0.46 | -0.08+0.14+0.24 | +0.41+0.41+0.00 | +0.44+0.54+0.00 | +0.38+0.51+0.00 | -0.08+0.13+0.25 | -0.31-0.39+0.00 |
| IC.LGL.CRED.XQ | +0.11-0.11+0.17 | +0.09-0.08+0.21 | +0.21+0.15+0.03 | -0.15-0.07+0.09 | -0.00-0.10+0.48 | -0.09-0.13+0.20 | -0.27-0.32+0.01 | -0.27-0.29+0.01 | +0.00-0.11+0.49 | +0.11+0.07+0.16 |
| NE.RSB.GNFS.ZS | +0.07+0.17+0.23 | +0.04+0.08+0.31 | -0.18-0.19+0.05 | -0.27-0.25+0.01 | +0.09+0.23+0.19 | +0.01+0.14+0.50 | +0.05+0.06+0.34 | +0.03-0.05+0.40 | +0.09+0.23+0.20 | +0.21+0.31+0.03 |
| NE.TRD.GNFS.ZS | -0.31-0.55+0.00 | -0.17-0.20+0.04 | -0.10-0.13+0.20 | -0.02+0.09+0.48 | -0.30-0.56+0.00 | -0.05-0.02+0.29 | -0.12-0.04+0.16 | -0.17-0.15+0.05 | -0.30-0.55+0.00 | +0.02-0.08+0.40 |
| Corruption.Index | -0.10+0.24+0.16 | -0.26+0.00+0.01 | -0.05-0.04+0.34 | +0.04+0.10+0.36 | -0.12+0.19+0.13 | +0.23+0.21+0.02 | +0.38+0.46+0.00 | +0.44+0.47+0.00 | -0.13+0.19+0.15 | -0.48-0.37+0.00 |

| | BM.GSR.INSF.ZS | BM.GSR.TRVL.ZS | BX.GSR.CMCP.ZS | BX.KLT.DINV.WD.GD.ZS | EG.GDP.PUSE.KO.PP | EG.GDP.PUSE.KO.PP.KD | EG.USE.COMM.KT.OE | EG.USE.COMM.GD.PP.KD | EG.USE.ELEC.KH.PC | EN.ATM.CO2E.KD.GD |
|---|---|---|---|---|---|---|---|---|---|---|
| BM.GSR.TRVL.ZS | -0.17-0.22+0.05 | | | | | | | | | |
| BX.GSR.CMCP.ZS | -0.09-0.11+0.22 | -0.20-0.13+0.04 | | | | | | | | |
| BX.KLT.DINV.WD.GD.ZS | +0.55+0.16+0.01 | -0.15-0.09+0.06 | -0.07-0.17+0.33 | | | | | | | |
| EG.GDP.PUSE.KO.PP | +0.18+0.13+0.05 | -0.05-0.08+0.35 | -0.20-0.26+0.04 | +0.11+0.16+0.17 | | | | | | |
| EG.GDP.PUSE.KO.PP.KD | +0.20+0.20+0.05 | +0.02-0.02+0.43 | -0.27-0.32+0.01 | +0.04+0.10+0.32 | +0.94+0.93+0.00 | | | | | |
| EG.USE.COMM.KT.OE | +0.03-0.16+0.29 | +0.10+0.23+0.17 | +0.15+0.28+0.10 | -0.07-0.39+0.00 | -0.15-0.30+0.07 | -0.16-0.36+0.05 | | | | |
| EG.USE.COMM.GD.PP.KD | -0.12-0.18+0.10 | -0.05+0.04+0.38 | +0.03+0.24+0.36 | -0.05-0.01+0.35 | -0.61-0.87+0.00 | -0.60-0.86+0.00 | +0.06+0.34+0.22 | | | |
| EG.USE.ELEC.KH.PC | -0.02-0.13+0.48 | +0.21+0.32+0.04 | +0.05+0.08+0.31 | +0.22+0.12+0.03 | -0.13-0.06+0.10 | -0.14-0.08+0.09 | +0.12+0.40+0.08 | +0.09+0.15+0.18 | | |
| EN.ATM.CO2E.KD.GD | -0.07-0.07+0.26 | +0.02+0.07+0.43 | +0.03+0.09+0.39 | -0.06-0.02+0.33 | -0.47-0.49+0.00 | -0.46-0.53+0.00 | +0.07+0.19+0.20 | +0.89+0.63+0.00 | -0.13-0.07+0.05 | |
| EN.ATM.CO2E.PC | +0.14-0.04+0.10 | +0.17+0.29+0.06 | +0.03-0.03+0.38 | +0.39+0.13+0.01 | -0.15-0.08+0.10 | -0.15-0.05+0.07 | +0.28+0.31+0.02 | +0.10+0.17+0.19 | +0.57+0.84+0.00 | +0.05+0.01+0.27 |
| EN.ATM.PM10.MC.M3 | -0.00+0.08+0.56 | -0.10-0.14+0.19 | +0.24+0.09+0.01 | -0.13-0.10+0.04 | -0.01-0.08+0.49 | -0.07-0.08+0.29 | -0.00+0.08+0.56 | +0.19+0.14+0.05 | -0.27-0.34+0.00 | +0.24+0.34+0.04 |
| ER.H2O.INTR.K3 | -0.02-0.03+0.51 | +0.11+0.16+0.18 | +0.14+0.24+0.12 | -0.08-0.39+0.01 | -0.05-0.09+0.37 | -0.06-0.16+0.29 | +0.57+0.51+0.00 | -0.01+0.10+0.57 | +0.11+0.12+0.13 | +0.01-0.02+0.35 |
| ER.H2O.INTR.PC | -0.10-0.12+0.13 | +0.08+0.14+0.21 | +0.13+0.01+0.12 | -0.00-0.17+0.71 | -0.10+0.16+0.20 | -0.12+0.08+0.12 | -0.07+0.00+0.20 | +0.07-0.10+0.18 | +0.63+0.05+0.00 | -0.10-0.15+0.15 |
| FM.LBL.MQMY.GD.ZS | +0.53+0.24+0.00 | -0.03+0.06+0.40 | -0.03-0.01+0.43 | +0.80+0.30+0.00 | +0.12+0.17+0.14 | +0.10+0.17+0.16 | +0.11-0.04+0.09 | -0.18-0.16+0.01 | +0.33+0.43+0.02 | -0.20-0.23+0.00 |
| FS.AST.PRVT.GD.ZS | +0.16+0.15+0.10 | +0.11+0.15+0.17 | -0.05-0.10+0.35 | +0.25+0.25+0.04 | -0.03+0.09+0.42 | -0.03+0.10+0.40 | +0.31+0.12+0.02 | -0.03-0.06+0.43 | +0.69+0.56+0.00 | -0.20-0.26+0.01 |
| IC.CRD.PRVT.ZS | -0.03-0.01+0.45 | +0.24+0.25+0.03 | -0.06+0.02+0.31 | -0.13-0.20+0.01 | +0.04+0.06+0.37 | +0.03+0.07+0.36 | +0.18+0.40+0.06 | -0.17-0.10+0.04 | +0.42+0.31+0.00 | -0.28-0.27+0.00 |
| IC.EXP.DURS | -0.10-0.09+0.19 | -0.12-0.12+0.15 | +0.01-0.03+0.43 | -0.15-0.19+0.01 | -0.36-0.37+0.00 | -0.33-0.35+0.00 | -0.12-0.05+0.10 | +0.66+0.38+0.00 | -0.28-0.51+0.00 | +0.65+0.45+0.00 |
| IC.LGL.CRED.XQ | +0.03+0.01+0.40 | +0.14+0.12+0.10 | +0.08+0.11+0.22 | +0.09+0.19+0.25 | -0.12-0.08+0.13 | -0.10-0.06+0.19 | +0.12+0.04+0.14 | -0.10-0.01+0.21 | +0.27+0.33+0.01 | -0.16-0.19+0.07 |
| NE.RSB.GNFS.ZS | +0.11+0.02+0.15 | -0.01+0.09+0.45 | +0.11+0.09+0.15 | +0.23+0.04+0.06 | -0.12-0.11+0.13 | -0.11-0.10+0.15 | +0.02+0.25+0.43 | +0.26+0.23+0.02 | +0.23+0.38+0.03 | +0.22+0.09+0.03 |
| NE.TRD.GNFS.ZS | +0.29+0.03+0.01 | -0.13-0.07+0.11 | -0.03-0.09+0.41 | +0.56+0.64+0.00 | +0.17+0.17+0.09 | +0.10+0.11+0.18 | -0.24-0.44+0.00 | +0.03+0.01+0.38 | +0.05+0.12+0.28 | +0.07+0.18+0.24 |
| Corruption.Index | -0.16-0.04+0.09 | -0.13-0.22+0.13 | +0.04+0.12+0.36 | -0.22-0.21+0.03 | -0.19-0.31+0.05 | -0.19-0.33+0.05 | -0.11-0.00+0.15 | +0.22+0.24+0.02 | -0.59-0.59+0.00 | +0.31+0.42+0.00 |

| | EN.ATM.CO2E.PC | EN.ATM.PM10.MC.M3 | ER.H2O.INTR.K3 | ER.H2O.INTR.PC | FM.LBL.MQMY.GD.ZS | FS.AST.PRVT.GD.ZS | IC.CRD.PRVT.ZS | IC.EXP.DURS | IC.LGL.CRED.XQ | NE.RSB.GNFS.ZS | NE.TRD.GNFS.ZS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EN.ATM.PM10.MC.M3 | -0.22-0.29+0.01 | | | | | | | | | | |
| ER.H2O.INTR.K3 | +0.16+0.03+0.09 | -0.13-0.17+0.08 | | | | | | | | | |
| ER.H2O.INTR.PC | +0.01-0.03+0.41 | -0.19-0.31+0.00 | +0.07+0.69+0.18 | | | | | | | | |
| FM.LBL.MQMY.GD.ZS | +0.54+0.52+0.00 | -0.16-0.11+0.03 | +0.02-0.09+0.31 | +0.00-0.13+0.39 | | | | | | | |
| FS.AST.PRVT.GD.ZS | +0.56+0.65+0.00 | -0.27-0.30+0.00 | +0.16+0.06+0.10 | +0.36-0.03+0.01 | +0.62+0.80+0.00 | | | | | | |
| IC.CRD.PRVT.ZS | +0.29+0.29+0.01 | -0.24-0.21+0.00 | +0.25+0.47+0.02 | +0.18+0.29+0.06 | -0.00+0.10+0.54 | +0.36+0.38+0.00 | | | | | |
| IC.EXP.DURS | -0.33-0.55+0.00 | +0.42+0.36+0.00 | -0.09-0.07+0.19 | -0.10-0.12+0.12 | -0.37-0.57+0.00 | -0.45-0.64+0.00 | -0.35-0.40+0.00 | | | | |
| IC.LGL.CRED.XQ | +0.28+0.24+0.00 | -0.25-0.25+0.02 | -0.05-0.03+0.31 | +0.01-0.13+0.47 | +0.23+0.31+0.01 | +0.35+0.34+0.00 | +0.15+0.17+0.09 | -0.26-0.33+0.01 | | | |
| NE.RSB.GNFS.ZS | +0.44+0.46+0.00 | +0.00-0.08+0.49 | +0.09+0.19+0.20 | +0.07+0.14+0.25 | +0.15+0.03+0.10 | +0.09+0.16+0.20 | +0.08+0.18+0.24 | +0.02-0.15+0.42 | -0.14-0.18+0.11 | | |
| NE.TRD.GNFS.ZS | +0.18+0.09+0.06 | -0.22-0.17+0.02 | -0.30-0.48+0.00 | +0.08-0.15+0.19 | +0.42+0.15+0.00 | +0.10+0.10+0.17 | -0.32-0.36+0.00 | -0.16-0.15+0.05 | +0.02+0.05+0.46 | +0.18+0.12+0.06 | |
| Corruption.Index | -0.60-0.60+0.00 | +0.41+0.43+0.00 | -0.09-0.04+0.21 | -0.18-0.18+0.06 | -0.42-0.48+0.00 | -0.59-0.60+0.00 | -0.51-0.37+0.00 | +0.56+0.67+0.00 | -0.27-0.31+0.00 | -0.15-0.12+0.08 | -0.19-0.19+0.04 |

### 4.4 Corruption.Index are values of random variables satyfing linear equations

Looking at pairs Figure-2 we can see that most of the predictor variables influence response variable in linear way. Later I am going to try to apply quadratic transformation of predictors to see if it can improve the model.

### 4.5 Normality assumptions of errors

$\epsilon_i$ mutually independent random variables with mean 0 and variance $\sigma^2$. For testing purposes we assume that $\epsilon_i \sim N(0, \sigma^2)$ (mean 0 and constant variance). The distribution assumption is needed for testing purposes and for the least sqares estimates to be optimal.

Checking normality assumption for errors:

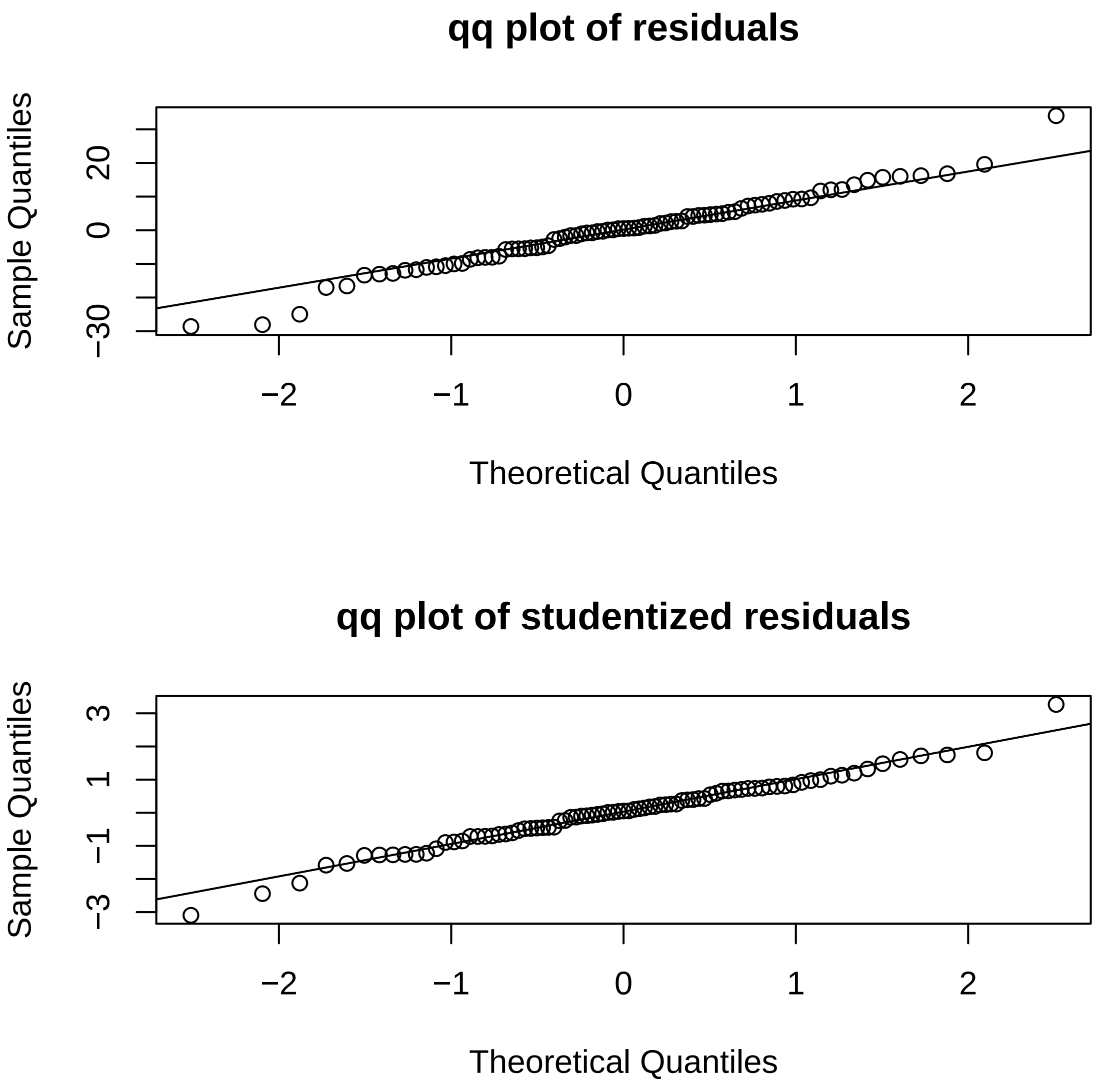


Figure 4: QQ plot

QQ plots depicts quite good linear trend so we can assume residuals have normal distribution. The long-tailed distributions of errors could pose problems for tools we used here but as we see there are no visible long tails on qqplot. We can also run Shapiro-Wilk normality test on residuals which $H_0$ states that data is normal. The p-value from the test is 0.23 so we cannot reject $H_0$.

Residuals are estimators for the error term on the regression model so they have to fullfil requirements imposed on the error term.

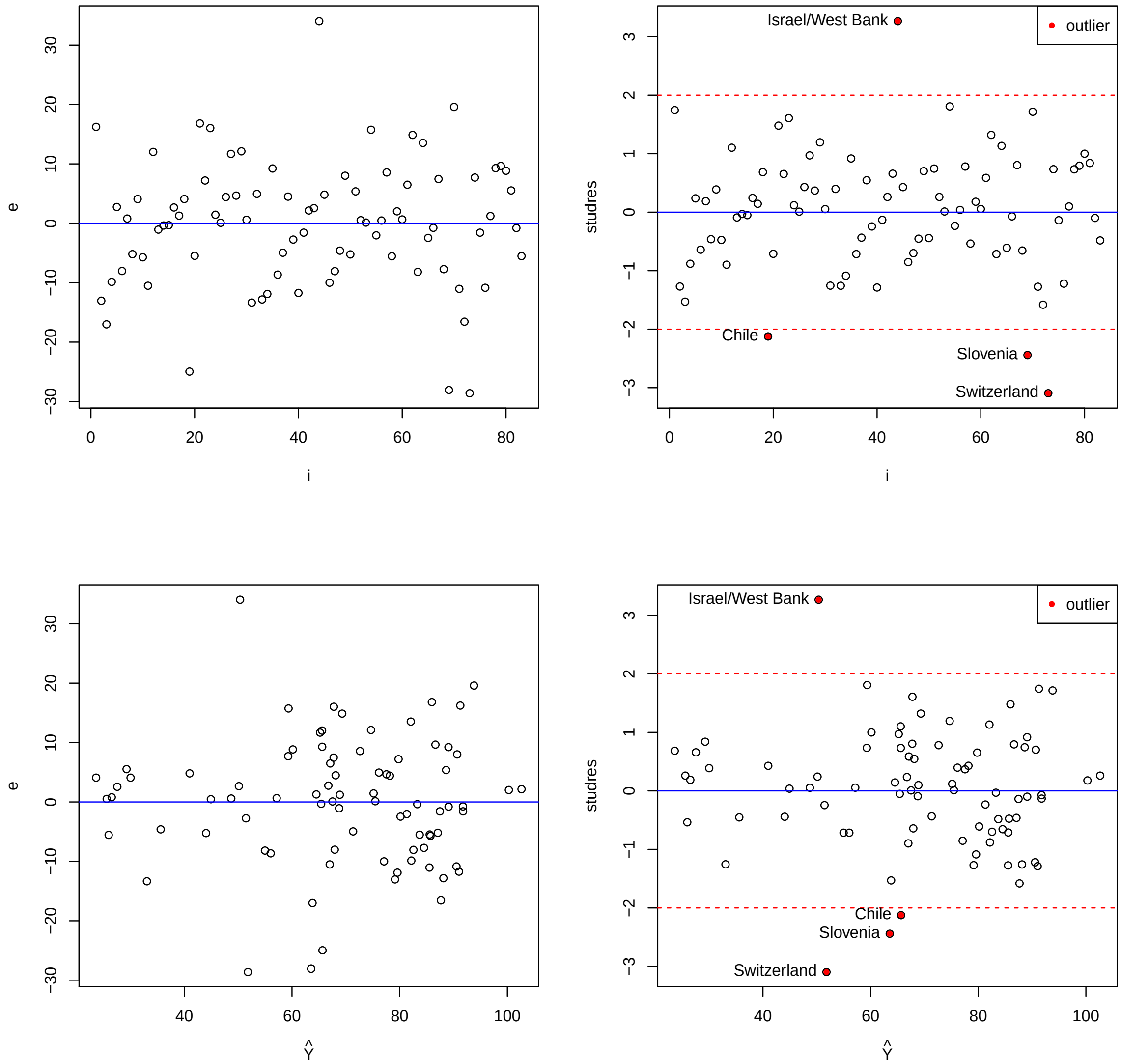


Figure 5: Residual plots and potential outliers

From residual plot $\hat{\epsilon} \sim \hat{Y}$ we can conclude that variance for observations with $\hat{Y} < 50$ is smaller then variance for cases with $\hat{Y} > 50$. Because I assumed independence of errors but it looks that errors are not identically distributed we can try two approaches i.e. try to fit two models for those two groups separately or use generalized linear model like weighted least squares. P-value of F test comparing those two variances equals: 0.008.

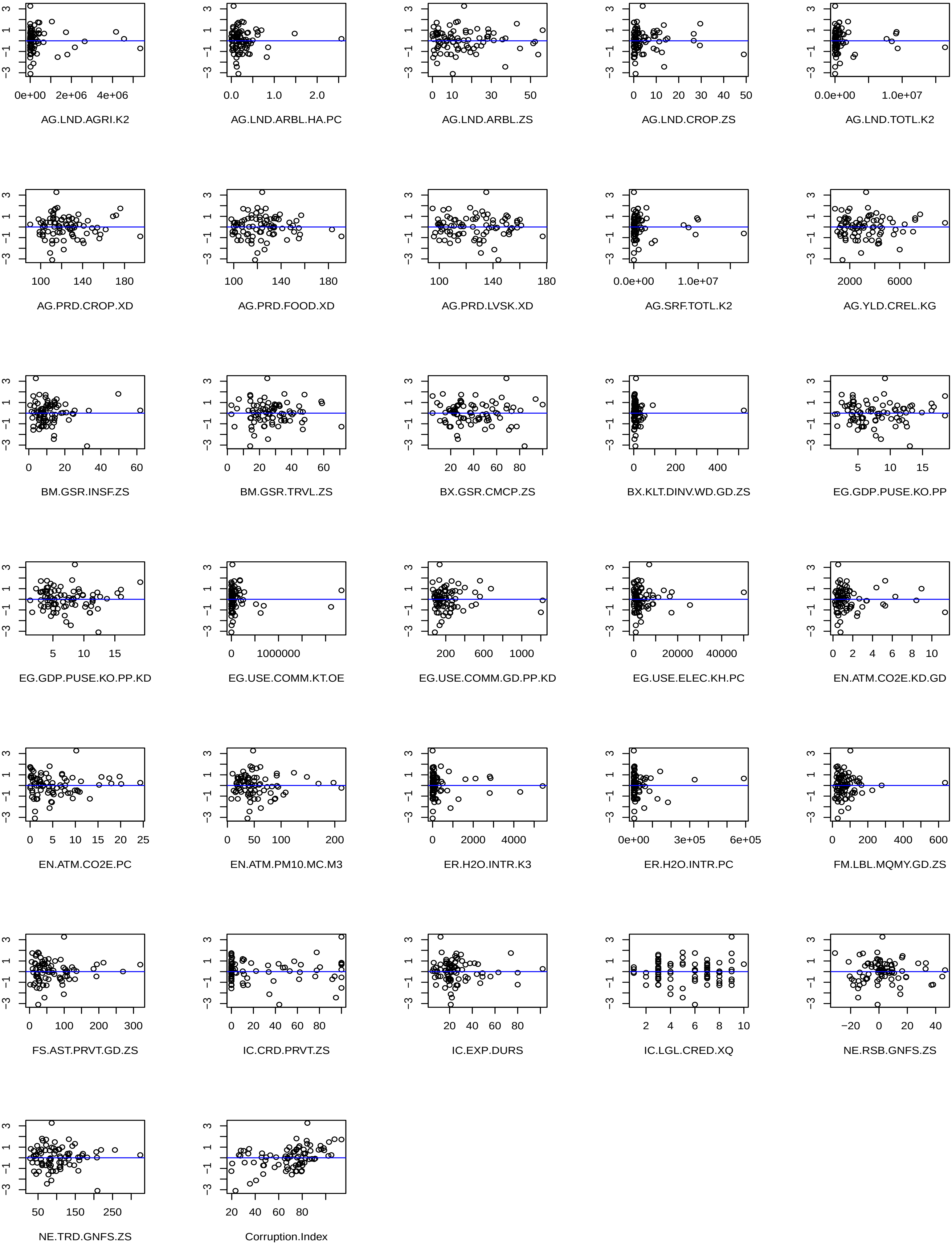


Figure 6: Studentized residual plots for predictors

We can check if there is linear pattern in the errors by fitting the regression line to residuals against fitted values:

```
                     Estimate Std. Error   t value   Pr(>|t|)
(Intercept)        6.29350551  2.8065569 2.2424293 0.02766685
fitted(model.full) 0.02297039  0.0391972 0.5860212 0.55949073
```

The coefficient for the fitted values isn't significant so we can conclude there is no liner relationship in the residuals.

Using partial regression and partial residual plots we can look for transformations of the predictor variables that could be benefitial to the model:

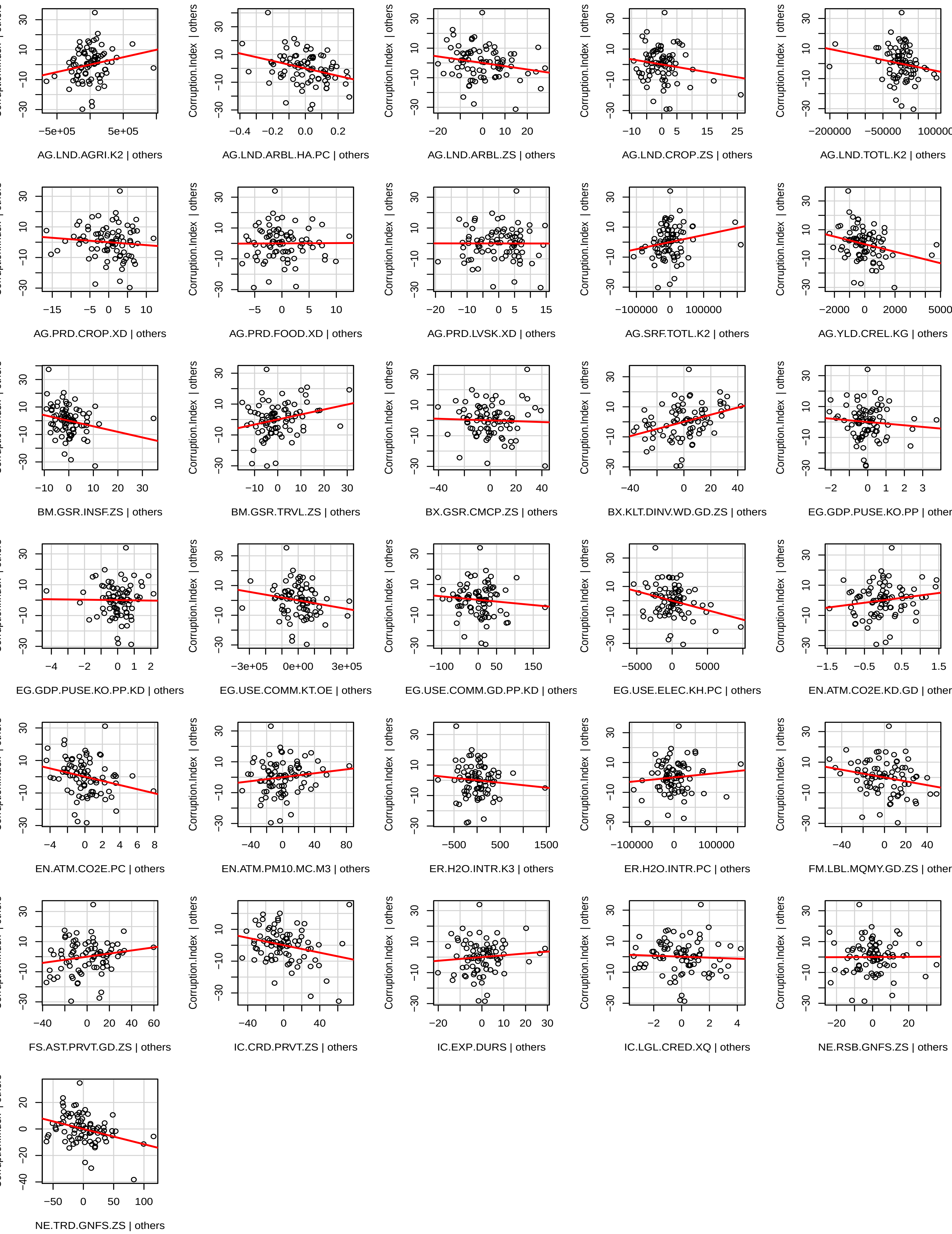


Figure 7: Partial regression plots

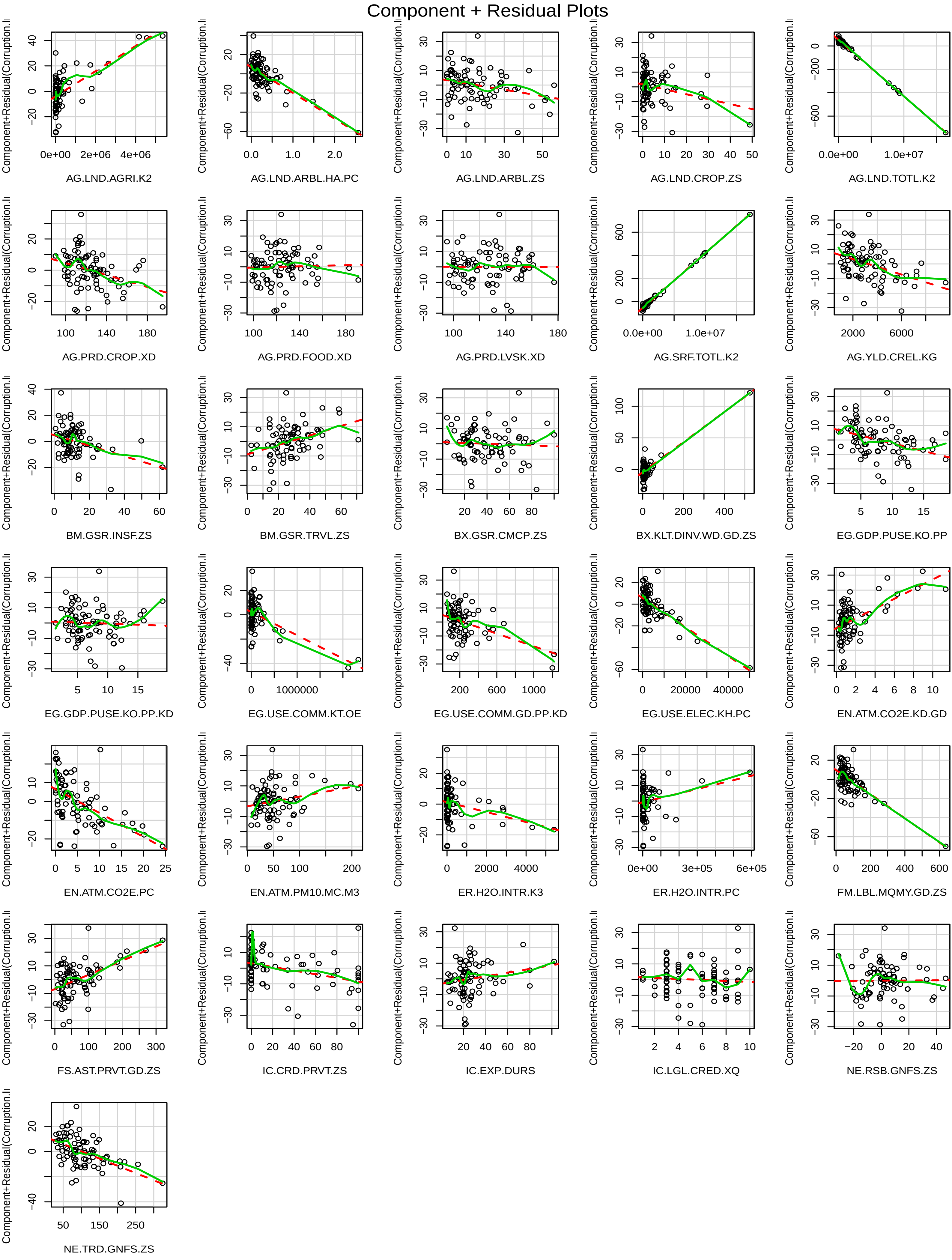


Figure 8: Partial residual plots

## 4.6 Additive impact of each predictor variable on explained variable.

This means that there are no interection effects between predictor variables included in the model.

## 4.7 Collinearity

Columns of experiment matrix have to be alegbraically independent otherwise we might not find solution or the solution is not stable.

The first check is to look at correlation matrix Table-1. The table marks all significant correlations in red font. The following pairs of variables have correlation bigger than 0.8:

```
[1] "AG.SRF.TOTL.K2 AG.LND.TOTL.K2"           "ER.H2O.INTR.K3 AG.LND.TOTL.K2"
[3] "ER.H2O.INTR.K3 AG.SRF.TOTL.K2"           "EG.GDP.PUSE.KO.PP.KD EG.GDP.PUSE.KO.PP"
[5] "EN.ATM.CO2E.KD.GD EG.USE.COMM.GD.PP.KD"
```

Checking coolinearity by computing rank of the matrix:

```
> X<-model.matrix(model.full)
> ncol(X)

[1] 32

> as.integer(rankMatrix(X))

[1] 32
```

We can also check if predictors are not collinear by checking that eigenvalues of the $X'X$ are not close to zero. The convenient way is to use statistic $\kappa = \sqrt{\frac{\lambda_l}{\lambda_p}}$ where $\lambda_l$ is the largest eigenvalue and $\lambda_p$ are other lambda values. Values greater equal than 30 are considered as a problem [3]. As we can see we observe very large values of this statistic:

```
 [1] 6.678691e+07 3.210068e+07 7.978557e+06 5.688988e+06 2.318929e+06 1.781263e+06 1.358323e+06
 [8] 9.400946e+05 6.765676e+05 5.585418e+05 5.268543e+05 5.005712e+05 3.706954e+05 3.510966e+05
[15] 2.801852e+05 2.742668e+05 2.050105e+05 1.746508e+05 1.406485e+05 1.049815e+05 9.558464e+04
[22] 5.262181e+04 3.943306e+04 1.909145e+04 9.664668e+03 1.698878e+03 7.600963e+02 1.026057e+02
[29] 5.196613e+01 2.020892e+01 7.122399e+00 1.000000e+00
```

We can also check for multiple collinearity by computing variance inflation factor (VIF) and removing variables with VIF $\geq 10$. The VIF is $\frac{1}{1-R_i^2}$ where $R_i^2$ is multiple coefficient of determination for the model $X_i \sim X_1 + ... + X_{i-1} + X_{i+1} + X_{p-1}$. The standard error of coefficient $\beta_i$ is proportional to the $\sqrt{VIF_i}$ ($SE_{\hat{\beta}_i} = \sigma\sqrt{VIF_i}\frac{1}{\sqrt{S_{x_ix_i}}}$ [3]). VIF values:

```
> vif(model.full)

      AG.LND.AGRI.K2    AG.LND.ARBL.HA.PC        AG.LND.ARBL.ZS       AG.LND.CROP.ZS
           18.632807             6.706063              1.790606             2.416085
      AG.LND.TOTL.K2       AG.PRD.CROP.XD        AG.PRD.FOOD.XD       AG.PRD.LVSK.XD
         3853.433087            12.846249             22.438777             7.816655
      AG.SRF.TOTL.K2       AG.YLD.CREL.KG        BM.GSR.INSF.ZS       BM.GSR.TRVL.ZS
         3826.562138             2.435269              2.432500             2.241709
   BX.GSR.CMCP.ZS BX.KLT.DINV.WD.GD.ZS     EG.GDP.PUSE.KO.PP EG.GDP.PUSE.KO.PP.KD
            1.863568            11.206991             19.456279            15.394738
EG.USE.COMM.KT.OE EG.USE.COMM.GD.PP.KD     EG.USE.ELEC.KH.PC    EN.ATM.CO2E.KD.GD
           12.309089            17.612909              8.322555            12.706842
      EN.ATM.CO2E.PC    EN.ATM.PM10.MC.M3        ER.H2O.INTR.K3       ER.H2O.INTR.PC
            6.394319             2.696709              8.827993             4.295018
   FM.LBL.MQMY.GD.ZS    FS.AST.PRVT.GD.ZS        IC.CRD.PRVT.ZS          IC.EXP.DURS
           15.285510             9.945347              2.353677             4.255378
      IC.LGL.CRED.XQ       NE.RSB.GNFS.ZS        NE.TRD.GNFS.ZS
            1.974724             1.878799              2.797101
```

We see that we should remove predictors from the model because there exists multicollinearity between them. We remove predictors one by one until all remaining predictors in the model have VIF<10. When all predictors have VID below 10 the model consists of following coefficients:

```
                     Estimate  Std. Error t value Pr(>|t|)
(Intercept)         1.1840e+02  2.0637e+01  5.7372 3.86e-07 ***
AG.LND.AGRI.K2      1.5983e-06  3.9521e-06  0.4044  0.68742
AG.LND.ARBL.HA.PC  -1.2532e+01  7.9489e+00 -1.5766  0.12042
AG.LND.ARBL.ZS     -1.7453e-01  1.3540e-01 -1.2890  0.20262
AG.LND.CROP.ZS     -1.5702e-01  2.5053e-01 -0.6268  0.53332
```

```
AG.PRD.CROP.XD       -1.3246e-01  1.2623e-01 -1.0493  0.29845
AG.PRD.LVSK.XD        3.7008e-02  1.0989e-01  0.3368  0.73753
AG.YLD.CREL.KG       -2.9888e-03  1.1981e-03 -2.4946  0.01553 *
BM.GSR.INSF.ZS       -3.6618e-01  2.1425e-01 -1.7091  0.09287 .
BM.GSR.TRVL.ZS        2.2159e-01  1.4753e-01  1.5020  0.13862
BX.GSR.CMCP.ZS       -4.8897e-02  9.1271e-02 -0.5357  0.59423
BX.KLT.DINV.WD.GD.ZS  1.1236e-01  4.3433e-02  2.5871  0.01226 *
EG.GDP.PUSE.KO.PP    -1.0611e+00  4.9750e-01 -2.1329  0.03725 *
EG.USE.COMM.KT.OE     2.6057e-06  9.1950e-06  0.2834  0.77791
EG.USE.ELEC.KH.PC    -1.0097e-03  5.6903e-04 -1.7745  0.08132 .
EN.ATM.CO2E.KD.GD     1.7046e+00  1.3800e+00  1.2353  0.22180
EN.ATM.CO2E.PC       -1.2679e+00  6.4869e-01 -1.9546  0.05554 .
EN.ATM.PM10.MC.M3     7.9668e-02  5.7451e-02  1.3867  0.17093
ER.H2O.INTR.K3       -1.6113e-04  2.5301e-03 -0.0637  0.94944
ER.H2O.INTR.PC        3.7347e-05  3.7404e-05  0.9985  0.32228
FS.AST.PRVT.GD.ZS    -1.7201e-02  4.5527e-02 -0.3778  0.70697
IC.CRD.PRVT.ZS       -1.0732e-01  5.9733e-02 -1.7967  0.07768 .
IC.EXP.DURS           4.5480e-02  1.5607e-01  0.2914  0.77180
IC.LGL.CRED.XQ       -1.5262e-02  8.1848e-01 -0.0186  0.98519
NE.RSB.GNFS.ZS        3.6597e-02  1.3291e-01  0.2754  0.78404
NE.TRD.GNFS.ZS       -1.0128e-01  4.4178e-02 -2.2925  0.02559 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting model has comparable adjusted $R^2$ =0.75 to the full model. It doesn't explain variance worse than the full model (p-value of F-statistic comparing both models: 0.41). The list of coefficients included in the model selected by the VIF method doesn't contain simultaneously both predictors that correlation is bigger than 0.8 (VIF method caters for pairwise collinearity and additionaly for multiple collinearity). The resulting model has still some coefficients which are not significant.

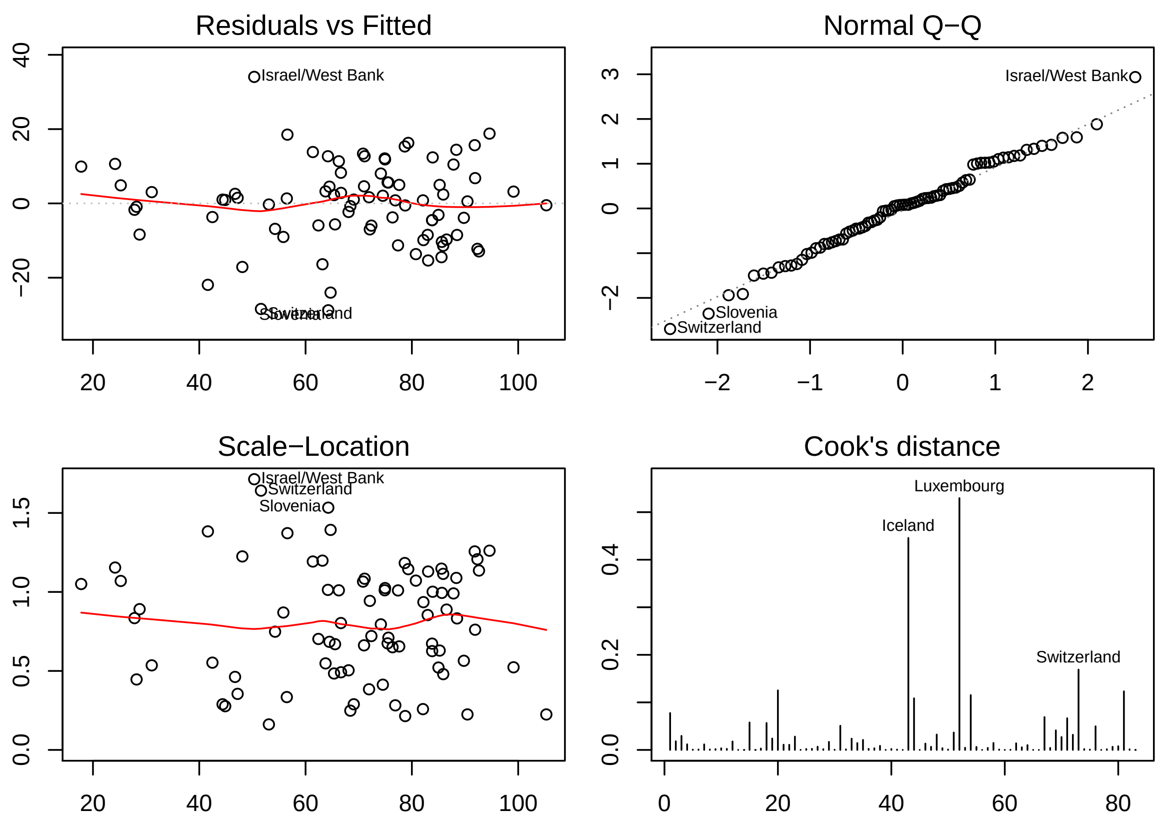


Figure 9: Diagnostic plots for model selected by VIF method

Checking if it would be beneficial to add quadratic terms:

```
[1] FALSE
```

## 4.8 $p \leq n$

```
> length(all.predictor.names)+1
[1] 32
```

```
> nrow(kaggle.data)
```

```
[1] 83
```

### 4.9 is specification of the structural equation of the model correct?

The divergence from the structural assumptions can be check by analyzing plots:

- Residual plots Figure-5, Figure-6,
- Response varaible against each predictor variable Figure-2
- Partial regression Figure-7. This allowes us to visualize relationshop between response variable and specific predictor taking out the effects of other predictors. I cannot see any outstanding nonlinearity there which could have meant that we should change the functional contribution of the predictor into the model.
- Partial residuals Figure-8 is better for spotting nonlinearity. The plots show partial residual plots and two lines i.e. least sqares and nonparametric smooth which allows us to compare linear and curved aproximation of the data. We can see that all sub figures show the linear aproximation is good.

One of the possibilties to deal with the heteroscedasticity of variance in error term is to transform response variable. One of the procedures is Box-Cox transformation, that computes the likelihood of model given transformation of the response variable:

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases}$$

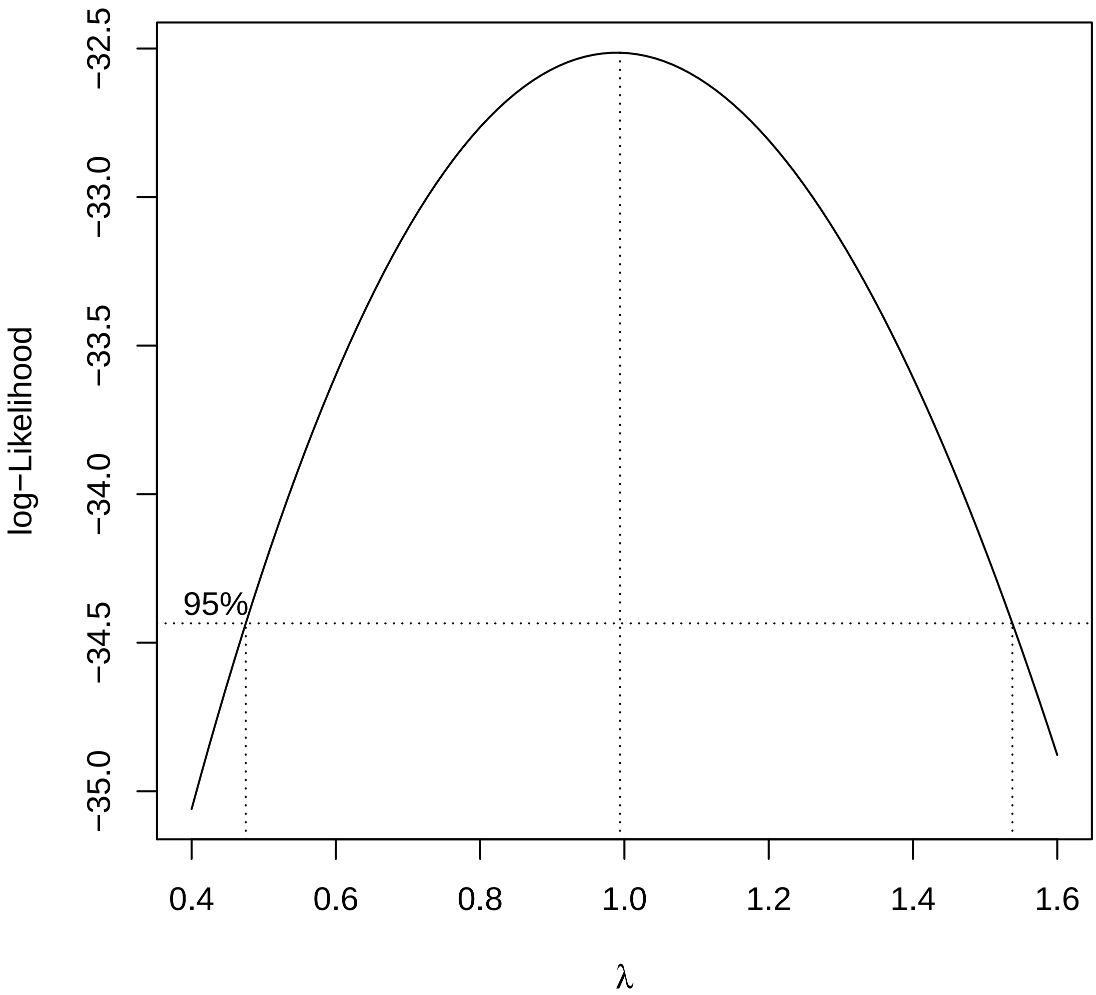


Figure 10: Box-Cox graph showing log-likelihood of data depending on different values $\lambda$

It shows that we can stay with the not transformed response variable because log-likelihood of the the data has maximum near $\lambda = 1$. I am also trying to tranform predictor variables with polynomials for different models in this project.

### 4.10 Dependence of errors

Because this is not a timeseries and it is difficult to order observations in any way I assume that there is no dependence of errors.

### 4.11 Occurance of outliers or influential observations

Potential outliers were shown on the residual plots but we have to remember that influential obvservations can draw regression line towards them and influential observations can be undetected when looking only at outliers (moreover influetial observations don't have to be outliers).
Leverages can be used to find potential influential observations. Using heuristic rule that observation is potentially influential if $h_{ii} \geq \frac{2p}{n}$ we can try to identify them:

```
[1] "Austria"    "Brazil"     "Canada"     "Iceland"    "Luxembourg" "Russia"
```

But we have to remember that $h_i$ depends only on X so even if the observation is far away from the mean(in terms of Mahalanpbis distance) it can still fit into the model. Therefore better tool to diagnose observation as influential is Cook's distance. The outliers for leverages can be also analized on half normal plot on which we see that we shouldn't be worried about unusual leverage values in our data:

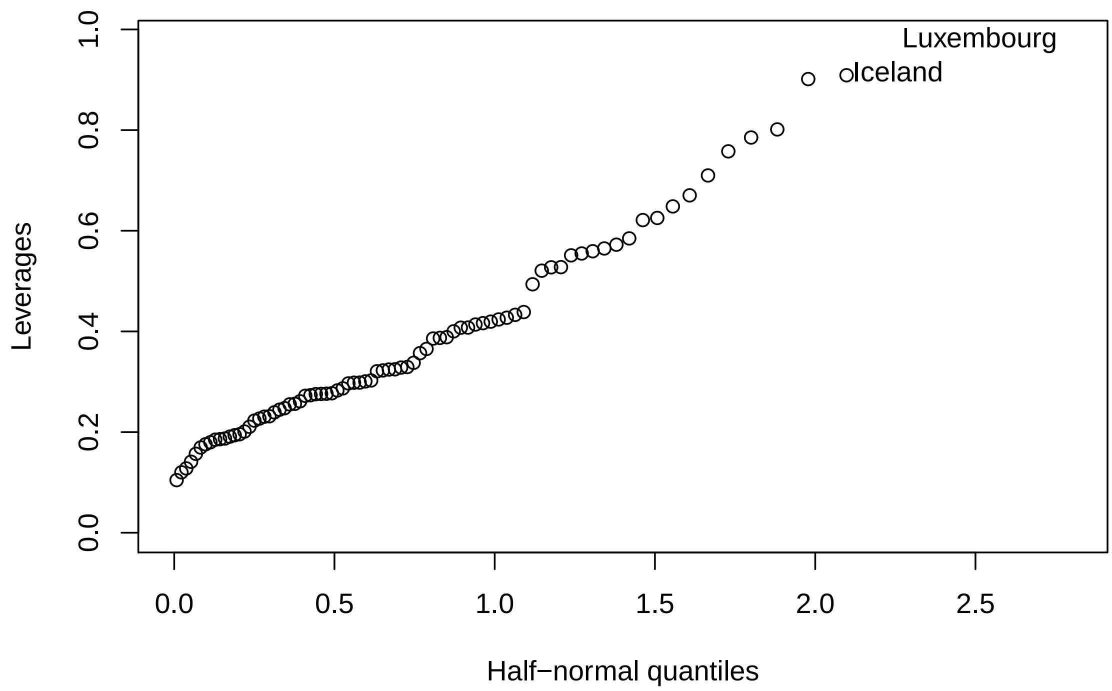


Figure 11: Half normal plot for the leverage values from the full model

Another test for outliers uses jacknife residuals and Bonferroni critical value. The maximum jacknife residual is 3.27 for Israel/West Bank and the Bonferroni critical value is -3.66 so we conclude that it is not an outlier because jacknife residual is less than absolute value of critical value.
To identify influential observations we can look at halfnormal plot of the Cook's distances:

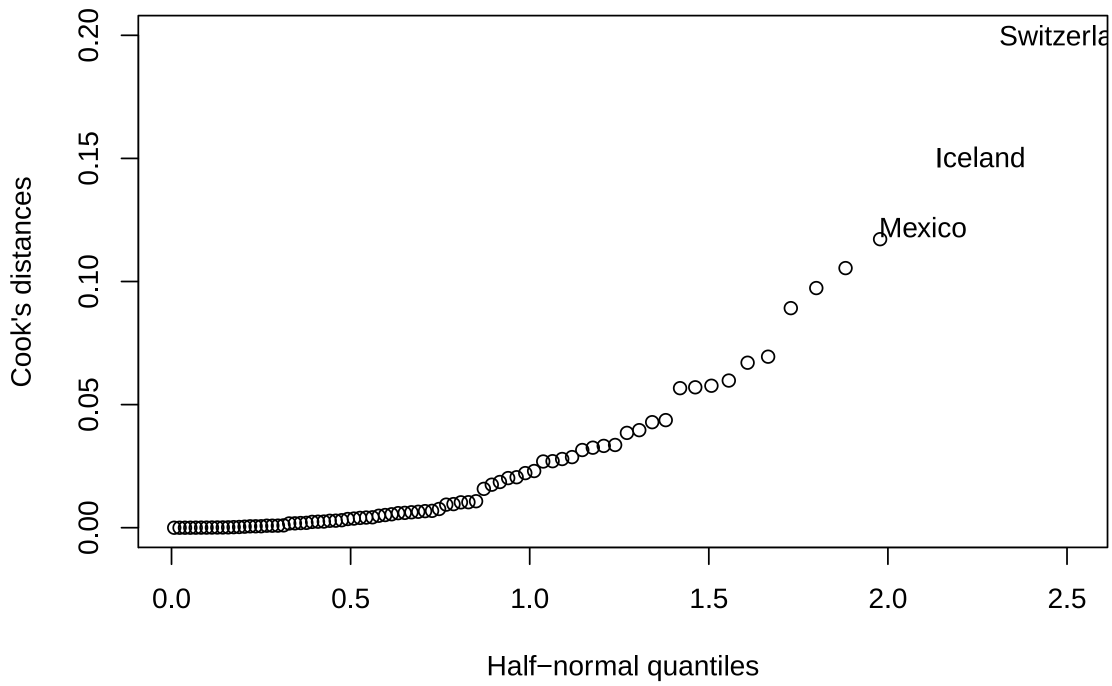


Figure 12: Half normal plot of the Cook statistics

### 4.12 All significant predictors are included

We can be sure that there are some predictors out there (and not included in the provided data set) that would be beneficial to explain the response variable. I am not going to look for them.

## 5 Selection of variables

In this section I will fit different models based basen of different methods for selection of variables:

### 5.1 Backward elimination procedure based on t-tests

Select sub-model (starting from the full model) until all coefficients are significant at $\alpha_{crit}$ =0.2. In each step we remove least significant predictor with the biggest p-value > $\alpha_{crit}$. We have to remember the removing variable from the model doesn't mean that the variable doesn't explain predicted variable. It means that all other variables already in the model have the same information as removed variable. Sometimes for the meaningfulness sake of the model it is better to retain variable even if its coefficient is nonsignificant. The model constructed this way:

```
                        Estimate  Std. Error t value  Pr(>|t|)
(Intercept)           1.2270e+02  1.2967e+01  9.4626 4.968e-14 ***
AG.LND.ARBL.HA.PC    -1.0192e+01  5.0617e+00 -2.0137  0.048005 *
AG.LND.ARBL.ZS       -2.6730e-01  1.0722e-01 -2.4931  0.015104 *
AG.PRD.CROP.XD       -1.4715e-01  9.8167e-02 -1.4990  0.138497
AG.YLD.CREL.KG       -2.2646e-03  9.8107e-04 -2.3083  0.024030 *
BM.GSR.TRVL.ZS        2.9820e-01  1.3472e-01  2.2135  0.030223 *
BX.KLT.DINV.WD.GD.ZS  1.4586e-01  4.8594e-02  3.0017  0.003752 **
EG.GDP.PUSE.KO.PP    -1.2133e+00  4.1566e-01 -2.9190  0.004758 **
EG.USE.ELEC.KH.PC    -7.8406e-04  2.8641e-04 -2.7376  0.007893 **
EN.ATM.CO2E.KD.GD     1.7789e+00  9.6981e-01  1.8343  0.070988 .
EN.ATM.CO2E.PC       -1.2229e+00  4.1895e-01 -2.9191  0.004757 **
EN.ATM.PM10.MC.M3     8.3564e-02  4.2777e-02  1.9535  0.054877 .
FM.LBL.MQMY.GD.ZS    -7.1500e-02  3.7050e-02 -1.9298  0.057801 .
IC.CRD.PRVT.ZS       -1.3469e-01  5.3555e-02 -2.5150  0.014273 *
NE.TRD.GNFS.ZS       -1.1227e-01  3.5577e-02 -3.1559  0.002385 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

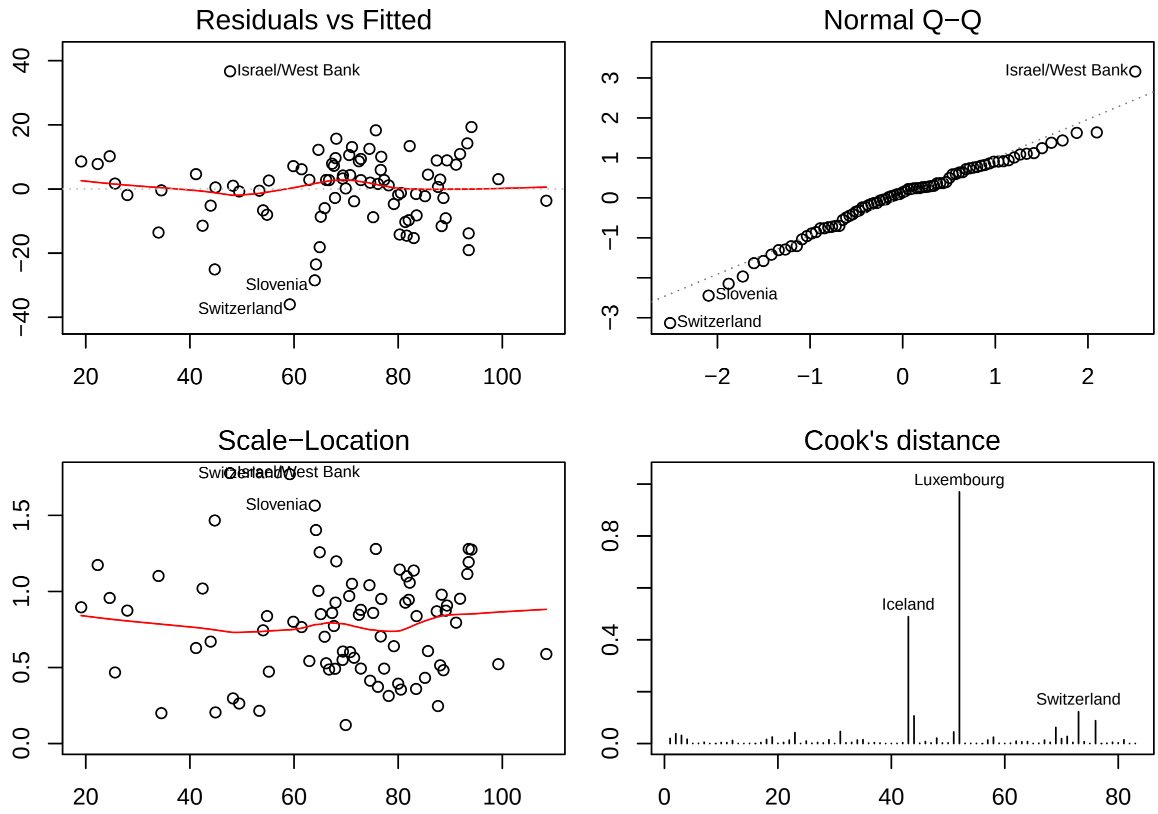


Figure 13: Diagnostic plots for model selected by backward elimination procedure based on t-tests

The final model with all significant predictors is not significantly worse than the full model because p value of F statistic comparing this model to the full model equals 0.93 so we should use the smaller model. Also the selected model explains data better then constant model because p value of F statistics comparing this model to the constant model equals 1.23e-14. Adjusted R-squared: 0.68.

We see that it would be good to reomve Luxembourg because Cook's distance is bigger than one. After removing Luxembourg:

```
> printCoefmat(model.backward.t.test.sum$coefficients)

                         Estimate  Std. Error t value  Pr(>|t|)
(Intercept)            1.2233e+02  1.3016e+01  9.3990 7.425e-14 ***
AG.LND.ARBL.HA.PC     -1.0477e+01  5.0911e+00 -2.0580  0.043487 *
AG.LND.ARBL.ZS        -2.7383e-01  1.0789e-01 -2.5380  0.013479 *
AG.PRD.CROP.XD        -1.4576e-01  9.8488e-02 -1.4800  0.143557
AG.YLD.CREL.KG        -2.3049e-03  9.8553e-04 -2.3388  0.022339 *
BM.GSR.TRVL.ZS         3.1812e-01  1.3764e-01  2.3112  0.023904 *
BX.KLT.DINV.WD.GD.ZS   2.3876e-01  1.3128e-01  1.8187  0.073428 .
EG.GDP.PUSE.KO.PP     -1.2423e+00  4.1868e-01 -2.9671  0.004164 **
EG.USE.ELEC.KH.PC     -8.3881e-04  2.9614e-04 -2.8325  0.006095 **
EN.ATM.CO2E.KD.GD      1.7941e+00  9.7302e-01  1.8439  0.069625 .
EN.ATM.CO2E.PC        -1.1987e+00  4.2145e-01 -2.8441  0.005900 **
EN.ATM.PM10.MC.M3      8.3616e-02  4.2909e-02  1.9487  0.055524 .
FM.LBL.MQMY.GD.ZS     -6.6861e-02  3.7660e-02 -1.7754  0.080376 .
IC.CRD.PRVT.ZS        -1.2987e-01  5.4093e-02 -2.4009  0.019138 *
NE.TRD.GNFS.ZS        -1.2466e-01  3.9211e-02 -3.1791  0.002237 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

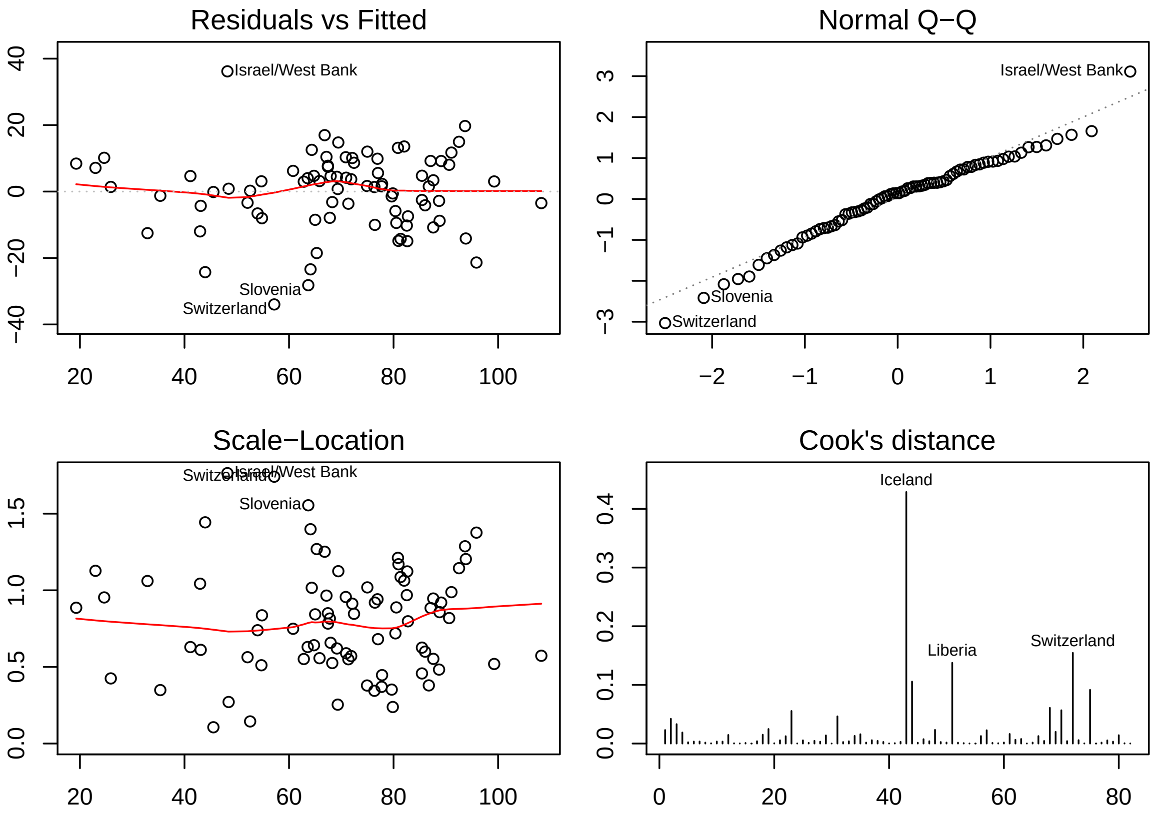


Figure 14: Diagnostic plots for model selected by backward elimination procedure based on t-tests after removing Luxembourg

Adjusted R-squared: 0.67. We see also now that there is no need to remove any more data.
There is a benefit in adding quadratic term for EG.GDP.PUSE.KO.PP:

```
                      Estimate  Std. Error t value  Pr(>|t|)
(Intercept)          1.4520e+02 1.5514e+01  9.3592 1.005e-13 ***
AG.LND.ARBL.HA.PC   -1.2474e+01 4.9671e+00 -2.5112 0.0144866 *
AG.LND.ARBL.ZS      -3.2730e-01 1.0607e-01 -3.0856 0.0029673 **
AG.PRD.CROP.XD      -1.8056e-01 9.5856e-02 -1.8837 0.0640119 .
AG.YLD.CREL.KG      -2.1606e-03 9.5078e-04 -2.2725 0.0263197 *
BM.GSR.TRVL.ZS       3.7830e-01 1.3471e-01  2.8082 0.0065450 **
BX.KLT.DINV.WD.GD.ZS 3.3713e-01 1.3240e-01  2.5463 0.0132283 *
EG.GDP.PUSE.KO.PP   -6.1343e+00 1.9974e+00 -3.0711 0.0030963 **
EG.USE.ELEC.KH.PC   -1.1618e-03 3.1306e-04 -3.7111 0.0004258 ***
EN.ATM.CO2E.KD.GD    1.0369e+00 9.8470e-01  1.0531 0.2961548
EN.ATM.CO2E.PC      -9.8227e-01 4.1497e-01 -2.3671 0.0208693 *
EN.ATM.PM10.MC.M3    6.0344e-02 4.2355e-02  1.4247 0.1589460
FM.LBL.MQMY.GD.ZS   -5.4510e-02 3.6600e-02 -1.4893 0.1411593
IC.CRD.PRVT.ZS      -9.1053e-02 5.4353e-02 -1.6752 0.0986244 .
NE.TRD.GNFS.ZS      -1.2991e-01 3.7818e-02 -3.4351 0.0010287 **
I(EG.GDP.PUSE.KO.PP^2) 2.4597e-01 9.8364e-02  2.5006 0.0148894 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted R-squared: 0.69. Comparing the model with quadratic term to model without it results in significant difference with p-value 0.015.

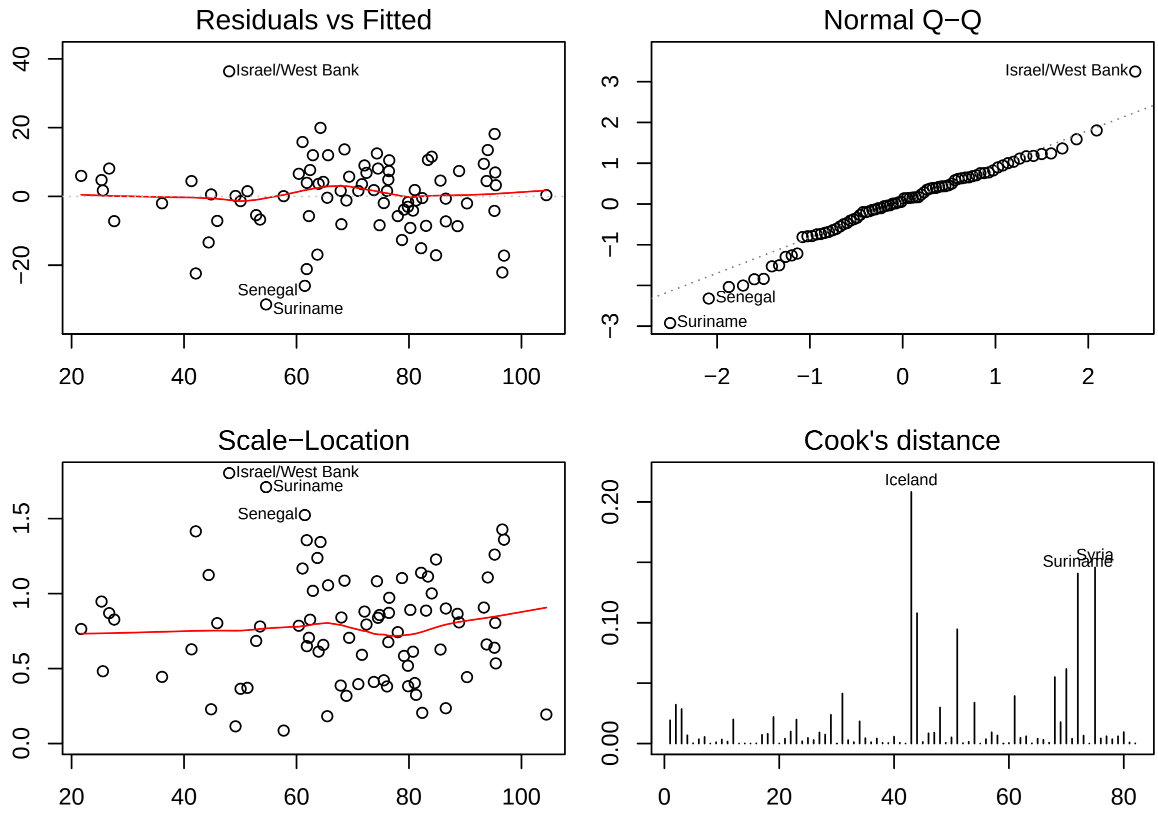


Figure 15: Diagnostic plots for model selected by backward elimination procedure based on t-tests after removing Luxembourg and with quadratic term on EG.GDP.PUSE.KO.PP

## 5.2 Forward selection procedure based on t-tests

Starting from the constant model select extended model iteratively as far as all coefficients are significant at $\alpha = 0.1$, the selected model:

```
                      Estimate  Std. Error t value  Pr(>|t|)
(Intercept)          95.87541621  7.40806456 12.9420 < 2.2e-16 ***
EN.ATM.CO2E.PC       -1.24009647  0.35510551 -3.4922 0.0008068 ***
IC.EXP.DURS           0.24747231  0.10384817  2.3830 0.0197042 *
IC.CRD.PRVT.ZS       -0.17788245  0.05226977 -3.4032 0.0010718 **
EG.USE.ELEC.KH.PC    -0.00089759  0.00028492 -3.1503 0.0023430 **
NE.TRD.GNFS.ZS       -0.07498509  0.03106866 -2.4135 0.0182402 *
EG.GDP.PUSE.KO.PP.KD -1.09084464  0.45203438 -2.4132 0.0182559 *
AG.LND.ARBL.ZS       -0.21287596  0.10853296 -1.9614 0.0535450 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

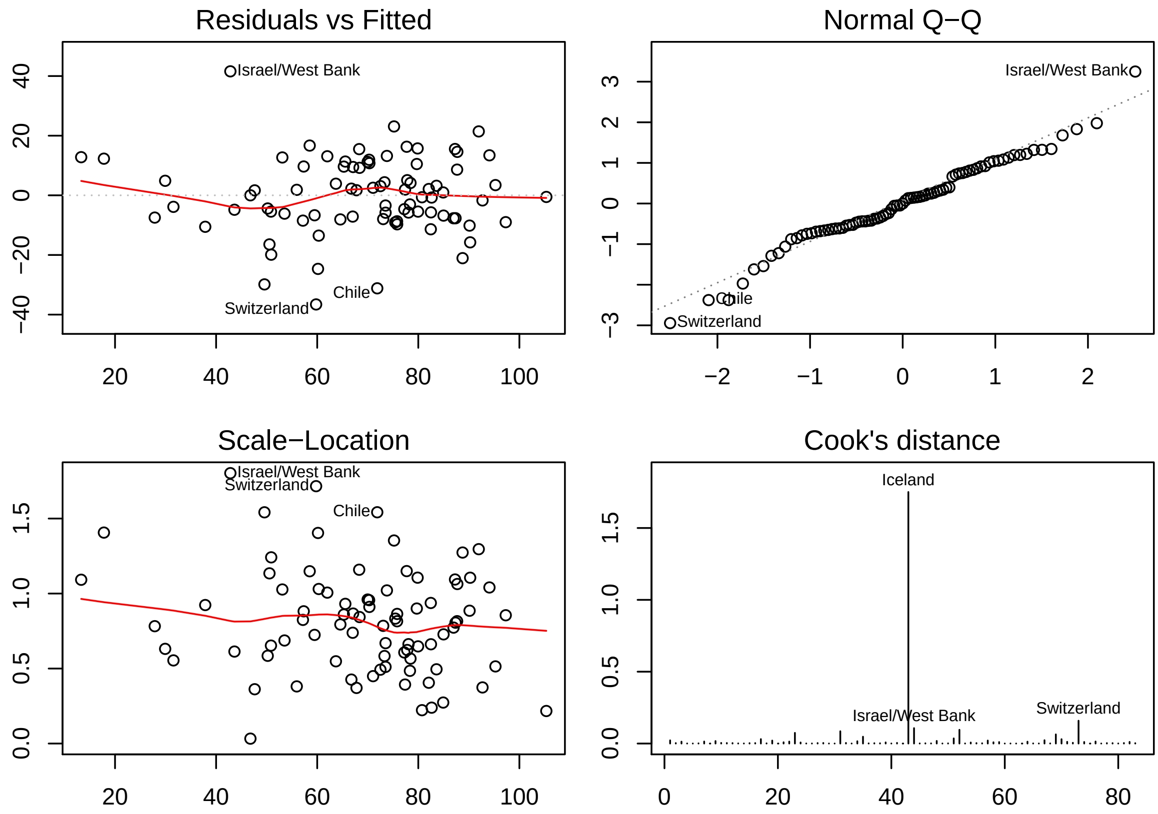


Figure 16: Diagnostic plots for model selected by forward selection procedure based on t-tests

The selected model by the forward procedure is even smaller than the model selected by the backwards procedure but is still insignificantly worse than the full model because p value of F statistic comparing this model to the full model equals 0.5. We cannot compare directly (using for example F-test) the models created by the forward and backward procedures because none of them is the subset of the other one. Also the selected model explains data better then constant model because p value of F statistics comparing this model to the constant model equals 8.25e-16

We see that it would be good to reomve Iceland because Cook's distance is bigger than one. After removing Icleland:

```
                        Estimate   Std. Error t value  Pr(>|t|)
(Intercept)          96.23855689  7.26302874 13.2505 < 2.2e-16 ***
EN.ATM.CO2E.PC       -0.60477410  0.46928872 -1.2887  0.201516
IC.EXP.DURS           0.22894628  0.10219686  2.2402  0.028077 *
IC.CRD.PRVT.ZS       -0.16972368  0.05138995 -3.3027  0.001477 **
EG.USE.ELEC.KH.PC    -0.00188657  0.00056401 -3.3449  0.001295 **
NE.TRD.GNFS.ZS       -0.07894455  0.03051418 -2.5871  0.011640 *
EG.GDP.PUSE.KO.PP.KD -1.02322531  0.44431344 -2.3029  0.024095 *
AG.LND.ARBL.ZS       -0.21884654  0.10641656 -2.0565  0.043260 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

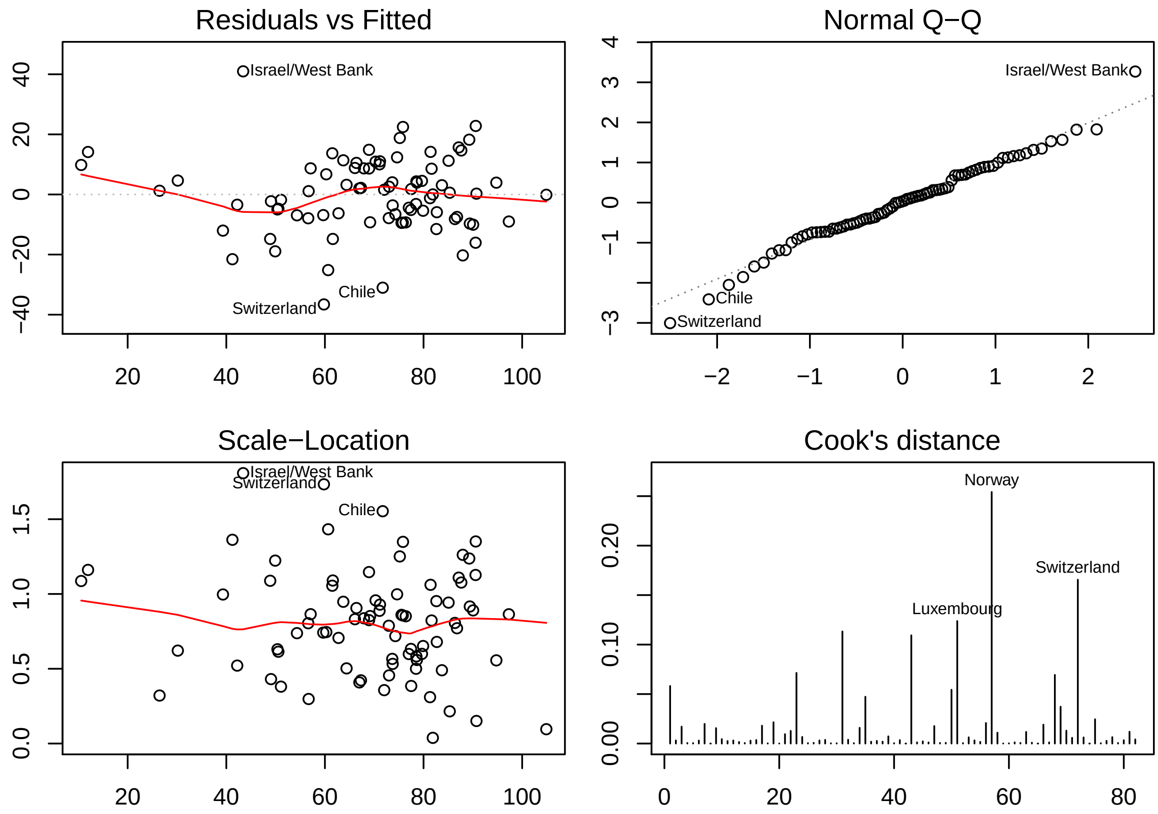


Figure 17: Diagnostic plots for model selected by forward selection procedure based on t-tests after removing Iceland

Adjusted R-squared: 0.67. We see also now that there is no need to remove any more data.
Checking if it would be beneficial to add quadratic terms:

```
[1] FALSE
```

## 5.3 Backward based on AIC criterion

The selected model:

```
                       Estimate  Std. Error t value  Pr(>|t|)
(Intercept)           1.2919e+02  1.3351e+01  9.6767 2.776e-14 ***
AG.LND.ARBL.HA.PC    -1.1068e+01  5.0564e+00 -2.1890  0.032141 *
AG.LND.ARBL.ZS       -2.8288e-01  1.0695e-01 -2.6450  0.010200 *
AG.PRD.CROP.XD       -1.8762e-01  9.9964e-02 -1.8769  0.064949 .
AG.YLD.CREL.KG       -2.6056e-03  9.9531e-04 -2.6179  0.010960 *
BM.GSR.INSF.ZS       -2.3762e-01  1.8647e-01 -1.2743  0.207022
BM.GSR.TRVL.ZS        3.4640e-01  1.3973e-01  2.4790  0.015738 *
BX.KLT.DINV.WD.GD.ZS  2.1704e-01  6.5934e-02  3.2918  0.001601 **
EG.GDP.PUSE.KO.PP    -1.1943e+00  4.1326e-01 -2.8899  0.005210 **
EG.USE.ELEC.KH.PC    -1.1414e-03  3.7093e-04 -3.0771  0.003042 **
EN.ATM.CO2E.KD.GD     1.8800e+00  9.6871e-01  1.9407  0.056569 .
EN.ATM.CO2E.PC       -1.2294e+00  4.1858e-01 -2.9371  0.004558 **
EN.ATM.PM10.MC.M3     9.8251e-02  4.3246e-02  2.2719  0.026358 *
FM.LBL.MQMY.GD.ZS    -1.1943e-01  5.8365e-02 -2.0463  0.044719 *
FS.AST.PRVT.GD.ZS     7.4541e-02  5.7148e-02  1.3043  0.196645
IC.CRD.PRVT.ZS       -1.3302e-01  5.4393e-02 -2.4455  0.017140 *
NE.TRD.GNFS.ZS       -1.1798e-01  3.5523e-02 -3.3213  0.001463 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

optimizes criterion AIC = -2max loglikelihood + 2p.

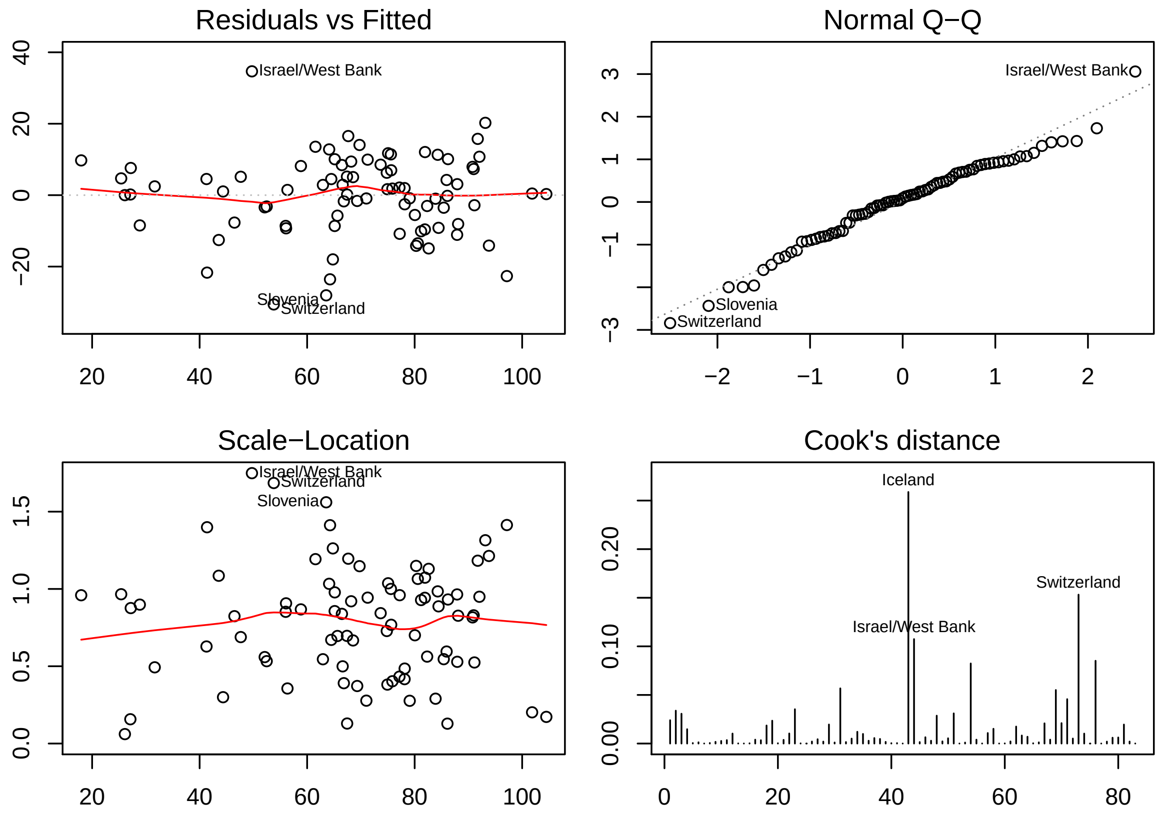


Figure 18: Diagnostic plots for model selected by Backward based on AIC criterion

The final model is not significantly worse than the full model because p value of F statistic comparing this model to the full model equals 0.97 so we should use the smaller model. Also the selected model explains data better then constant model because p value of F statistics comparing this model to the constant model equals 3.92e-14.

Adjusted R-squared: 0.69.

Checking if it would be beneficial to add quadratic terms:

```
[1] FALSE
```

## 5.4 Forward based on AIC criterion

The selected model:

```
                         Estimate  Std. Error t value  Pr(>|t|)
(Intercept)            1.0335e+02  9.0541e+00 11.4144 < 2.2e-16 ***
EN.ATM.CO2E.PC        -5.1420e-01  4.9519e-01 -1.0384  0.302767
IC.EXP.DURS            1.2524e-01  1.0981e-01  1.1405  0.258072
IC.CRD.PRVT.ZS        -1.2962e-01  5.3329e-02 -2.4305  0.017721 *
EG.USE.ELEC.KH.PC     -1.3578e-03  4.1751e-04 -3.2522  0.001784 **
NE.TRD.GNFS.ZS        -1.1061e-01  3.5546e-02 -3.1116  0.002720 **
EG.GDP.PUSE.KO.PP.KD -1.3995e+00  4.4945e-01 -3.1138  0.002702 **
AG.LND.ARBL.ZS        -1.7859e-01  1.1298e-01 -1.5808  0.118573
AG.YLD.CREL.KG        -2.4845e-03  1.0252e-03 -2.4234  0.018045 *
AG.LND.ARBL.HA.PC     -1.2453e+01  5.2381e+00 -2.3773  0.020258 *
BM.GSR.TRVL.ZS         2.4588e-01  1.2025e-01  2.0448  0.044751 *
BX.KLT.DINV.WD.GD.ZS   1.0403e-01  4.6455e-02  2.2394  0.028401 *
ER.H2O.INTR.PC         4.4463e-05  3.0713e-05  1.4477  0.152290
EN.ATM.PM10.MC.M3      5.9475e-02  4.2278e-02  1.4067  0.164059
FM.LBL.MQMY.GD.ZS     -4.9525e-02  3.5719e-02 -1.3865  0.170111
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

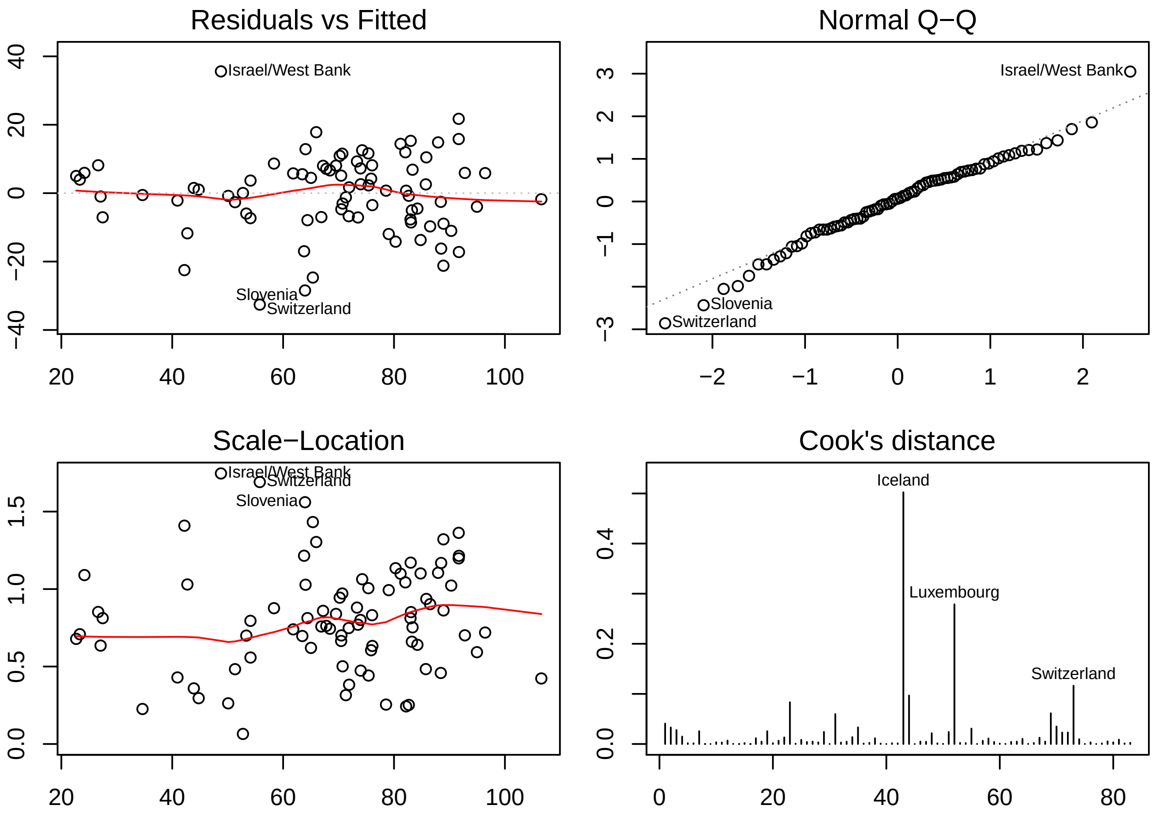


Figure 19: Diagnostic plots for model selected by Forward based on AIC criterion

The final model is not significantly worse than the full model because p value of F statistic comparing this model to the full model equals 0.91 so we should use the smaller model. Also the selected model explains data better then constant model because p value of F statistics comparing this model to the constant model equals 1.62e-14.

Adjusted R-squared: 0.68.

There is a benefit in adding quadratic term for EG.GDP.PUSE.KO.PP.KD:

```
                            Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                1.2335e+02 1.1485e+01 10.7396  3.29e-16 ***
EN.ATM.CO2E.PC            -5.4082e-01 4.7460e-01 -1.1395 0.2585434
IC.EXP.DURS                2.6164e-02 1.1163e-01  0.2344 0.8154081
IC.CRD.PRVT.ZS            -9.2974e-02 5.2928e-02 -1.7566 0.0835525 .
EG.USE.ELEC.KH.PC         -1.4664e-03 4.0214e-04 -3.6465 0.0005205 ***
NE.TRD.GNFS.ZS            -1.1005e-01 3.4062e-02 -3.2308 0.0019146 **
EG.GDP.PUSE.KO.PP.KD      -6.1690e+00 1.8458e+00 -3.3421 0.0013634 **
AG.LND.ARBL.ZS            -2.4154e-01 1.1082e-01 -2.1796 0.0328041 *
AG.YLD.CREL.KG            -2.5461e-03 9.8265e-04 -2.5910 0.0117322 *
AG.LND.ARBL.HA.PC         -1.4515e+01 5.0788e+00 -2.8579 0.0056773 **
BM.GSR.TRVL.ZS             2.7525e-01 1.1575e-01  2.3779 0.0202704 *
BX.KLT.DINV.WD.GD.ZS       1.0340e-01 4.4514e-02  2.3228 0.0232372 *
ER.H2O.INTR.PC             3.5281e-05 2.9631e-05  1.1907 0.2379839
EN.ATM.PM10.MC.M3          6.2571e-02 4.0528e-02  1.5439 0.1273261
FM.LBL.MQMY.GD.ZS         -3.6719e-02 3.4563e-02 -1.0624 0.2918841
I(EG.GDP.PUSE.KO.PP.KD^2)  2.5284e-01 9.5150e-02  2.6572 0.0098404 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted R-squared: 0.71. Comparing the model with quadratic term to model without it results in significant difference with p-value 0.01.

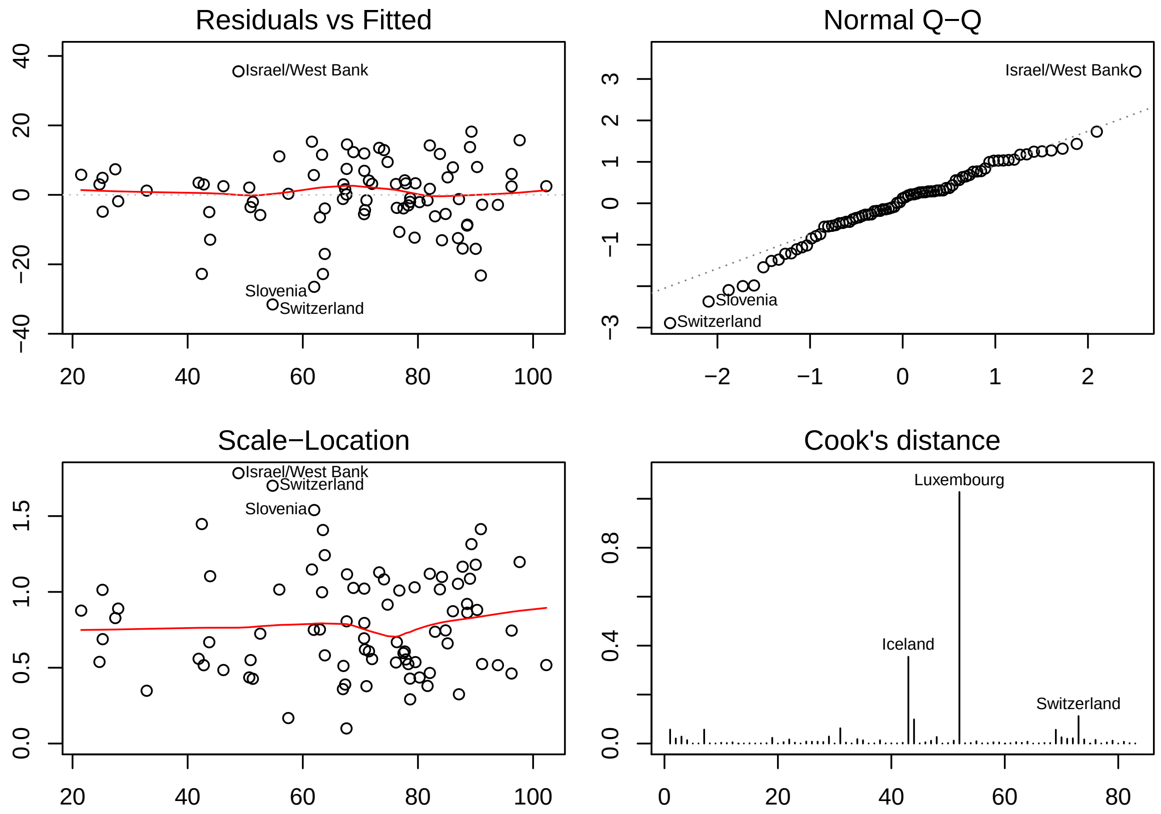


Figure 20: Diagnostic plots for model selected by Forward based on AIC criterion and with quadratic term on EG.GDP.PUSE.KO.PP.KD

We see that it would be good to reomve Luxembourg because Cook's distance is bigger than one. After removing Luxembourg:

```
                             Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                1.2282e+02  1.1537e+01 10.6459 5.762e-16 ***
EN.ATM.CO2E.PC            -4.8498e-01  4.8116e-01 -1.0079 0.3171659
IC.EXP.DURS                3.7611e-02  1.1288e-01  0.3332 0.7400416
IC.CRD.PRVT.ZS            -8.6694e-02  5.3669e-02 -1.6153 0.1110073
EG.USE.ELEC.KH.PC         -1.5442e-03  4.1511e-04 -3.7199 0.0004138 ***
NE.TRD.GNFS.ZS            -1.2298e-01  3.7880e-02 -3.2467 0.0018364 **
EG.GDP.PUSE.KO.PP.KD      -6.3205e+00  1.8610e+00 -3.3964 0.0011606 **
AG.LND.ARBL.ZS            -2.4692e-01  1.1134e-01 -2.2177 0.0300215 *
AG.YLD.CREL.KG            -2.5629e-03  9.8565e-04 -2.6003 0.0114841 *
AG.LND.ARBL.HA.PC         -1.4813e+01  5.1072e+00 -2.9005 0.0050562 **
BM.GSR.TRVL.ZS             2.9628e-01  1.1909e-01  2.4878 0.0153860 *
BX.KLT.DINV.WD.GD.ZS       1.9754e-01  1.2727e-01  1.5521 0.1254123
ER.H2O.INTR.PC             3.7101e-05  2.9804e-05  1.2448 0.2175982
EN.ATM.PM10.MC.M3          6.1596e-02  4.0661e-02  1.5149 0.1345820
FM.LBL.MQMY.GD.ZS         -3.1198e-02  3.5358e-02 -0.8823 0.3807913
I(EG.GDP.PUSE.KO.PP.KD^2)  2.6325e-01  9.6325e-02  2.7329 0.0080475 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

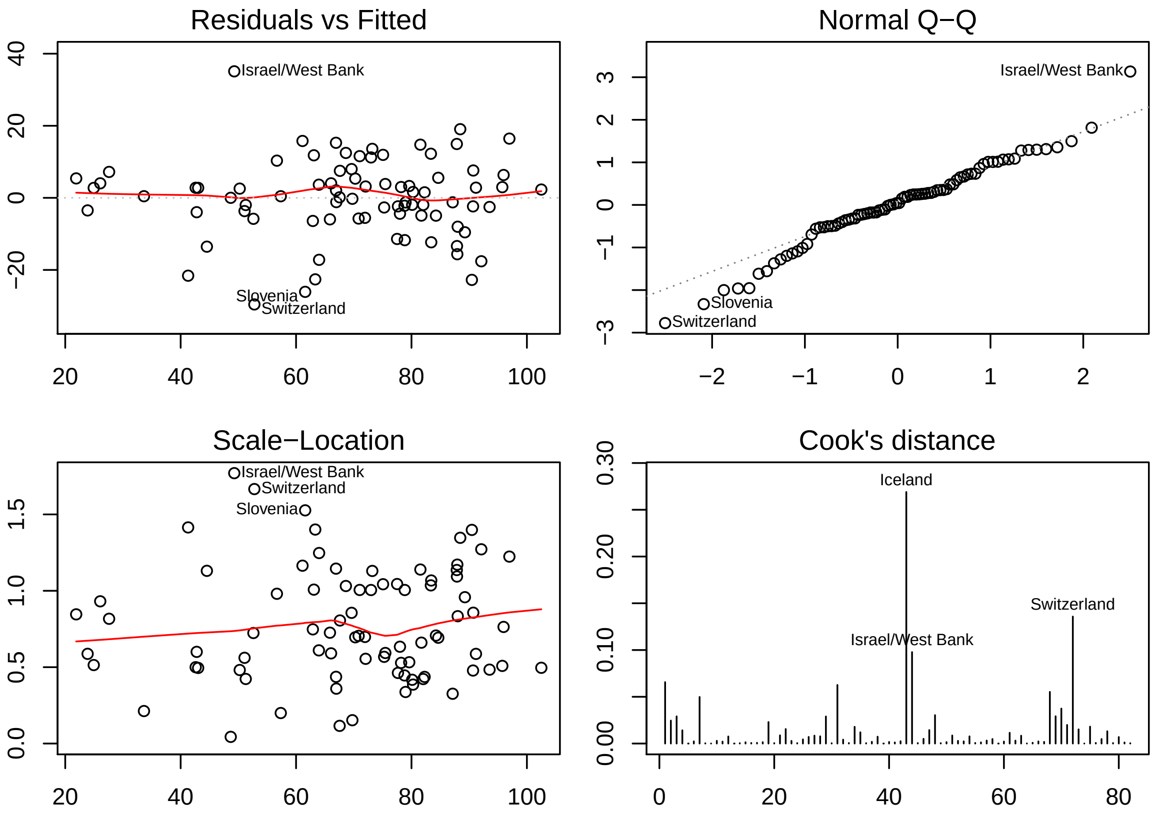


Figure 21: Diagnostic plots for model selected by Forward based on AIC criterion after adding quadratic term and removing Luxembourg

Adjusted R-squared: 0.69. We see also now that there is no need to remove any more data.

## 5.5 Backward based on BIC criterion

The selected model:

```
                    Estimate   Std. Error t value  Pr(>|t|)
(Intercept)       94.84170805  5.24636179 18.0776 < 2.2e-16 ***
AG.YLD.CREL.KG    -0.00212845  0.00099221 -2.1452 0.0351339 *
EG.GDP.PUSE.KO.PP -1.49648844  0.37739030 -3.9654 0.0001644 ***
EG.USE.ELEC.KH.PC -0.00065991  0.00029538 -2.2341 0.0284172 *
EN.ATM.CO2E.PC    -1.45618044  0.35431882 -4.1098 9.909e-05 ***
EN.ATM.PM10.MC.M3  0.11828346  0.04099303  2.8855 0.0050843 **
IC.CRD.PRVT.ZS    -0.12563130  0.04774311 -2.6314 0.0102920 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

is smaller than the one selected by AIC criterion because BIC more havily penalizes bigger models BIC = -2max loglikelihood + plog(n) (where log(n) in our case = 4.42).

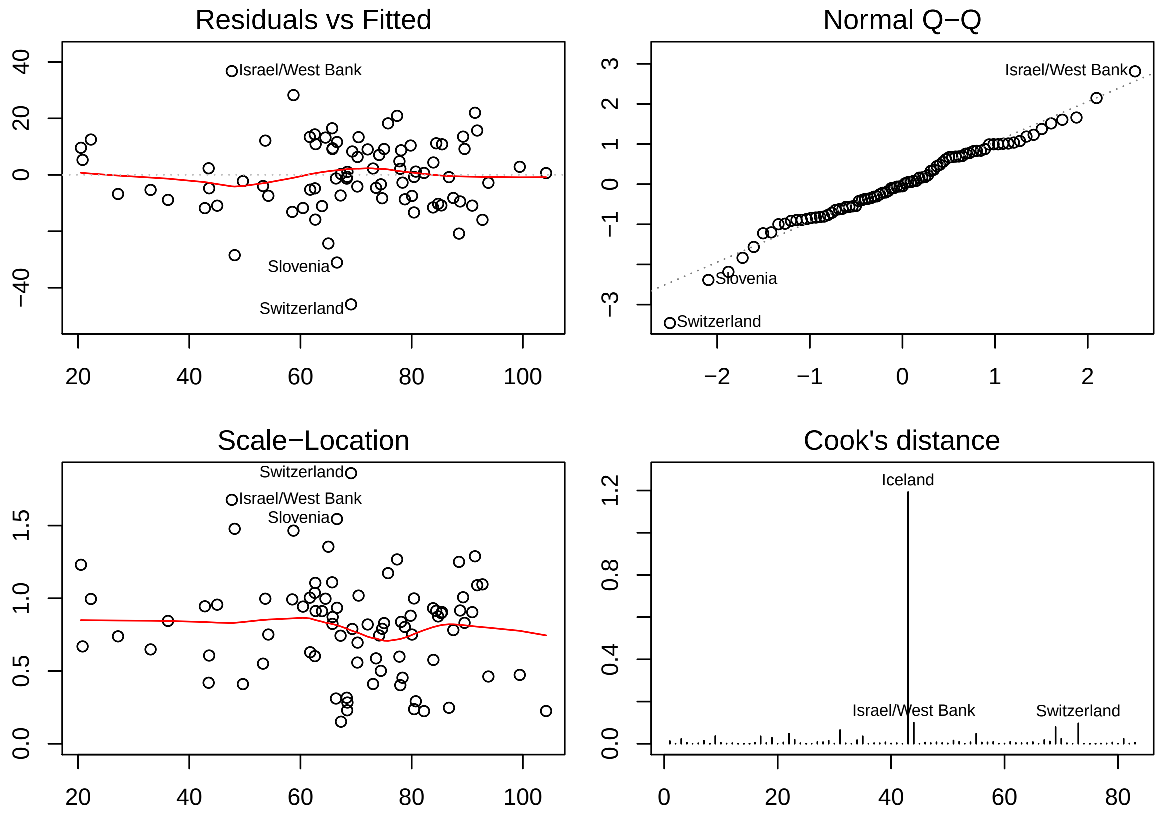


Figure 22: Diagnostic plots for model selected by Backward based on BIC criterion

The final model is not significantly worse than the full model because p value of F statistic comparing this model to the full model equals 0.34 so we should use the smaller model. Also the selected model explains data better then constant model because p value of F statistics comparing this model to the constant model equals 1.56e-15. Adjusted R-squared: 0.62.

We see that it would be good to reomve Iceland because Cook's distance is bigger than one. After removing Icleland:

```
> printCoefmat(model.bckwrd.bic.sum$coefficients)

                    Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)       94.43161723  5.20782714 18.1326 < 2.2e-16 ***
AG.YLD.CREL.KG    -0.00214356  0.00098366 -2.1792 0.0324590 *  
EG.GDP.PUSE.KO.PP -1.41121993  0.37826030 -3.7308 0.0003688 ***
EG.USE.ELEC.KH.PC -0.00144336  0.00059050 -2.4443 0.0168624 *  
EN.ATM.CO2E.PC    -0.95117427  0.48231715 -1.9721 0.0522863 .  
EN.ATM.PM10.MC.M3  0.11112888  0.04090677  2.7166 0.0081838 ** 
IC.CRD.PRVT.ZS    -0.11694075  0.04767003 -2.4531 0.0164850 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

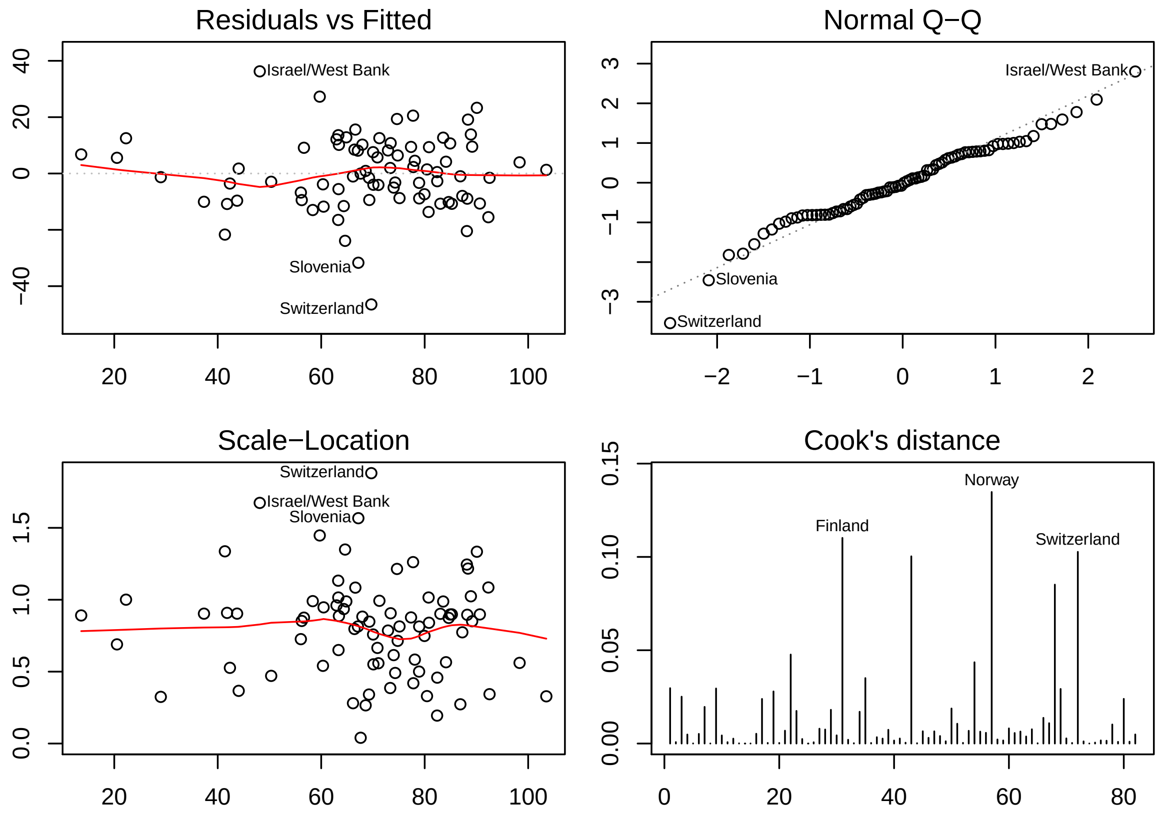


Figure 23: Diagnostic plots for model selected by Backward based on BIC criterion after removing Iceland

Adjusted R-squared: 0.62. We see also now that there is no need to remove any more data.
Checking if it would be beneficial to add quadratic terms:

```
[1] FALSE
```

## 5.6 Forward based on BIC criterion

The selected model:

```
                       Estimate  Std. Error t value  Pr(>|t|)
(Intercept)         91.56208614  7.20541551 12.7074 < 2.2e-16 ***
EN.ATM.CO2E.PC      -1.26113093  0.36153075 -3.4883 0.0008118 ***
IC.EXP.DURS          0.25315374  0.10573428  2.3942 0.0191202 *
IC.CRD.PRVT.ZS      -0.18368520  0.05315447 -3.4557 0.0009015 ***
EG.USE.ELEC.KH.PC   -0.00080268  0.00028599 -2.8067 0.0063559 **
NE.TRD.GNFS.ZS      -0.07133965  0.03158858 -2.2584 0.0267895 *
EG.GDP.PUSE.KO.PP.KD -1.03600146  0.45954177 -2.2544 0.0270501 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

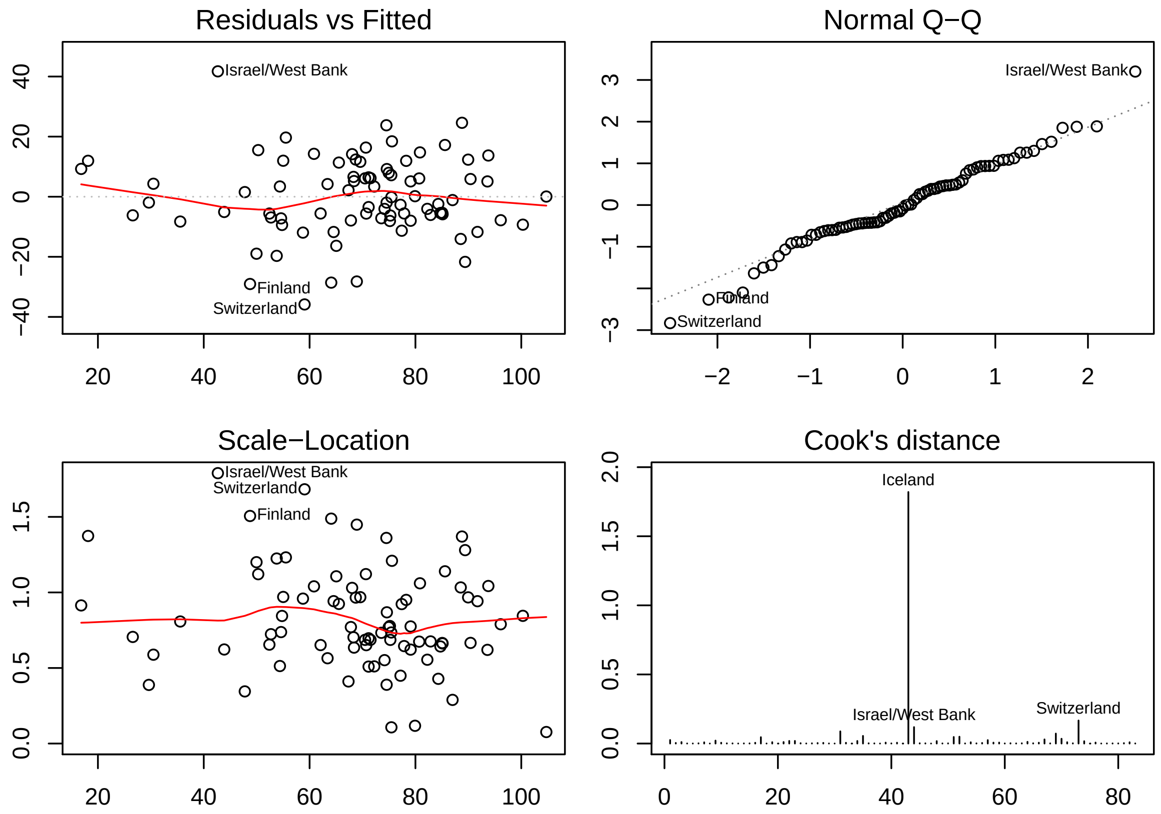


Figure 24: Diagnostic plots for model selected by Forward based on BIC criterion

The final model is not significantly worse than the full model because p value of F statistic comparing this model to the full model equals 0.38 so we should use the smaller model. Also the selected model explains data better then constant model because p value of F statistics comparing this model to the constant model equals 9.81e-16.

We see that it would be good to reomve Iceland because Cook's distance is bigger than one. After removing Icleland:

```
> printCoefmat(model.frwd.bic.sum$coefficients)

                        Estimate  Std. Error t value  Pr(>|t|)
(Intercept)          91.79738996  7.08225568 12.9616 < 2.2e-16 ***
EN.ATM.CO2E.PC       -0.64437873  0.47888144 -1.3456  0.182488
IC.EXP.DURS           0.23530733  0.10432597  2.2555  0.027018 *
IC.CRD.PRVT.ZS       -0.17591568  0.05239446 -3.3575  0.001238 **
EG.USE.ELEC.KH.PC    -0.00176106  0.00057265 -3.0753  0.002933 **
NE.TRD.GNFS.ZS       -0.07508760  0.03110525 -2.4140  0.018219 *
EG.GDP.PUSE.KO.PP.KD -0.96880297  0.45297230 -2.1388  0.035710 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

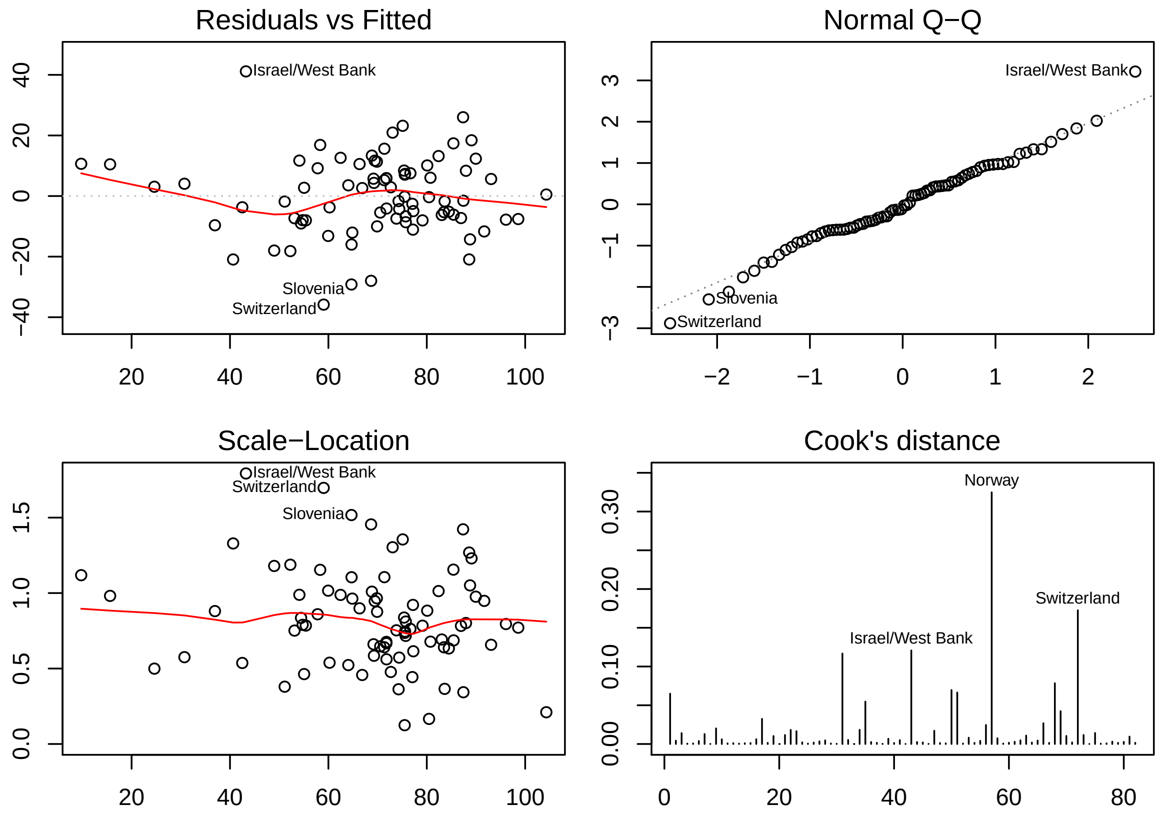


Figure 25: Diagnostic plots for model selected by Forward based on BIC criterion after removing Iceland

Adjusted R-squared: 0.63. We see also now that there is no need to remove any more data.
There is a benefit in adding quadratic term for EN.ATM.CO2E.P:

```
                       Estimate  Std. Error t value  Pr(>|t|)
(Intercept)          96.82967979  7.38886576 13.1048 < 2.2e-16 ***
EN.ATM.CO2E.PC       -2.11723759  0.87467105 -2.4206  0.017948 *
IC.EXP.DURS           0.21949283  0.10261681  2.1390  0.035739 *
IC.CRD.PRVT.ZS       -0.16982687  0.05147271 -3.2994  0.001492 **
EG.USE.ELEC.KH.PC    -0.00179809  0.00056189 -3.2001  0.002024 **
NE.TRD.GNFS.ZS       -0.08221327  0.03071251 -2.6769  0.009147 **
EG.GDP.PUSE.KO.PP.KD -1.07122234  0.44717496 -2.3955  0.019126 *
I(EN.ATM.CO2E.PC^2)   0.08028523  0.04022281  1.9960  0.049613 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted R-squared: 0.64. Comparing the model with quadratic term to model without it results in significant difference with p-value 0.05.

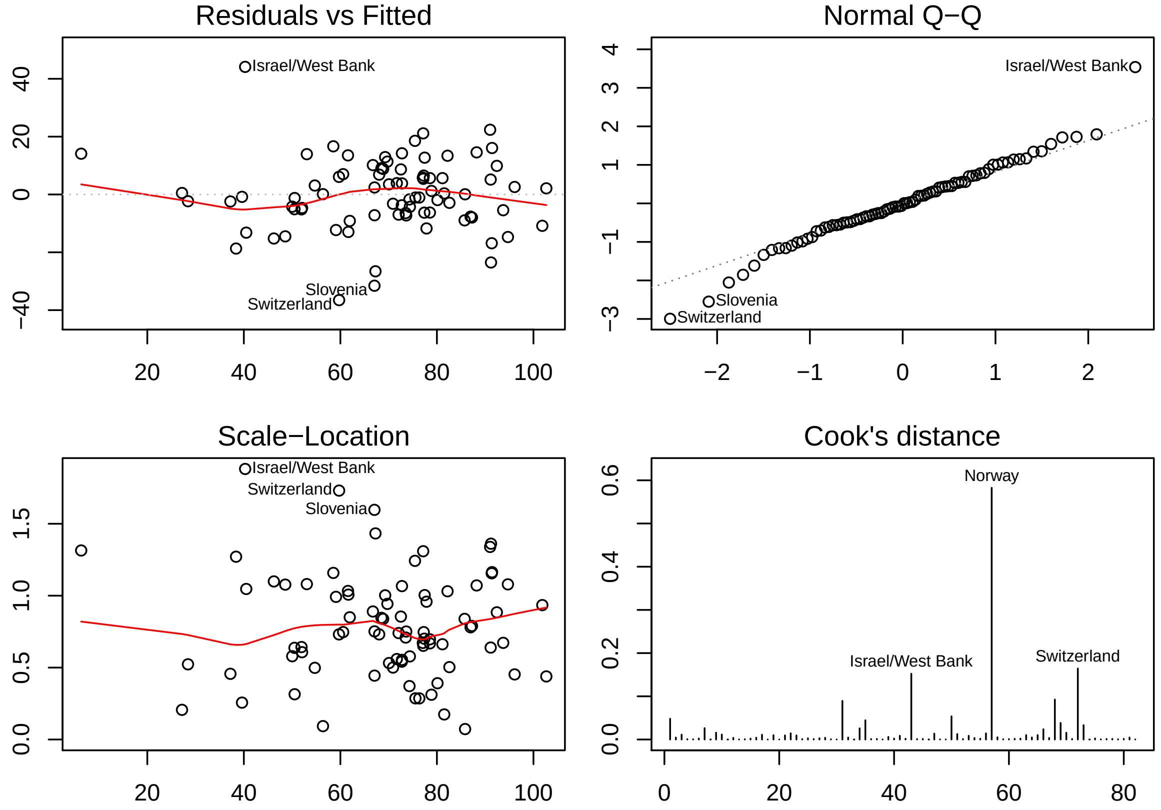


Figure 26: Diagnostic plots for model selected by Forward based on BIC criterion after removing Iceland and with quadratic term on EN.ATM.CO2E.P

## 5.7 Based on Adjusted R2 criterion

To select best model using $R^2$ we have to use adjusted $R^2$ becuase we have to take into considaration number of parameters in the model and not only how well model fits the data.

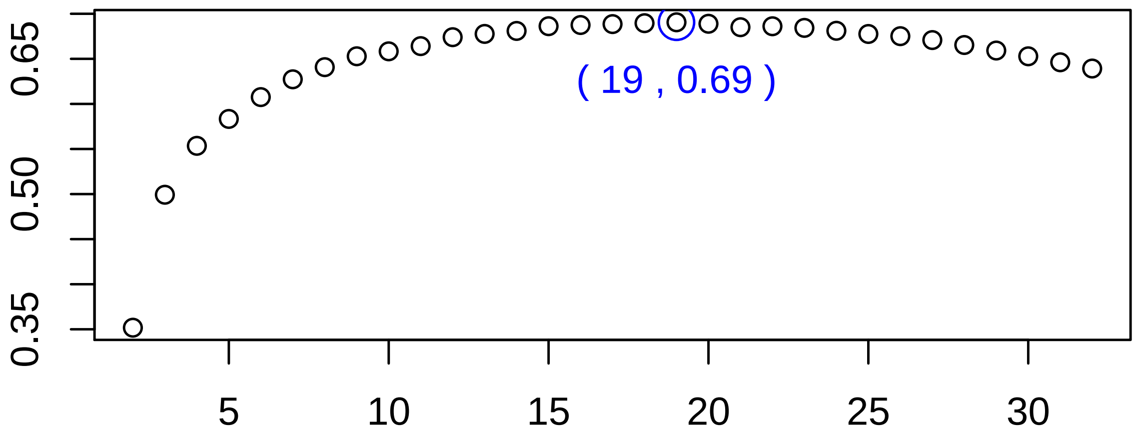


Figure 27: Adjusted $R^2$ against number of model parameters in selected best model.

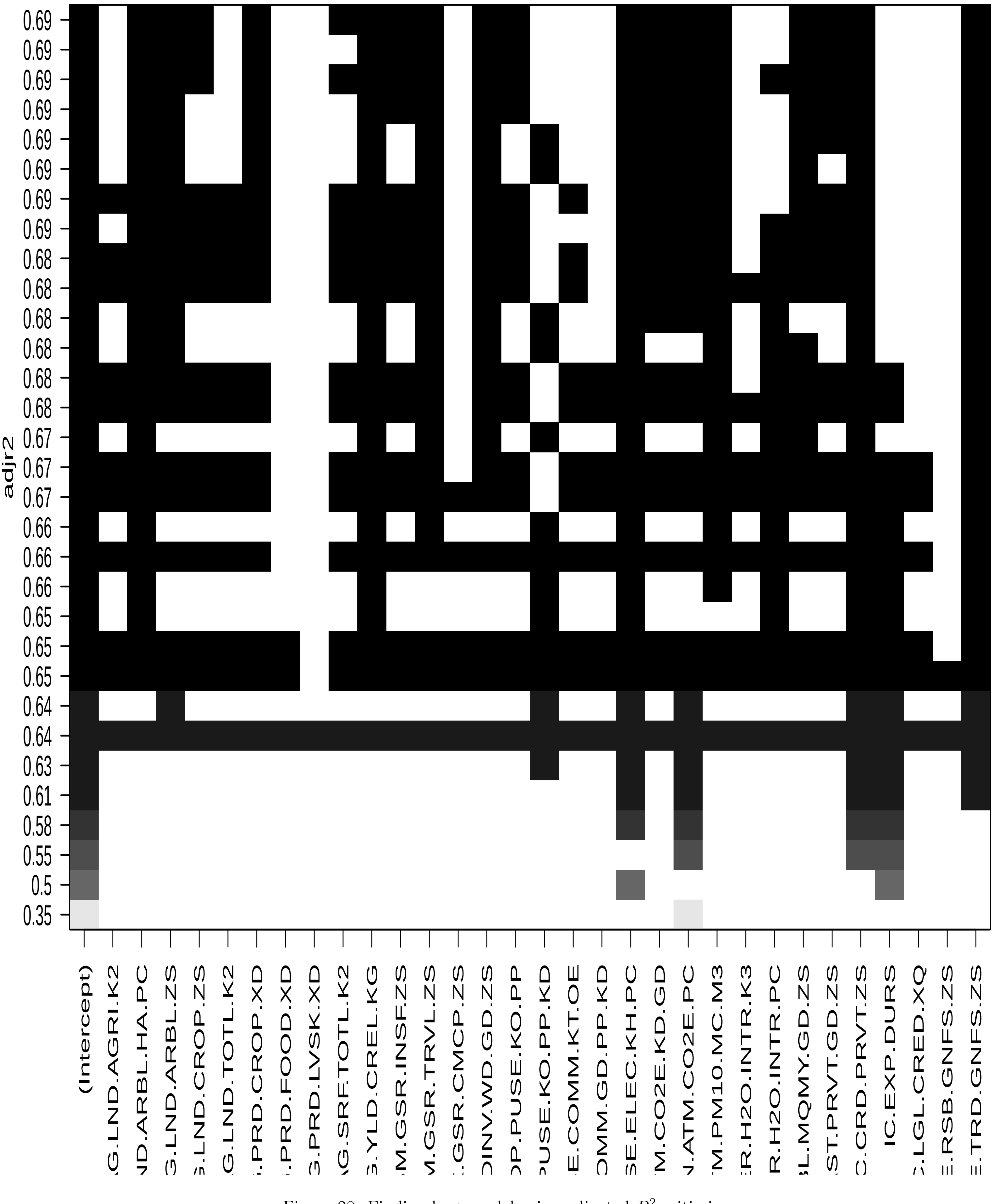


Figure 28: Finding best model using adjusted $R^2$ critierion.

Selected model with maximum adjusted $R^2$:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 1.3673e+02 | 1.4529e+01 | 9.4112 | 1.09e-13 | *** |
| AG.LND.ARBL.HA.PC | -1.4650e+01 | 5.6392e+00 | -2.5980 | 0.0116252 | * |
| AG.LND.ARBL.ZS | -2.4089e-01 | 1.1011e-01 | -2.1876 | 0.0323548 | * |
| AG.LND.CROP.ZS | -2.3289e-01 | 2.0662e-01 | -1.1272 | 0.2638836 | |
| AG.PRD.CROP.XD | -2.3006e-01 | 1.0336e-01 | -2.2257 | 0.0295625 | * |
| AG.SRF.TOTL.K2 | 7.1802e-07 | 6.5401e-07 | 1.0979 | 0.2763765 | |
| AG.YLD.CREL.KG | -2.9829e-03 | 1.0247e-03 | -2.9111 | 0.0049512 | ** |
| BM.GSR.INSF.ZS | -3.0973e-01 | 1.9282e-01 | -1.6063 | 0.1131260 | |
| BM.GSR.TRVL.ZS | 3.5855e-01 | 1.3951e-01 | 2.5701 | 0.0125057 | * |
| BX.KLT.DINV.WD.GD.ZS | 2.4313e-01 | 6.8819e-02 | 3.5329 | 0.0007687 | *** |
| EG.GDP.PUSE.KO.PP | -1.1890e+00 | 4.1210e-01 | -2.8853 | 0.0053244 | ** |
| EG.USE.ELEC.KH.PC | -1.0875e-03 | 3.7168e-04 | -2.9260 | 0.0047476 | ** |
| EN.ATM.CO2E.KD.GD | 1.7543e+00 | 9.6905e-01 | 1.8103 | 0.0749426 | . |
| EN.ATM.CO2E.PC | -1.4046e+00 | 4.3259e-01 | -3.2470 | 0.0018581 | ** |
| EN.ATM.PM10.MC.M3 | 1.0193e-01 | 4.3583e-02 | 2.3387 | 0.0224881 | * |
| FM.LBL.MQMY.GD.ZS | -1.3850e-01 | 5.9977e-02 | -2.3093 | 0.0241672 | * |
| FS.AST.PRVT.GD.ZS | 9.0793e-02 | 5.9530e-02 | 1.5252 | 0.1321457 | |
| IC.CRD.PRVT.ZS | -1.3711e-01 | 5.4743e-02 | -2.5046 | 0.0148185 | * |
| NE.TRD.GNFS.ZS | -1.1015e-01 | 3.7473e-02 | -2.9395 | 0.0045685 | ** |

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

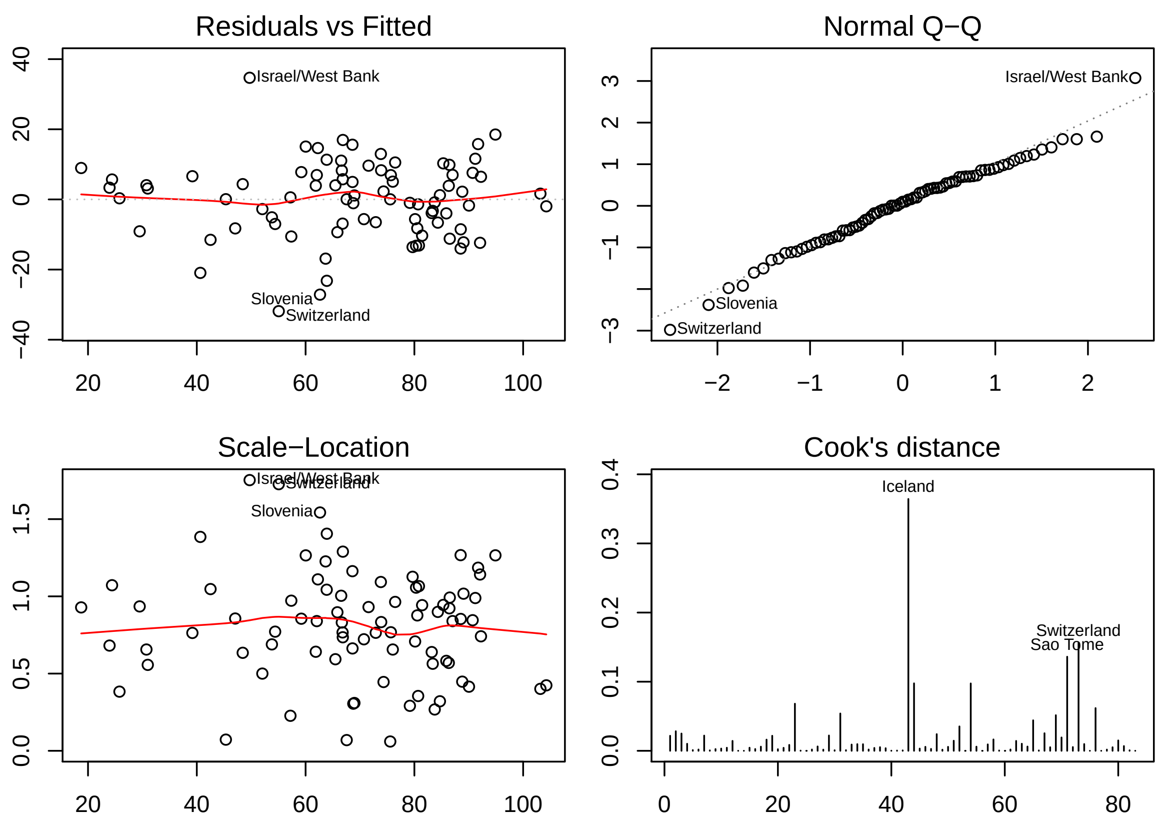


Figure 29: Diagnostic plots for model selected by adjusted $R^2$ criterion

Adjusted R-squared: 0.69.

There is a benefit in adding quadratic term for EG.GDP.PUSE.KO.PP:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 1.5496e+02 | 1.6873e+01 | 9.1836 | 3.12e-13 | *** |
| AG.LND.ARBL.HA.PC | -1.5900e+01 | 5.5469e+00 | -2.8665 | 0.0056358 | ** |
| AG.LND.ARBL.ZS | -2.7411e-01 | 1.0890e-01 | -2.5172 | 0.0143870 | * |
| AG.LND.CROP.ZS | -2.8110e-01 | 2.0338e-01 | -1.3822 | 0.1717967 | |
| AG.PRD.CROP.XD | -2.5661e-01 | 1.0189e-01 | -2.5184 | 0.0143417 | * |
| AG.SRF.TOTL.K2 | 6.6411e-07 | 6.3978e-07 | 1.0380 | 0.3032235 | |
| AG.YLD.CREL.KG | -2.8471e-03 | 1.0038e-03 | -2.8364 | 0.0061290 | ** |
| BM.GSR.INSF.ZS | -2.6678e-01 | 1.8968e-01 | -1.4065 | 0.1644875 | |

```
BM.GSR.TRVL.ZS          3.8708e-01  1.3709e-01  2.8235 0.0063518 **
BX.KLT.DINV.WD.GD.ZS    2.3983e-01  6.7282e-02  3.5645 0.0007019 ***
EG.GDP.PUSE.KO.PP      -4.9478e+00  1.9224e+00 -2.5737 0.0124252 *
EG.USE.ELEC.KH.PC      -1.2579e-03  3.7313e-04 -3.3713 0.0012824 **
EN.ATM.CO2E.KD.GD       1.1155e+00  9.9953e-01  1.1160 0.2686415
EN.ATM.CO2E.PC         -1.2659e+00  4.2846e-01 -2.9545 0.0043991 **
EN.ATM.PM10.MC.M3       8.1030e-02  4.3860e-02  1.8475 0.0693738 .
FM.LBL.MQMY.GD.ZS      -1.3396e-01  5.8664e-02 -2.2836 0.0257774 *
FS.AST.PRVT.GD.ZS       8.9576e-02  5.8186e-02  1.5395 0.1286934
IC.CRD.PRVT.ZS         -1.1485e-01  5.4650e-02 -2.1016 0.0395952 *
NE.TRD.GNFS.ZS         -1.0557e-01  3.6696e-02 -2.8770 0.0054739 **
I(EG.GDP.PUSE.KO.PP^2)  1.8998e-01  9.5006e-02  1.9996 0.0498577 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted R-squared: 0.7. Comparing the model with quadratic term to model without it results in significant difference with p-value 0.05.

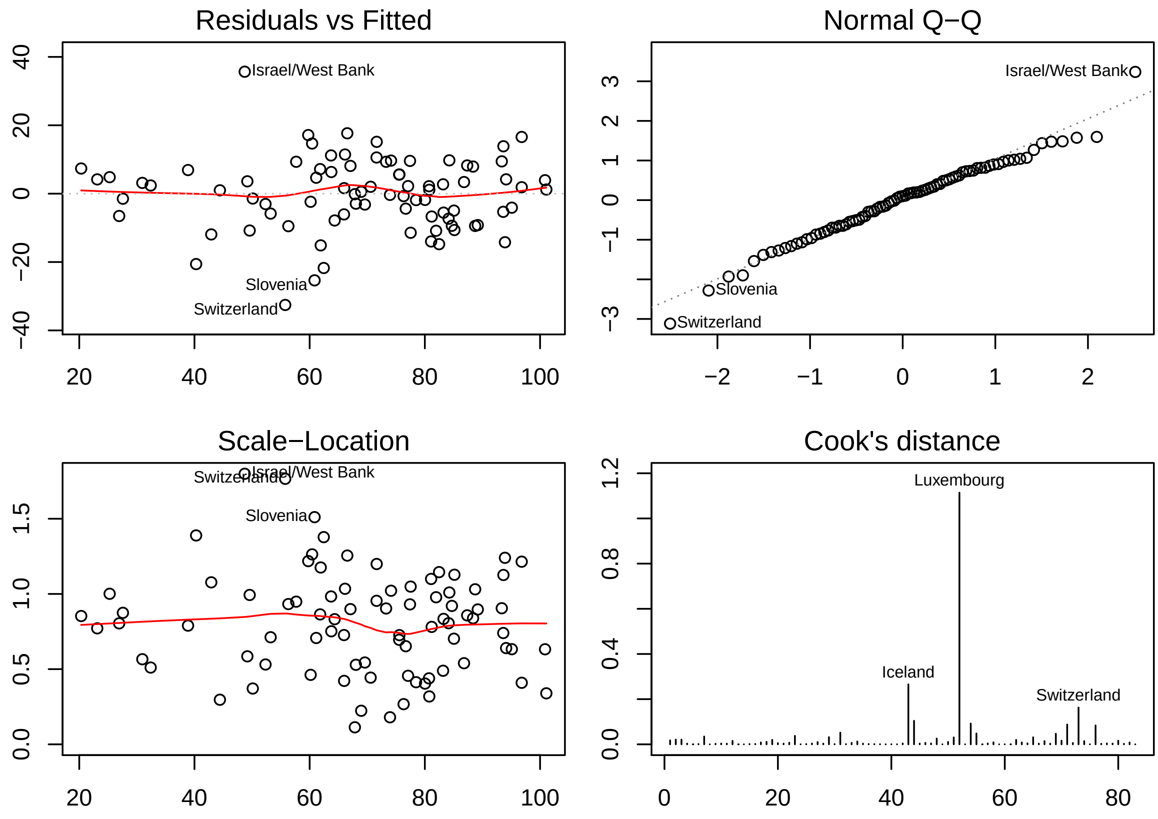


Figure 30: Diagnostic plots for model selected by adjusted $R^2$ criterion and with quadratic term on EG.GDP.PUSE.KO.PP

We see that it would be good to reomve Luxembourg because Cook's distance is bigger than one. After removing Luxembourg:

```
                         Estimate  Std. Error t value  Pr(>|t|)
(Intercept)             1.5606e+02  1.6991e+01  9.1847 3.589e-13 ***
AG.LND.ARBL.HA.PC      -1.6057e+01  5.5692e+00 -2.8831  0.005406 **
AG.LND.ARBL.ZS         -2.8621e-01  1.1041e-01 -2.5923  0.011875 *
AG.LND.CROP.ZS         -2.8151e-01  2.0406e-01 -1.3796  0.172679
AG.PRD.CROP.XD         -2.4866e-01  1.0277e-01 -2.4197  0.018480 *
AG.SRF.TOTL.K2          6.0943e-07  6.4591e-07  0.9435  0.349079
AG.YLD.CREL.KG         -2.7760e-03  1.0115e-03 -2.7446  0.007915 **
BM.GSR.INSF.ZS         -2.0882e-01  2.0494e-01 -1.0189  0.312196
BM.GSR.TRVL.ZS          4.0353e-01  1.3924e-01  2.8982  0.005182 **
BX.KLT.DINV.WD.GD.ZS    3.3210e-01  1.3861e-01  2.3959  0.019612 *
EG.GDP.PUSE.KO.PP      -5.6541e+00  2.1399e+00 -2.6422  0.010413 *
EG.USE.ELEC.KH.PC      -1.2730e-03  3.7490e-04 -3.3956  0.001200 **
EN.ATM.CO2E.KD.GD       1.0075e+00  1.0128e+00  0.9947  0.323753
EN.ATM.CO2E.PC         -1.2048e+00  4.3729e-01 -2.7553  0.007689 **
EN.ATM.PM10.MC.M3       7.3750e-02  4.5031e-02  1.6378  0.106535
FM.LBL.MQMY.GD.ZS      -1.1505e-01  6.3879e-02 -1.8010  0.076561 .
```

```
FS.AST.PRVT.GD.ZS        7.1091e-02  6.3217e-02  1.1246  0.265112
IC.CRD.PRVT.ZS          -1.0506e-01  5.6317e-02 -1.8655  0.066848 .
NE.TRD.GNFS.ZS          -1.1873e-01  4.0661e-02 -2.9199  0.004876 **
I(EG.GDP.PUSE.KO.PP^2)   2.2378e-01  1.0514e-01  2.1285  0.037278 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

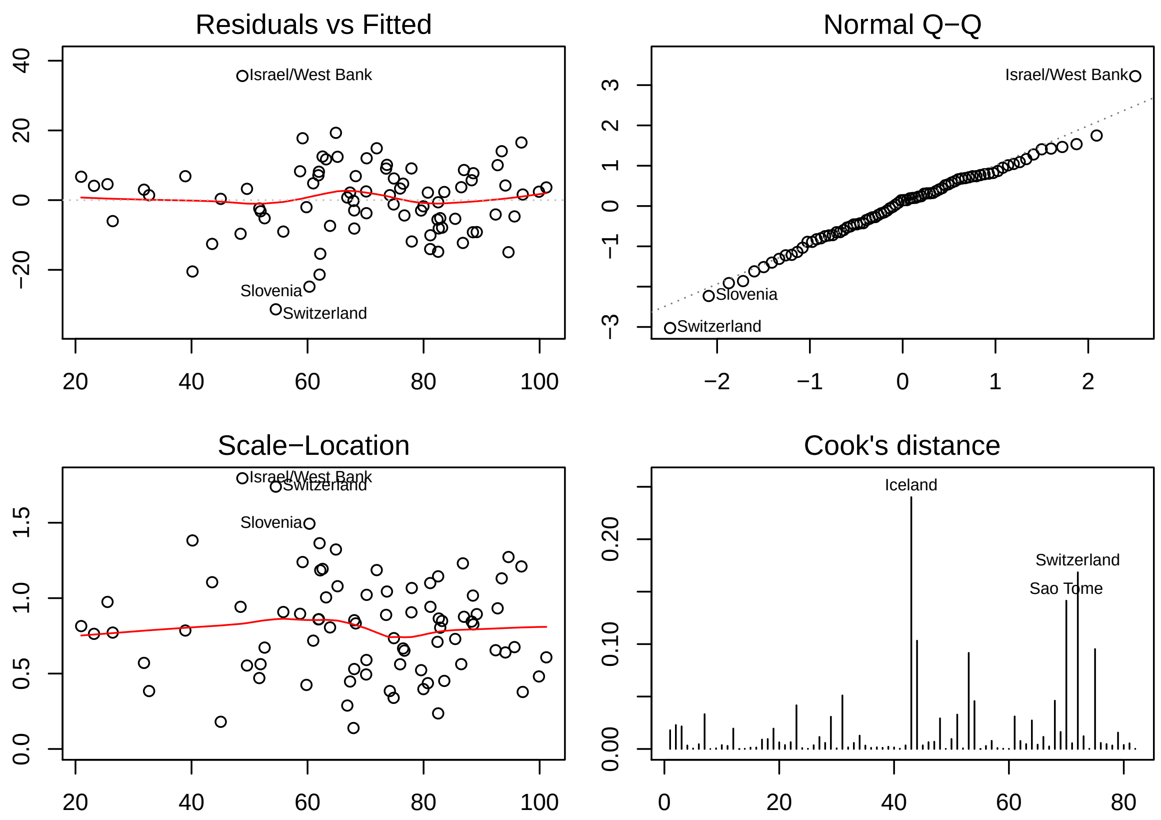


Figure 31: Diagnostic plots for model selected by adjusted $R^2$ criterion and with quadratic term on EG.GDP.PUSE.KO.PP and removing Luxembourg

Adjusted R-squared: 0.69. We see also now that there is no need to remove any more data.

## 5.8 Based on $C_p$ criterion

Mallow's $C_p$ criterion selects model that should predict well i.e. it tries to minimize average mean sqare error.

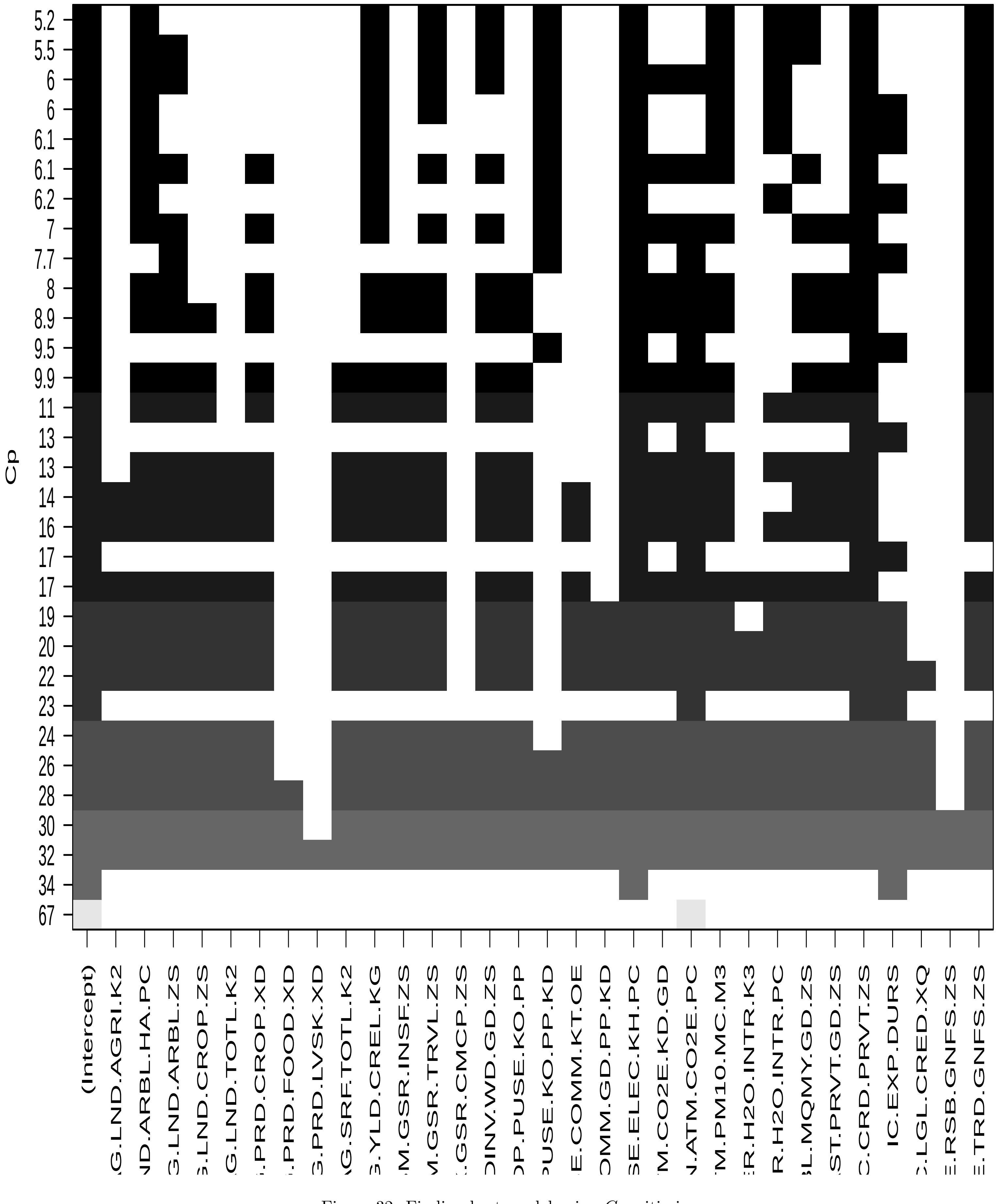


Figure 32: Finding best model using $C_p$ critierion.

Checking $C_p$ with respect to number of parameters (A model with good fit should have $C_p$ around or less then p):

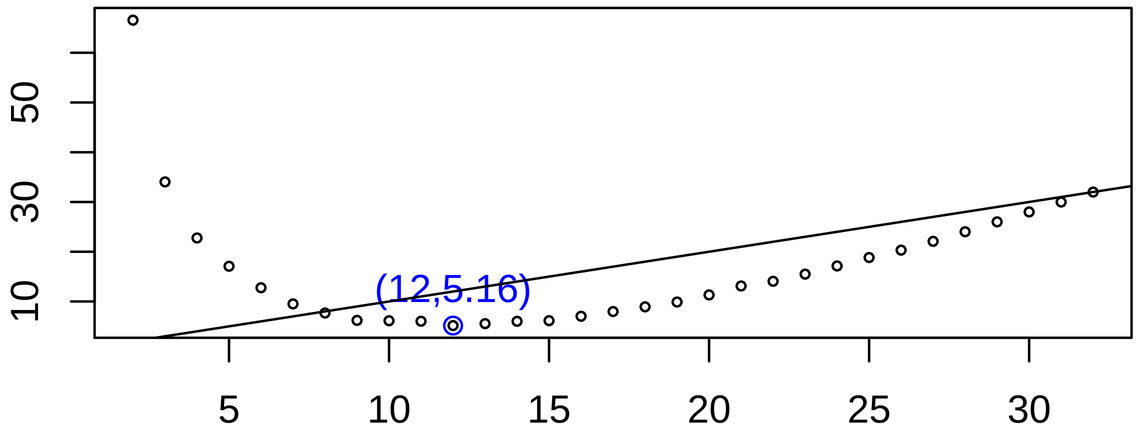


Figure 33: $C_p$ against number of model parameters.

We see that the model with 11 predictors (choosen by $C_p$ criterion) meets goodnes of fit requiremnt.
Selected model with minimum $C_p$:

```
                        Estimate  Std. Error t value  Pr(>|t|)
(Intercept)            1.0604e+02  7.0884e+00 14.9590 < 2.2e-16 ***
AG.LND.ARBL.HA.PC     -1.5073e+01  4.7113e+00 -3.1992  0.002061 **
AG.YLD.CREL.KG        -3.0971e-03  9.7807e-04 -3.1665  0.002274 **
BM.GSR.TRVL.ZS         2.2644e-01  1.2011e-01  1.8852  0.063493 .
BX.KLT.DINV.WD.GD.ZS  1.1227e-01  4.5017e-02  2.4939  0.014967 *
EG.GDP.PUSE.KO.PP.KD -1.4667e+00  4.1939e-01 -3.4973  0.000815 ***
EG.USE.ELEC.KH.PC     -1.5972e-03  3.3762e-04 -4.7308 1.105e-05 ***
EN.ATM.PM10.MC.M3      8.5535e-02  3.9993e-02  2.1387  0.035898 *
ER.H2O.INTR.PC         7.1739e-05  2.4657e-05  2.9095  0.004831 **
FM.LBL.MQMY.GD.ZS     -6.9397e-02  3.3049e-02 -2.0998  0.039299 *
IC.CRD.PRVT.ZS        -1.3999e-01  5.2572e-02 -2.6628  0.009581 **
NE.TRD.GNFS.ZS        -1.1717e-01  3.4629e-02 -3.3836  0.001168 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

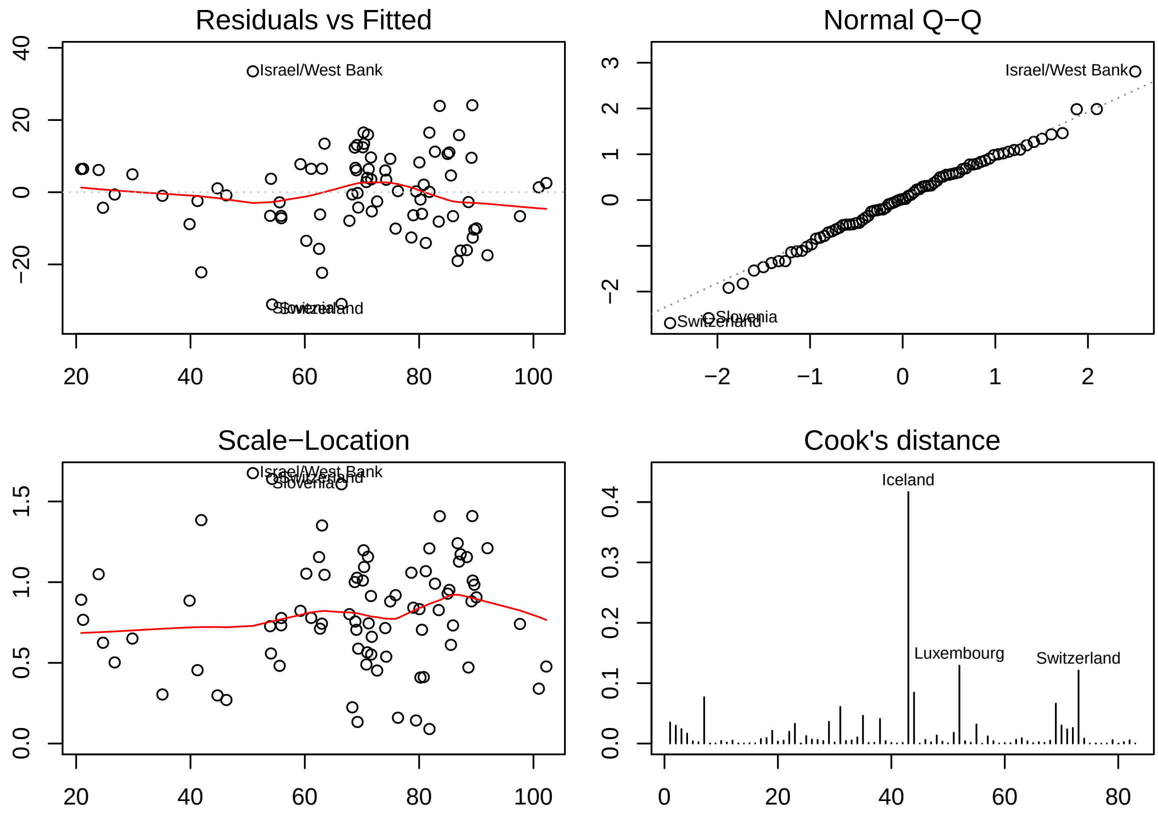


Figure 34: Diagnostic plots for model selected by $C_p$ criterion

Adjusted R-squared: 0.67.
There is a benefit in adding quadratic term for EG.GDP.PUSE.KO.PP.KD:

```
                          Estimate  Std. Error t value  Pr(>|t|)
(Intercept)             1.1831e+02  8.4215e+00 14.0488 < 2.2e-16 ***
AG.LND.ARBL.HA.PC      -1.6609e+01  4.5878e+00 -3.6202 0.0005526 ***
AG.YLD.CREL.KG         -3.0733e-03  9.4388e-04 -3.2561 0.0017427 **
BM.GSR.TRVL.ZS          2.4041e-01  1.1604e-01  2.0717 0.0419815 *
BX.KLT.DINV.WD.GD.ZS    9.9778e-02  4.3727e-02  2.2818 0.0255425 *
EG.GDP.PUSE.KO.PP.KD   -5.5264e+00  1.6742e+00 -3.3010 0.0015182 **
EG.USE.ELEC.KH.PC      -1.6939e-03  3.2809e-04 -5.1631 2.179e-06 ***
EN.ATM.PM10.MC.M3       8.0531e-02  3.8645e-02  2.0839 0.0408234 *
ER.H2O.INTR.PC          6.7411e-05  2.3857e-05  2.8257 0.0061436 **
FM.LBL.MQMY.GD.ZS      -4.8775e-02  3.2942e-02 -1.4806 0.1431917
IC.CRD.PRVT.ZS         -9.8790e-02  5.3343e-02 -1.8520 0.0682442 .
NE.TRD.GNFS.ZS         -1.1040e-01  3.3526e-02 -3.2930 0.0015563 **
I(EG.GDP.PUSE.KO.PP.KD^2)  2.2271e-01  8.9117e-02  2.4990 0.0148018 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted R-squared: 0.7. Comparing the model with quadratic term to model without it results in significant difference with p-value 0.015.

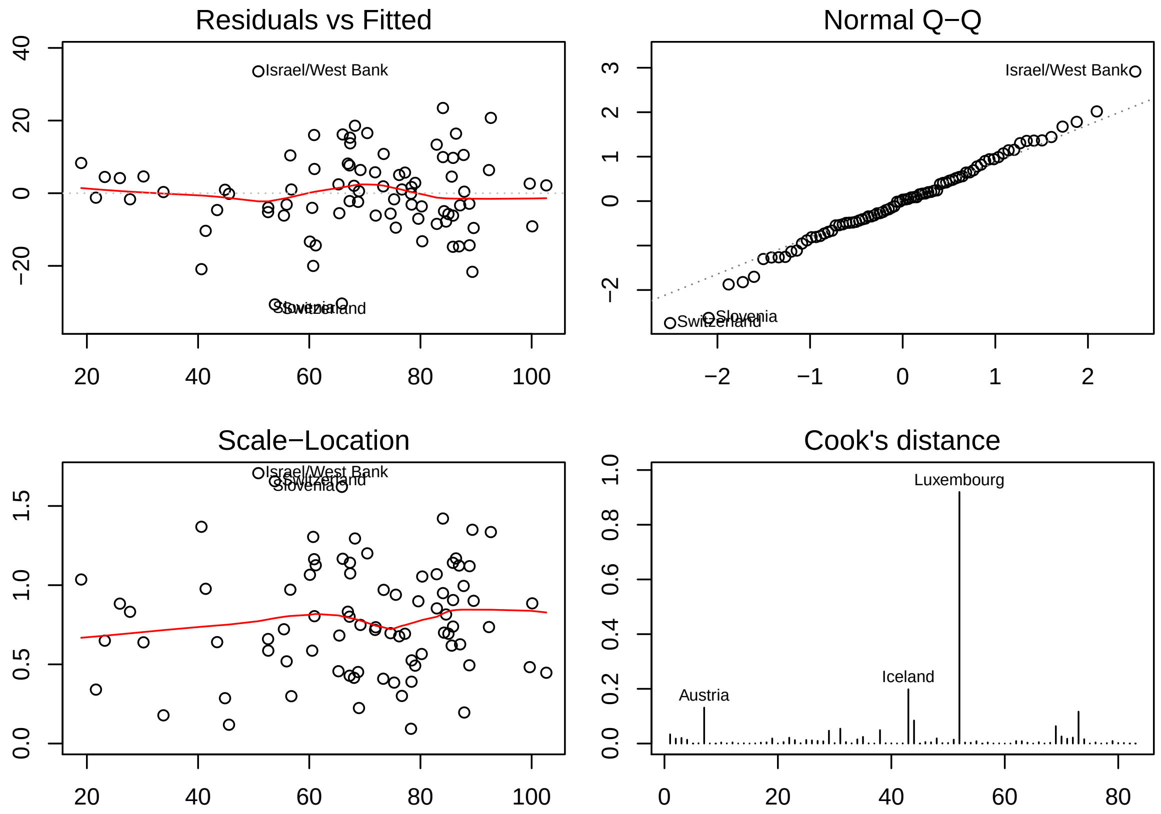


Figure 35: Diagnostic plots for model selected by $C_p$ criterion and with quadratic term on EG.GDP.PUSE.KO.PP.KD

We see that it would be good to reomve Luxembourg because Cook's distance is bigger than one. After removing Luxembourg:

```
                        Estimate  Std. Error t value  Pr(>|t|)
(Intercept)            1.5606e+02  1.6991e+01   9.1847 3.589e-13 ***
AG.LND.ARBL.HA.PC     -1.6057e+01  5.5692e+00 -2.8831  0.005406 **
AG.LND.ARBL.ZS        -2.8621e-01  1.1041e-01 -2.5923  0.011875 *
AG.LND.CROP.ZS        -2.8151e-01  2.0406e-01 -1.3796  0.172679
AG.PRD.CROP.XD        -2.4866e-01  1.0277e-01 -2.4197  0.018480 *
AG.SRF.TOTL.K2         6.0943e-07  6.4591e-07   0.9435  0.349079
AG.YLD.CREL.KG        -2.7760e-03  1.0115e-03 -2.7446  0.007915 **
BM.GSR.INSF.ZS        -2.0882e-01  2.0494e-01 -1.0189  0.312196
BM.GSR.TRVL.ZS         4.0353e-01  1.3924e-01   2.8982  0.005182 **
BX.KLT.DINV.WD.GD.ZS   3.3210e-01  1.3861e-01   2.3959  0.019612 *
EG.GDP.PUSE.KO.PP     -5.6541e+00  2.1399e+00 -2.6422  0.010413 *
EG.USE.ELEC.KH.PC     -1.2730e-03  3.7490e-04 -3.3956  0.001200 **
EN.ATM.CO2E.KD.GD      1.0075e+00  1.0128e+00   0.9947  0.323753
EN.ATM.CO2E.PC        -1.2048e+00  4.3729e-01 -2.7553  0.007689 **
EN.ATM.PM10.MC.M3      7.3750e-02  4.5031e-02   1.6378  0.106535
FM.LBL.MQMY.GD.ZS     -1.1505e-01  6.3879e-02 -1.8010  0.076561 .
FS.AST.PRVT.GD.ZS      7.1091e-02  6.3217e-02   1.1246  0.265112
IC.CRD.PRVT.ZS        -1.0506e-01  5.6317e-02 -1.8655  0.066848 .
NE.TRD.GNFS.ZS        -1.1873e-01  4.0661e-02 -2.9199  0.004876 **
I(EG.GDP.PUSE.KO.PP^2)  2.2378e-01  1.0514e-01   2.1285  0.037278 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

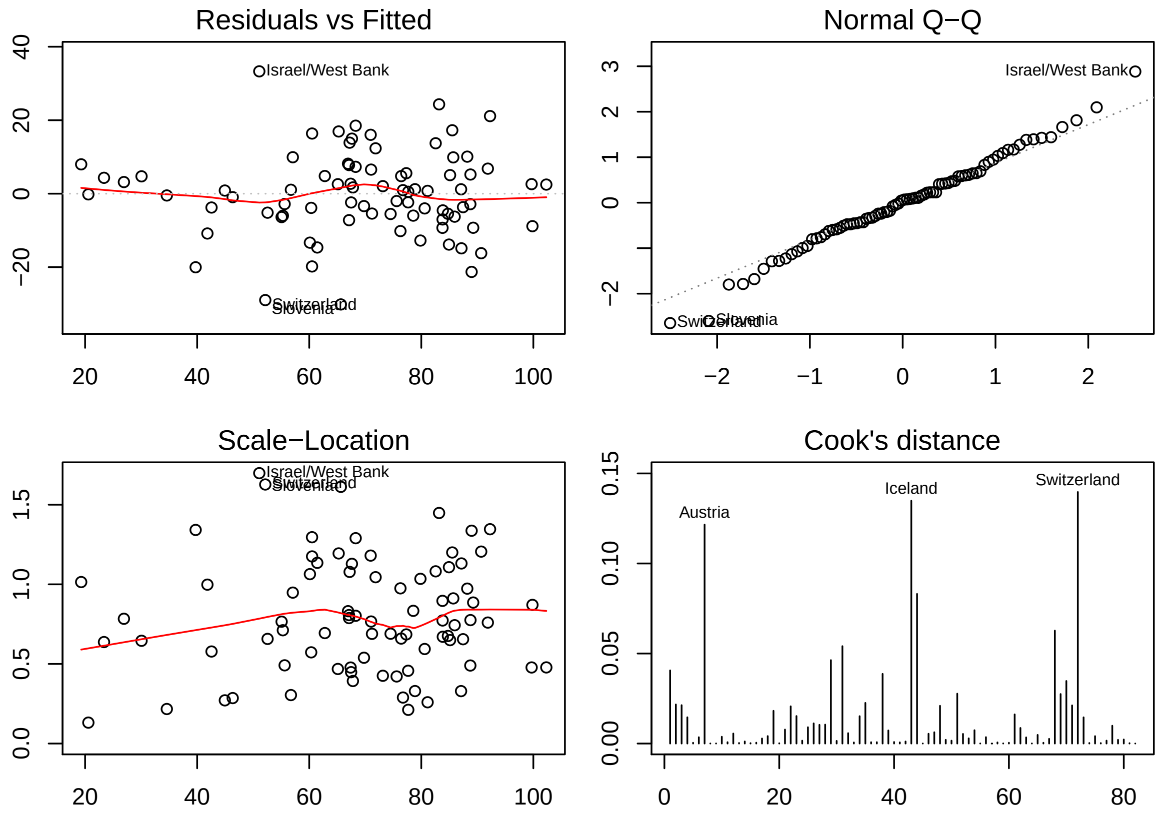


Figure 36: Diagnostic plots for model selected by $C_p$ criterion and with quadratic term on EG.GDP.PUSE.KO.PP.KD and removed Luxembourg

Adjusted R-squared: 0.69. We see also now that there is no need to remove any more data.

# 6 Shrinkage methods

Methods in this section can be used to automatically select the predictors.

## 6.1 Ridge regression

To choose biased estimate of the $\beta$ we can apply ridge regression which deals with collnerity of the predictors.

Choose the best value of penalty parameter using genralized crossvalidation:

```
modified HKB estimator is 0.1465822
modified L-W estimator is 13.65141
smallest value of GCV  at 39.5
```

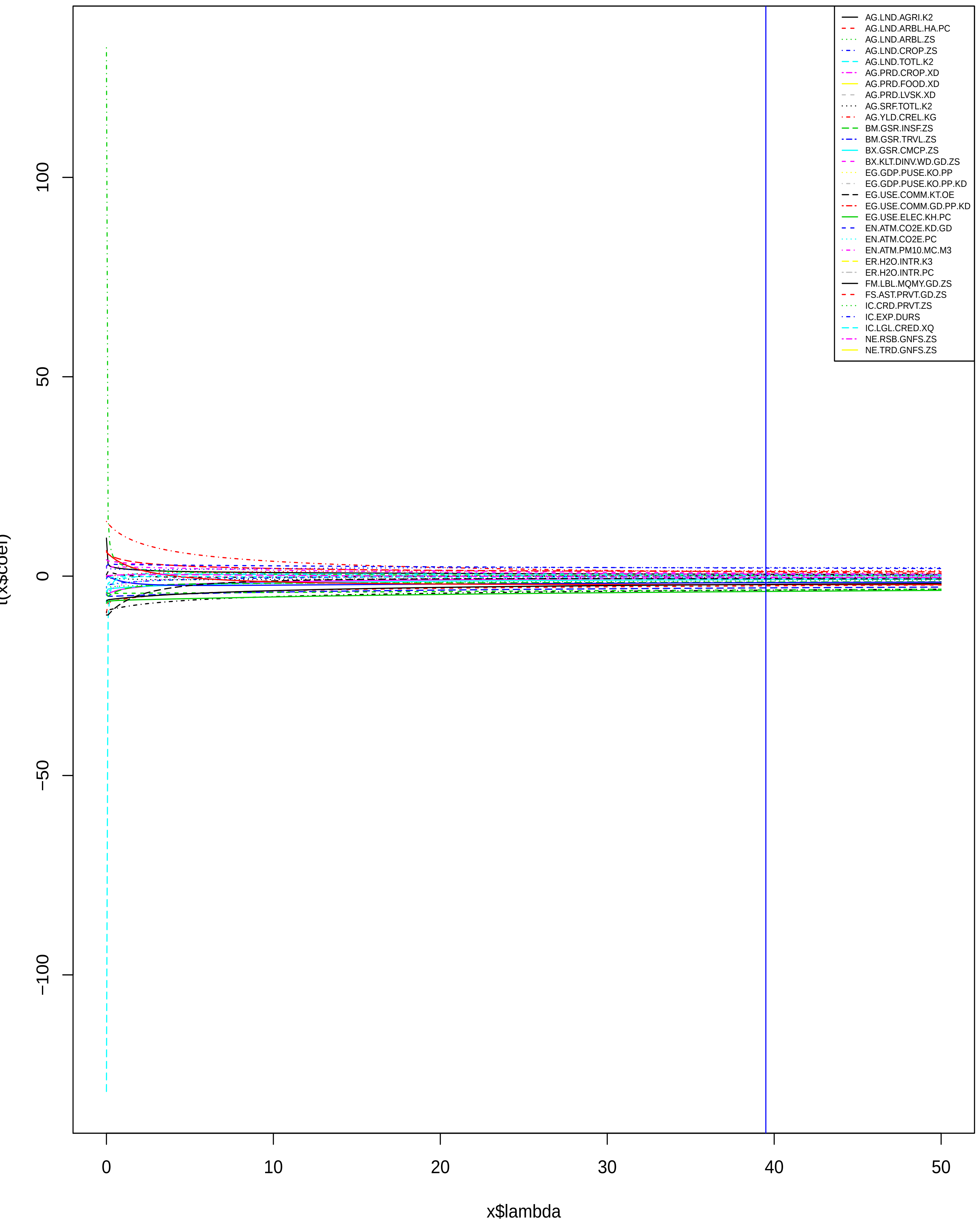


Figure 37: fitted values of coefficients (bi) as a function of parameter $\lambda$ with marked optimal lambda

The impact of lambda on GCV can be visualized:

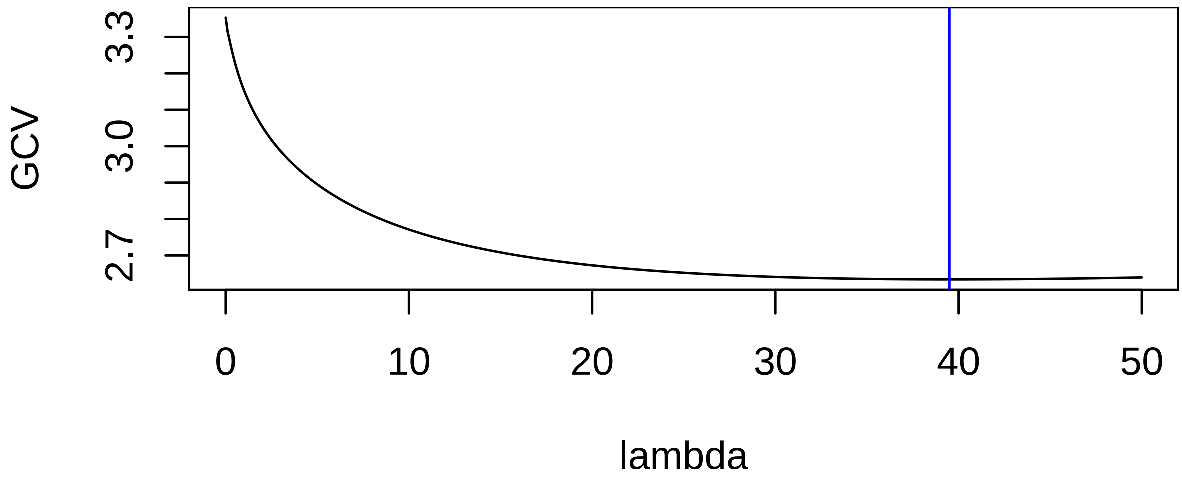


Figure 38: GCV with respect to lmbda

Coefficients of selected model:

```
                             AG.LND.AGRI.K2    AG.LND.ARBL.HA.PC        AG.LND.ARBL.ZS
        2.820289e-15           4.891585e-07        -7.116879e+00         -8.679093e-02
      AG.LND.CROP.ZS         AG.LND.TOTL.K2       AG.PRD.CROP.XD        AG.PRD.FOOD.XD
        8.121092e-03           4.340013e-08        -2.326247e-02          9.425381e-03
      AG.PRD.LVSK.XD         AG.SRF.TOTL.K2       AG.YLD.CREL.KG        BM.GSR.INSF.ZS
        6.884878e-02           4.132774e-08        -1.663944e-03         -9.167729e-02
      BM.GSR.TRVL.ZS         BX.GSR.CMCP.ZS BX.KLT.DINV.WD.GD.ZS     EG.GDP.PUSE.KO.PP
        5.776156e-02          -1.603808e-02         1.765687e-02         -3.708520e-01
EG.GDP.PUSE.KO.PP.KD      EG.USE.COMM.KT.OE EG.USE.COMM.GD.PP.KD     EG.USE.ELEC.KH.PC
       -4.498845e-01           1.204373e-06         1.260502e-03         -5.198139e-04
   EN.ATM.CO2E.KD.GD         EN.ATM.CO2E.PC    EN.ATM.PM10.MC.M3        ER.H2O.INTR.K3
        5.849059e-01          -7.249030e-01         5.265726e-02          6.155290e-05
      ER.H2O.INTR.PC      FM.LBL.MQMY.GD.ZS    FS.AST.PRVT.GD.ZS        IC.CRD.PRVT.ZS
        5.709176e-06          -8.496698e-03        -3.662162e-02         -9.630877e-02
         IC.EXP.DURS         IC.LGL.CRED.XQ       NE.RSB.GNFS.ZS        NE.TRD.GNFS.ZS
        1.238967e-01          -1.934809e-01        -1.876214e-02         -3.826836e-02
```

RMSE computed using leave one out is 22.61. We see that biased estimator which decreases the variance of estimator increasing the bias can give worse results in terms of prediction compared to other methods.

### 6.2 LASSO regression

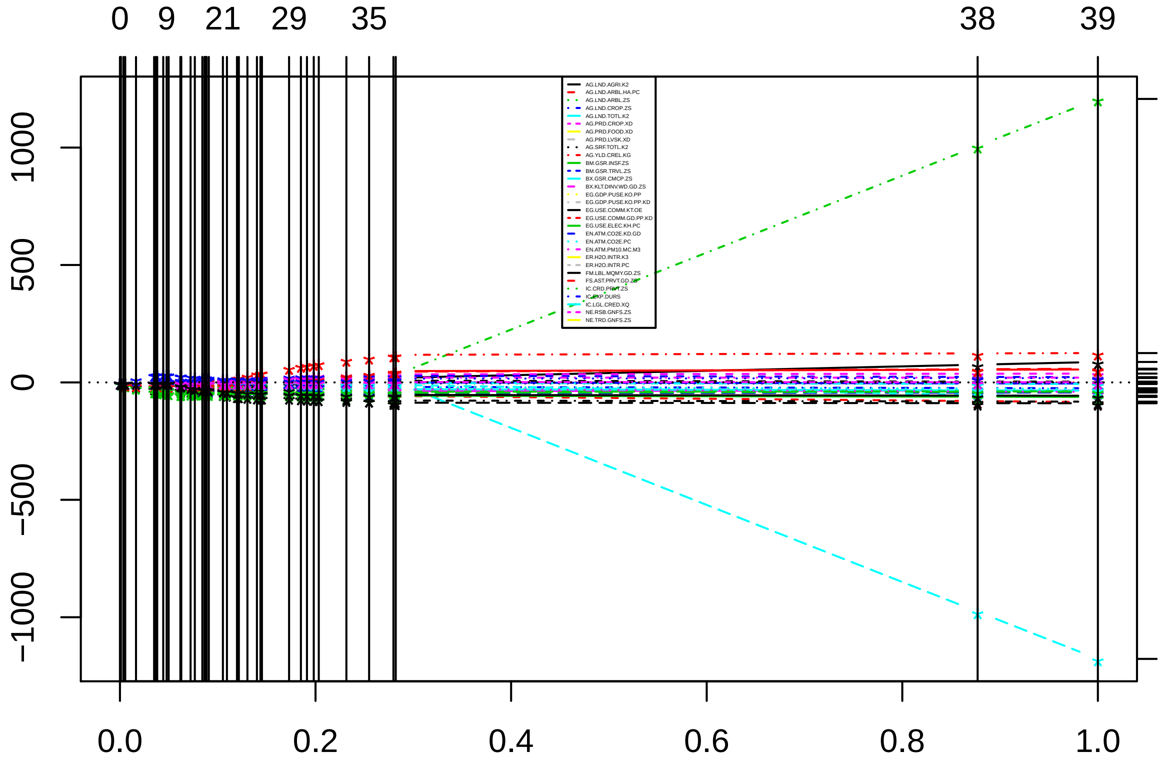


Figure 39: Visualization of the coefficients paths for LASSO

Choosing the best subset of predictors in LASSO regresssion on the basis of Mallows Cp criterion

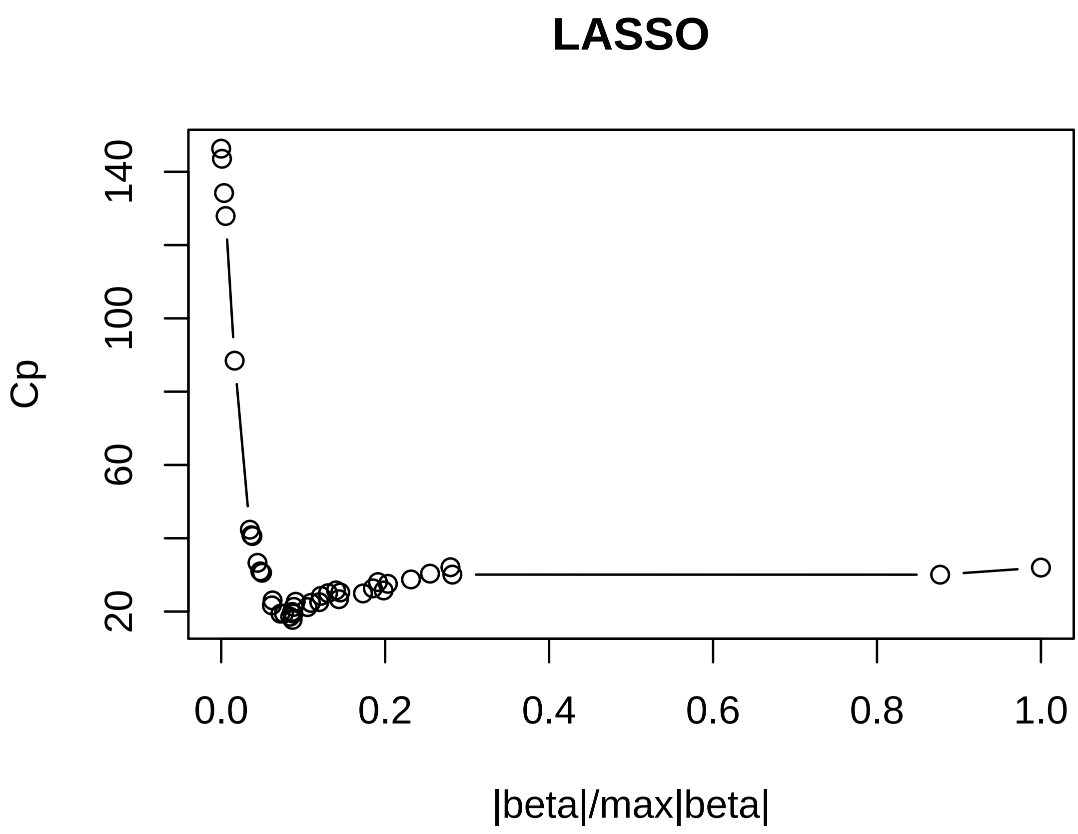


Figure 40: Visualization of Cp for LASSO

So to select the best coefficient based on $C_p$ criterium we select the set with the smallest $C_p$:

```
> (lasso.coef.best.idx<-as.numeric(which.min(model.lasso$Cp)))
```

```
[1] 18
```

values of fitted coefficients (bi) in LASSO regression for the chosen model

```
> model.lasso$beta[lasso.coef.best.idx,]
```

```
     AG.LND.AGRI.K2    AG.LND.ARBL.HA.PC       AG.LND.ARBL.ZS       AG.LND.CROP.ZS
       1.632643e-07        -5.946752e+00        -1.059844e-01         0.000000e+00
     AG.LND.TOTL.K2       AG.PRD.CROP.XD       AG.PRD.FOOD.XD       AG.PRD.LVSK.XD
       0.000000e+00         0.000000e+00         0.000000e+00         3.405054e-02
     AG.SRF.TOTL.K2       AG.YLD.CREL.KG       BM.GSR.INSF.ZS       BM.GSR.TRVL.ZS
       0.000000e+00        -1.630761e-03         0.000000e+00         5.545624e-02
     BX.GSR.CMCP.ZS BX.KLT.DINV.WD.GD.ZS    EG.GDP.PUSE.KO.PP EG.GDP.PUSE.KO.PP.KD
       0.000000e+00         0.000000e+00         0.000000e+00        -9.882746e-01
  EG.USE.COMM.KT.OE EG.USE.COMM.GD.PP.KD    EG.USE.ELEC.KH.PC    EN.ATM.CO2E.KD.GD
       0.000000e+00         0.000000e+00        -6.132116e-04         3.826532e-01
     EN.ATM.CO2E.PC    EN.ATM.PM10.MC.M3       ER.H2O.INTR.K3       ER.H2O.INTR.PC
      -9.547260e-01         4.250566e-02         0.000000e+00         0.000000e+00
  FM.LBL.MQMY.GD.ZS    FS.AST.PRVT.GD.ZS       IC.CRD.PRVT.ZS          IC.EXP.DURS
       0.000000e+00        -2.093679e-02        -1.241605e-01         1.477372e-01
     IC.LGL.CRED.XQ       NE.RSB.GNFS.ZS       NE.TRD.GNFS.ZS
       0.000000e+00         0.000000e+00        -5.170877e-02
```

Choosing best predictors in LASSO regresssion on the basis of crossvalidation:

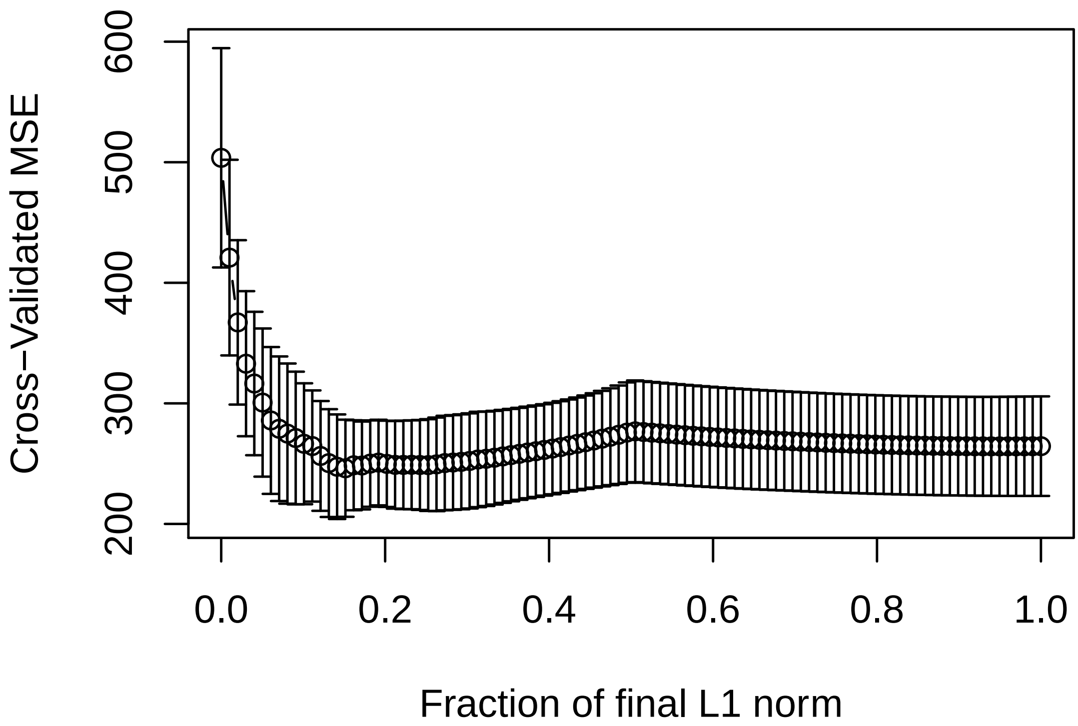


Figure 41: Visualisation of CV MSE

Getting coefficient for the minimum MSE:

```
> frac<-kaggle.cv.lars$index[kaggle.cv.lars$cv==min(kaggle.cv.lars$cv)]
> predict.lars(model.lasso, type="coefficients", mode="fraction", s=frac)$coef
```

```
     AG.LND.AGRI.K2    AG.LND.ARBL.HA.PC       AG.LND.ARBL.ZS       AG.LND.CROP.ZS
       1.075583e-06        -1.172119e+01        -1.598687e-01        -2.218017e-02
     AG.LND.TOTL.K2       AG.PRD.CROP.XD       AG.PRD.FOOD.XD       AG.PRD.LVSK.XD
       0.000000e+00        -6.754011e-02         0.000000e+00         1.710303e-02
     AG.SRF.TOTL.K2       AG.YLD.CREL.KG       BM.GSR.INSF.ZS       BM.GSR.TRVL.ZS
       1.200951e-07        -2.499797e-03        -1.686797e-01         1.981832e-01
     BX.GSR.CMCP.ZS BX.KLT.DINV.WD.GD.ZS    EG.GDP.PUSE.KO.PP EG.GDP.PUSE.KO.PP.KD
      -5.779094e-03         9.272045e-02        -3.228319e-01        -7.886943e-01
  EG.USE.COMM.KT.OE EG.USE.COMM.GD.PP.KD    EG.USE.ELEC.KH.PC    EN.ATM.CO2E.KD.GD
       3.794306e-07         0.000000e+00        -1.021712e-03         1.092142e+00
     EN.ATM.CO2E.PC    EN.ATM.PM10.MC.M3       ER.H2O.INTR.K3       ER.H2O.INTR.PC
      -9.807543e-01         6.430192e-02         0.000000e+00         2.456967e-05
  FM.LBL.MQMY.GD.ZS    FS.AST.PRVT.GD.ZS       IC.CRD.PRVT.ZS          IC.EXP.DURS
      -2.968208e-02         0.000000e+00        -1.165076e-01         8.675806e-02
     IC.LGL.CRED.XQ       NE.RSB.GNFS.ZS       NE.TRD.GNFS.ZS
       0.000000e+00         1.724057e-02        -8.841993e-02
```

## 6.3 Robust regression: M-estimators and Least Trimmed Squares

Methods that deals with vialoted assumptions of the linear regression like outliers, heterscedacticity of variance, fatter tails of errors. These methods are usefull when automatic and quick model fitting is required or to compare to LS model for validation (If they differ the source of dissimilarity should be investigated). Here M-estimators apply the method with the Huber function. The M-estimators uses special function on residuals when minimizing sum of those (possibly different than quadratic like in case of the LS). The Least Trimmed Squares ignores the biggest residuals in the optimization process. Both models are fitted using predictors selected by VIF method so we are not impacted by colinearity. Comparing those coeffcients to the LS model allows us to check what is the influence of the remaining outliers:

| | VIF | M-est | LTS |
|---|---|---|---|
| (Intercept) | 1.183955e+02 | 1.114423e+02 | 1.120662e+02 |
| AG.LND.AGRI.K2 | 1.598293e-06 | 7.333228e-07 | 3.650273e-05 |
| AG.LND.ARBL.HA.PC | -1.253239e+01 | -8.574811e+00 | 8.288523e+00 |
| AG.LND.ARBL.ZS | -1.745265e-01 | -1.611377e-01 | -1.186622e-01 |
| AG.LND.CROP.ZS | -1.570192e-01 | -1.044620e-01 | -2.215589e-01 |
| AG.PRD.CROP.XD | -1.324605e-01 | -1.448184e-01 | -3.320323e-02 |
| AG.PRD.LVSK.XD | 3.700824e-02 | 7.421651e-02 | 3.672683e-02 |
| AG.YLD.CREL.KG | -2.988844e-03 | -2.492744e-03 | 1.294025e-03 |
| BM.GSR.INSF.ZS | -3.661819e-01 | -2.222917e-01 | -2.817695e-01 |
| BM.GSR.TRVL.ZS | 2.215913e-01 | 2.072900e-01 | -2.731435e-01 |
| BX.GSR.CMCP.ZS | -4.889663e-02 | -3.623866e-02 | 3.490963e-02 |
| BX.KLT.DINV.WD.GD.ZS | 1.123641e-01 | 8.797489e-02 | -4.467767e-01 |
| EG.GDP.PUSE.KO.PP | -1.061146e+00 | -1.015587e+00 | -7.831947e-01 |
| EG.USE.COMM.KT.OE | 2.605741e-06 | 4.210267e-06 | -6.894729e-05 |
| EG.USE.ELEC.KH.PC | -1.009742e-03 | -7.513840e-04 | 2.430462e-04 |
| EN.ATM.CO2E.KD.GD | 1.704628e+00 | 1.474666e+00 | 4.844412e+00 |
| EN.ATM.CO2E.PC | -1.267936e+00 | -1.507218e+00 | -4.821782e+00 |
| EN.ATM.PM10.MC.M3 | 7.966807e-02 | 6.889363e-02 | -9.204514e-02 |
| ER.H2O.INTR.K3 | -1.611311e-04 | 1.176023e-04 | -5.023729e-05 |
| ER.H2O.INTR.PC | 3.734656e-05 | 2.162018e-05 | -1.248380e-04 |
| FS.AST.PRVT.GD.ZS | -1.720066e-02 | -2.576060e-02 | 1.282050e-01 |
| IC.CRD.PRVT.ZS | -1.073238e-01 | -1.140820e-01 | -1.521718e-01 |
| IC.EXP.DURS | 4.547959e-02 | 5.496453e-02 | 4.796941e-02 |
| IC.LGL.CRED.XQ | -1.526177e-02 | -1.773783e-01 | -4.785446e-01 |
| NE.RSB.GNFS.ZS | 3.659690e-02 | 5.853014e-02 | 9.089325e-02 |
| NE.TRD.GNFS.ZS | -1.012795e-01 | -7.978937e-02 | -6.599761e-02 |

As we can see in most of the cases the robust regression methods does not change coefficients considerably so we can conclude that outliers does not change model too much.

## 6.4 PCA

PCA tries to rotate model matrix into ortogonality hence simplifying testing and interpretation. I use prcomp function which should be more acurate than princomp (uses SVD). After performing PCA on kaggle data we obtain standard deviations for principal components:

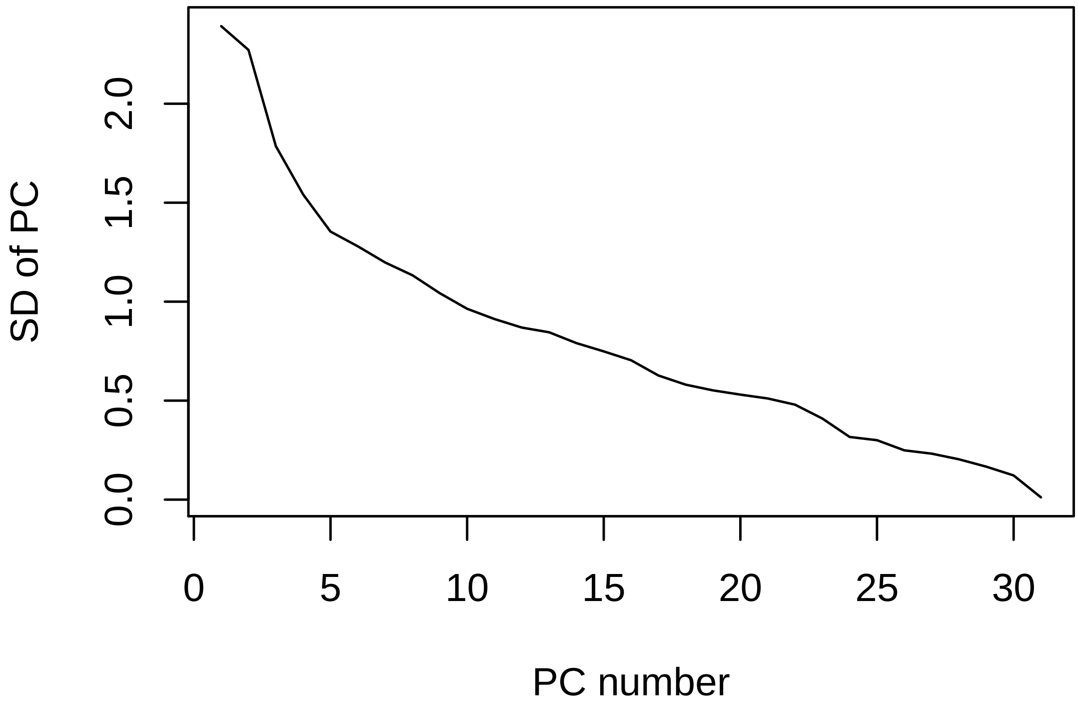


Figure 42: Principal components' variance of kaggle data

We see that first few components contain the majority of variation in the data.
Visualizing 3 first componets as composition of orginal predictors:

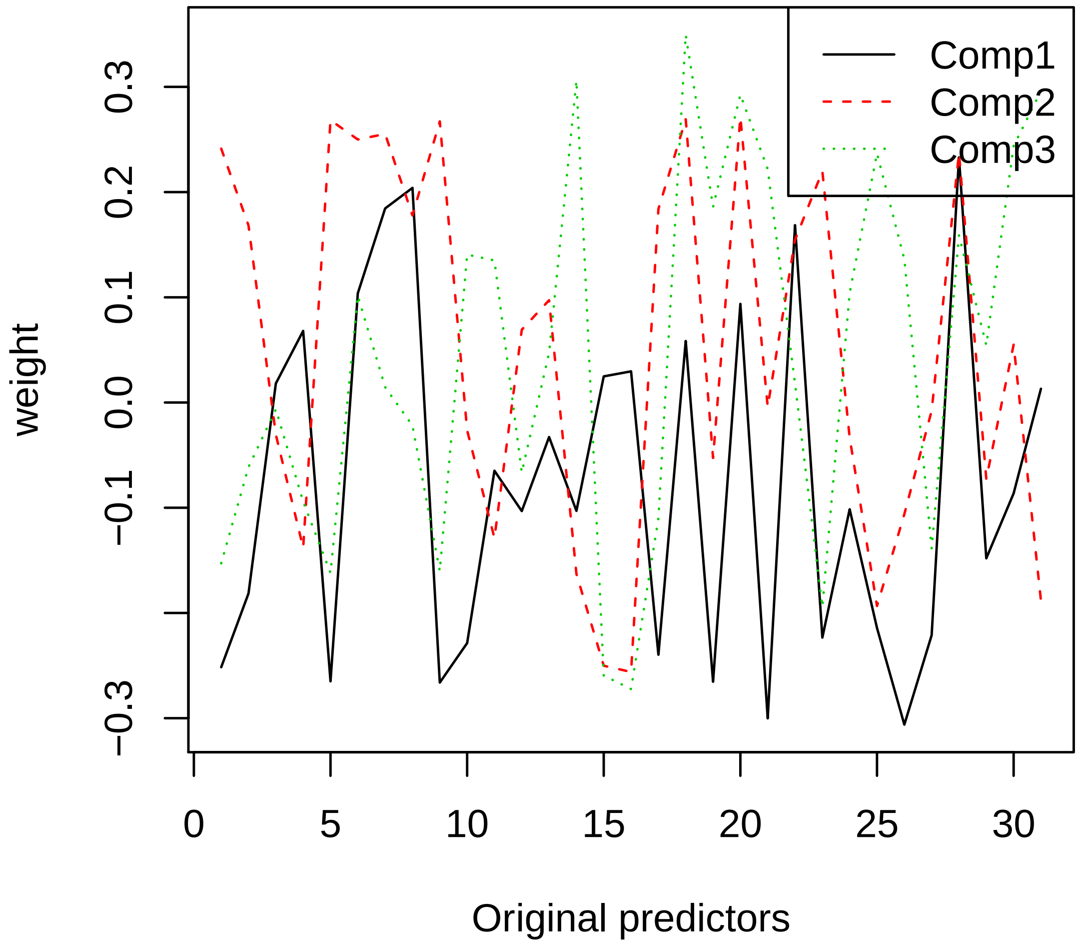


Figure 43: 3 first principal components as composition of original predicotrs.

It can be seen that first 3 principal components somehow weigh orginal predictors in differeent ways but it is difficult to find the systematic pattern there.

PCA can also be used to discover extreme observations in the data. Getting the biggest value for the first component we can find entreme observation:

```
> #Extream observation
> extreme.idx<-which.max(kaggle.pc$x[,1])
> kaggle.data$country[extreme.idx]

[1] Iraq
87 Levels: Afghanistan Albania Antigua & Barbuda Argentina Armenia Austria Azerbaijan ... Venezuela

> kaggle.data.scaled<-scale(kaggle.data[variable.names])
> kaggle.data.scaled[extreme.idx,]

     AG.LND.AGRI.K2     AG.LND.ARBL.HA.PC        AG.LND.ARBL.ZS        AG.LND.CROP.ZS
        -0.30791251           -0.11299188           -0.18701738           -0.53570288
     AG.LND.TOTL.K2        AG.PRD.CROP.XD        AG.PRD.FOOD.XD        AG.PRD.LVSK.XD
        -0.24024095            0.26282849           -0.04531691            1.21019539
     AG.SRF.TOTL.K2        AG.YLD.CREL.KG        BM.GSR.INSF.ZS        BM.GSR.TRVL.ZS
        -0.24513837           -1.07499847            1.44074075           -1.11903857
     BX.GSR.CMCP.ZS BX.KLT.DINV.WD.GD.ZS     EG.GDP.PUSE.KO.PP EG.GDP.PUSE.KO.PP.KD
         1.90884956           -0.25926840           -1.08844310           -0.92712124
  EG.USE.COMM.KT.OE EG.USE.COMM.GD.PP.KD     EG.USE.ELEC.KH.PC     EN.ATM.CO2E.KD.GD
        -0.22867533            1.65595716           -0.42149955            2.35924312
     EN.ATM.CO2E.PC     EN.ATM.PM10.MC.M3        ER.H2O.INTR.K3        ER.H2O.INTR.PC
        -0.29028135            3.74026447           -0.36467701           -0.34207405
```

```
FM.LBL.MQMY.GD.ZS     FS.AST.PRVT.GD.ZS          IC.CRD.PRVT.ZS             IC.EXP.DURS
      -0.26484286           -0.97149414             -0.72584404              4.33963016
   IC.LGL.CRED.XQ        NE.RSB.GNFS.ZS          NE.TRD.GNFS.ZS        Corruption.Index
      -0.97051999            0.44970581             -1.03782687              1.61652583
```

### 6.5 PCR

Principal component regression builds linear model based on the principal components computed like in the previous section. Performance of this model can be checked by computing RMSE (I am using here leave one out crossvalidation) for models built from various number of most important components:

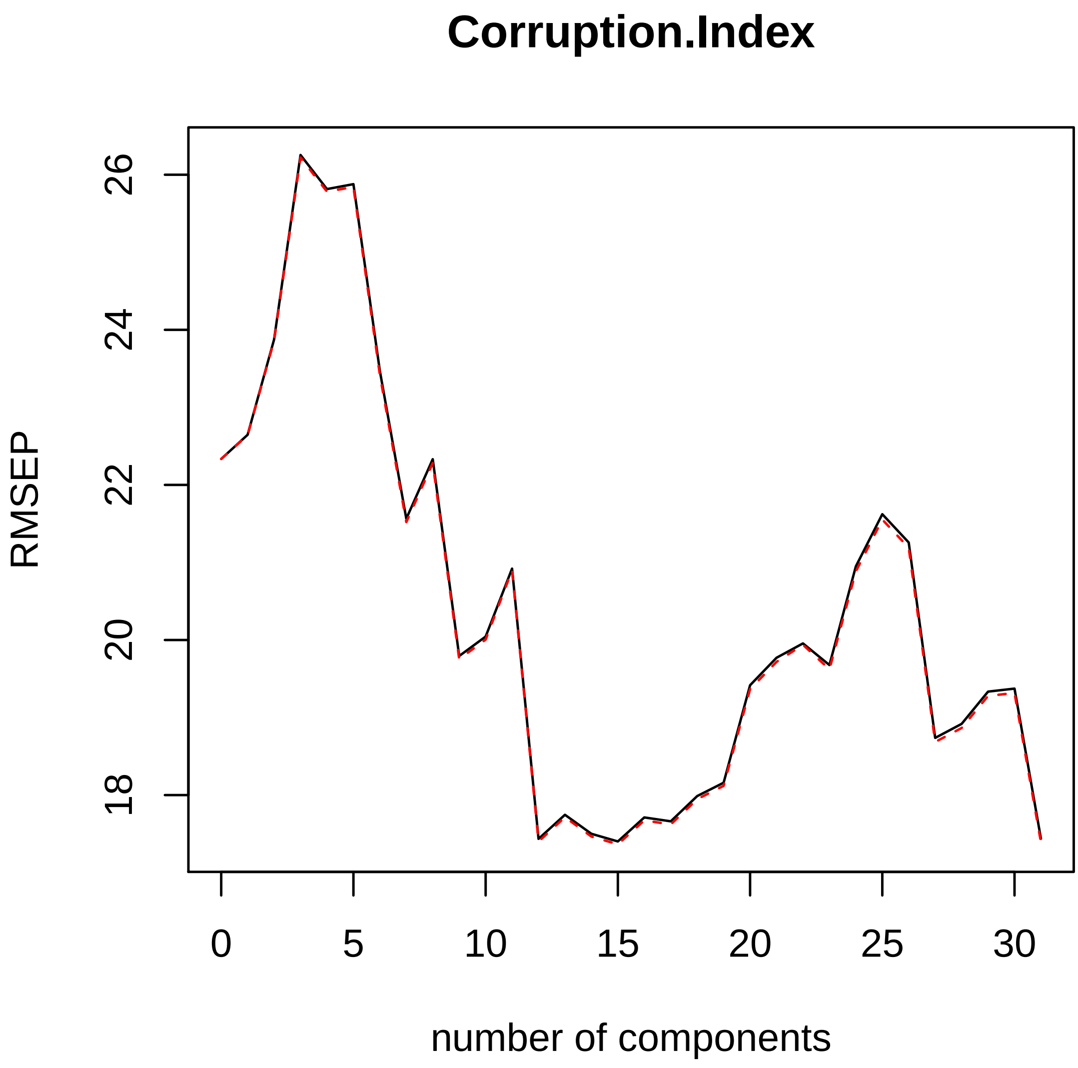


Figure 44: RMSEP ∼ number of componets using leave one out crossvalidation

We see that the best performance is achieved for 15 PCs with RMSE:

```
[1] 17.4021
```

Visualising the components as the linear function of orginal predictors.

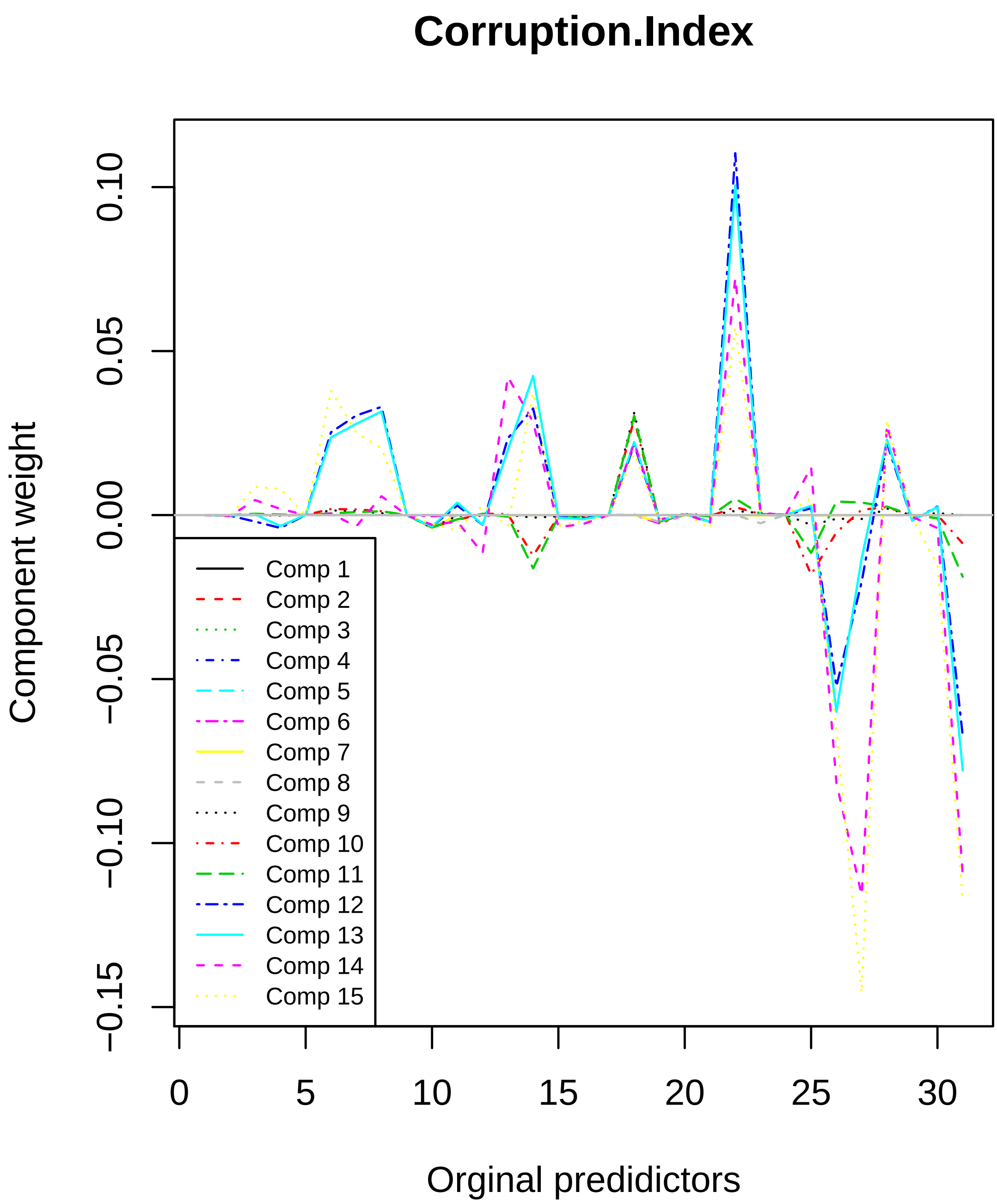


Figure 45: 15 first PCR components as function of orginal predictors

We can see that all components select similar orginal predictors but weigh them with different intensity. Checking if model doesn't violate OLS assumptions:

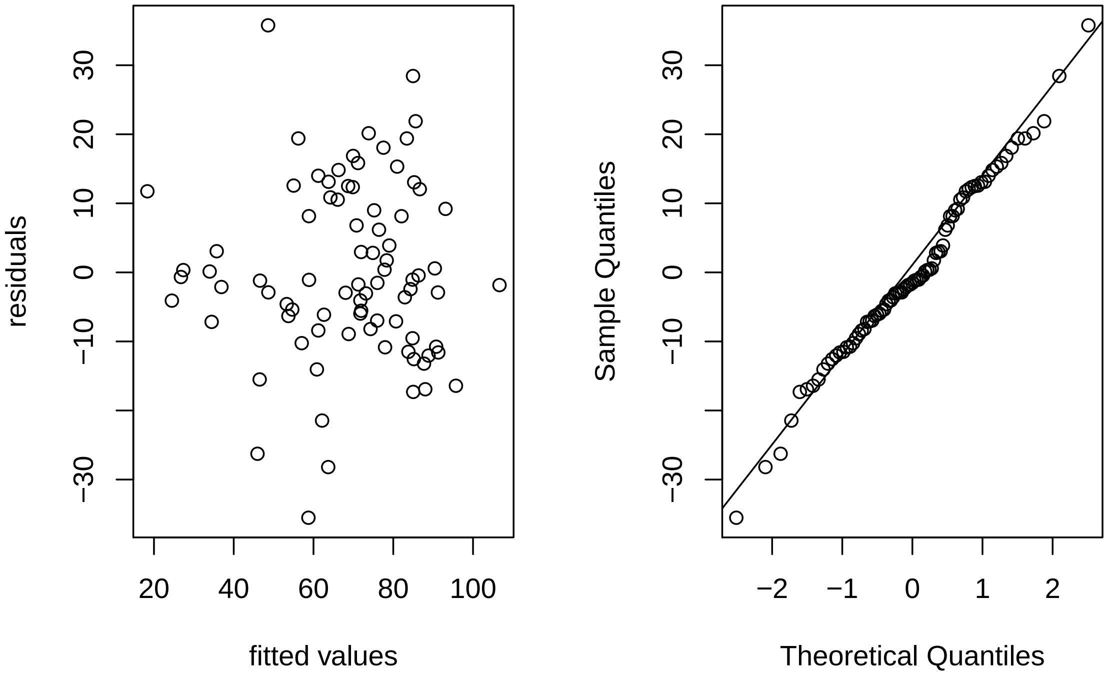


Figure 46: Diagnostic plots for PCR model with 15 components

We see that model is conformant with OLS assumptions like normal distribution of residuals, homogenous variance.

## 6.6 PLSR

Partial least squares regression is similar to PCR (builds predictors based on linear combination of original predictors) with one important difference that it uses information how predictors influence response variable (where PCR ignores this information). Performance of this model can be checked by computing RMSE (I am using here leave one out crossvalidation) for models built from various number of most important components:

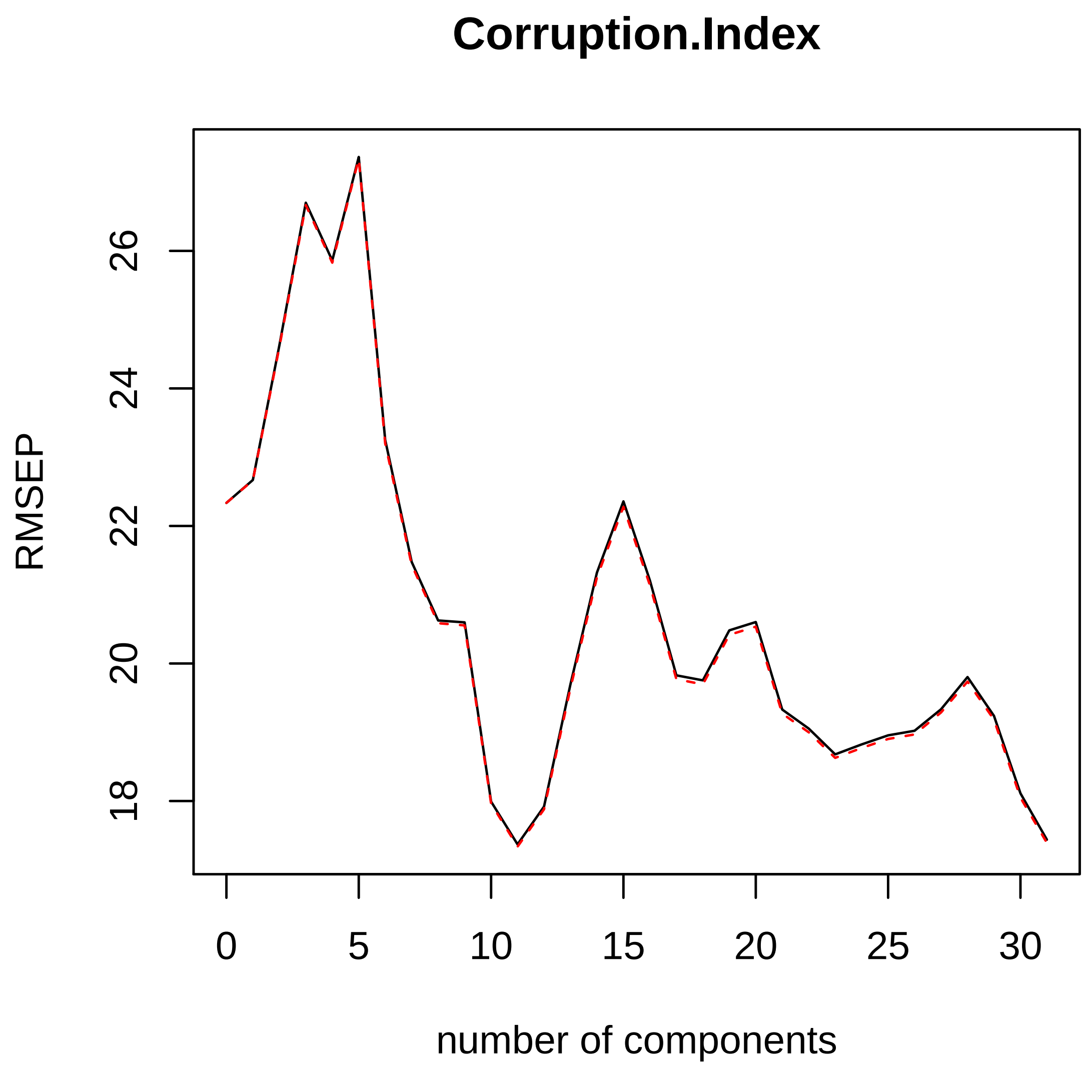


Figure 47: RMSEP ∼ number of componets

We see that the best performance is achieved for 11 components with RMSE:

```
[1] 17.36973
```

So using PLSR allowed us to decrease number of predictors and error as well compared to PCR.

Visualising the components as the linear function of orginal predictors.

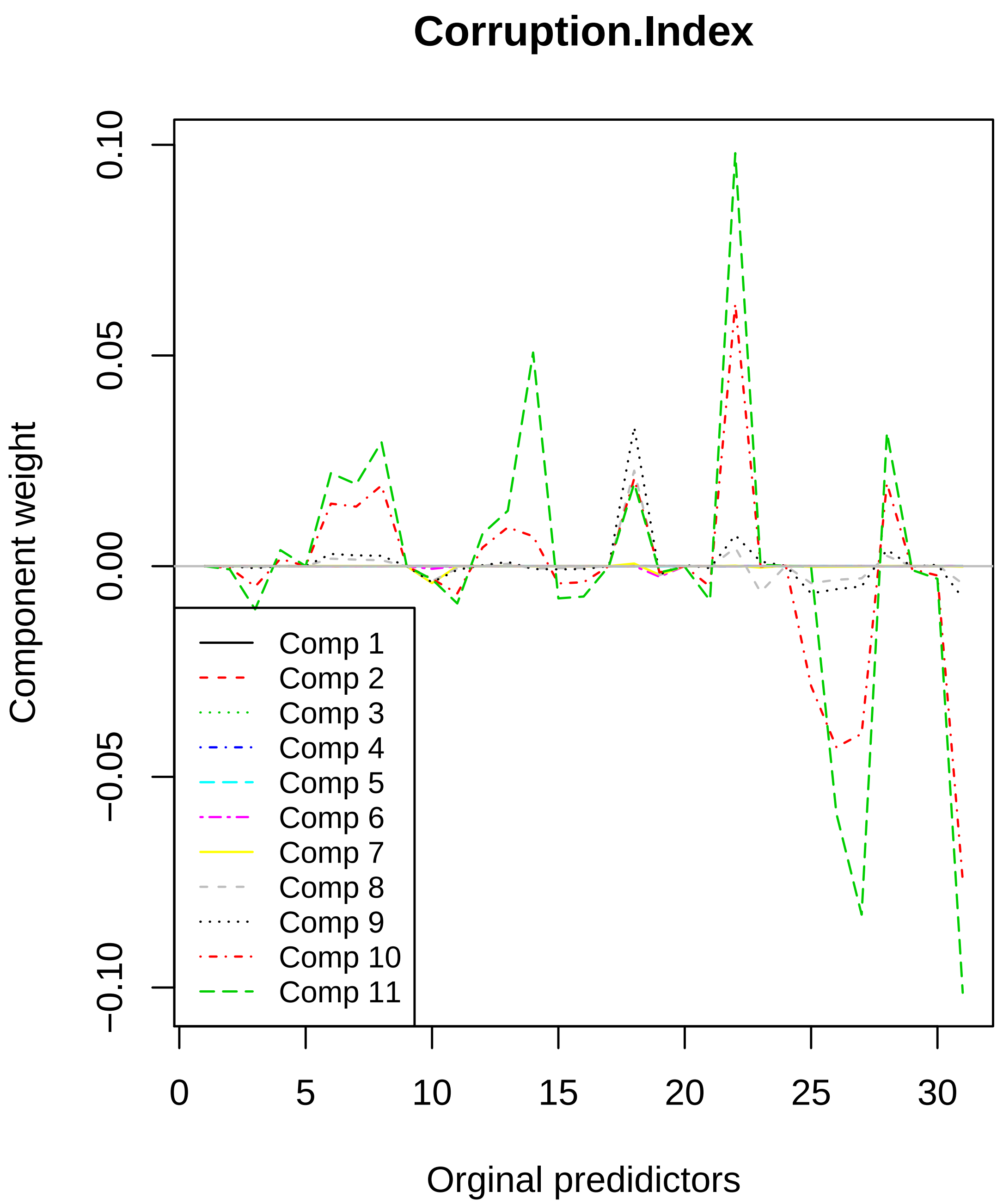


Figure 48: 11 first PLSR components as function of orginal predictors

We can see that all components select similar orginal predictors but weigh them with different intensity.
Checking if model doesn't violate OLS assumptions:

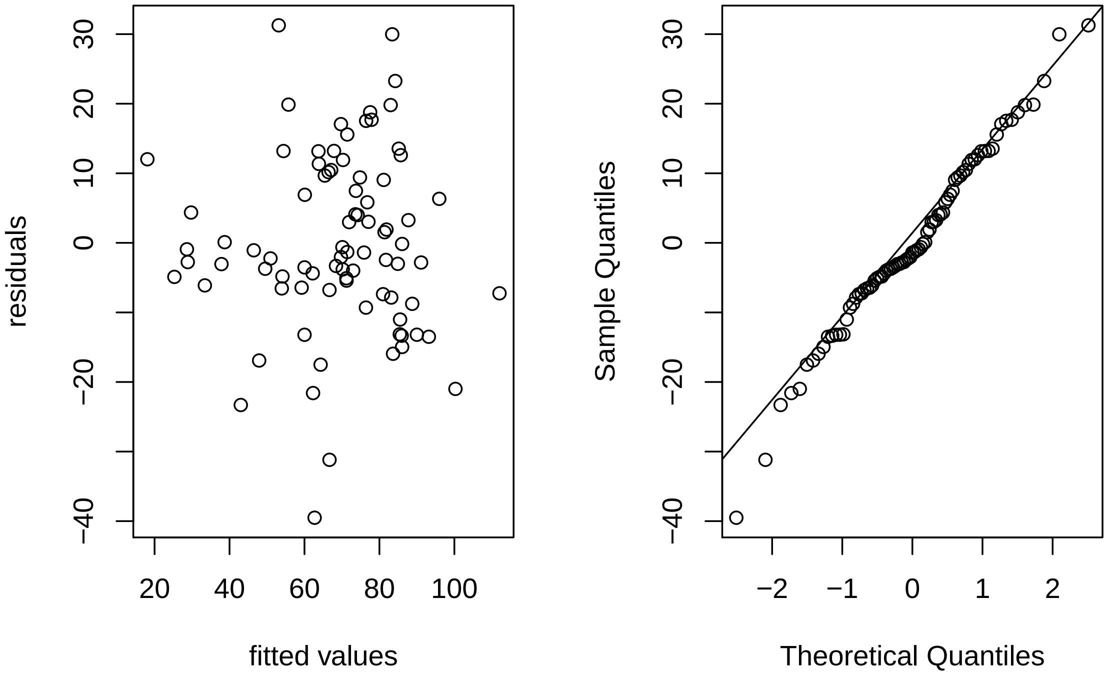


Figure 49: Diagnostic plots for PLSR model with 11 components

We see that model is conformant with OLS assumptions like normal distribution of residuals, homogenous variance.

# 7 Nonlinear regression

## 7.1 Regression tree

Tree having the minimum xerror (crossvalidated error: ratio of $R^{CV}(T)$ and SSE for root) has 3 splits and coresponding SE=0.115:

```
Regression tree:
rpart(formula = Corruption.Index ~ ., data = kaggle.data[variable.names])

Variables actually used in tree construction:
[1] AG.LND.ARBL.ZS       AG.LND.CROP.ZS       BX.KLT.DINV.WD.GD.ZS EG.GDP.PUSE.KO.PP.KD
[5] EG.USE.ELEC.KH.PC    EN.ATM.CO2E.PC       IC.EXP.DURS

Root node error: 40413/83 = 486.9

n= 83

        CP nsplit rel error  xerror    xstd
1 0.419584      0   1.00000 1.01450 0.14153
2 0.180685      1   0.58042 0.82332 0.14397
3 0.061492      2   0.39973 0.73449 0.14563
4 0.034589      3   0.33824 0.67533 0.11516
5 0.033165      4   0.30365 0.70010 0.11133
6 0.028088      5   0.27048 0.68418 0.11126
7 0.025676      6   0.24240 0.68418 0.11126
8 0.010000      7   0.21672 0.67899 0.11425
```

The same selection process depicted graphically:

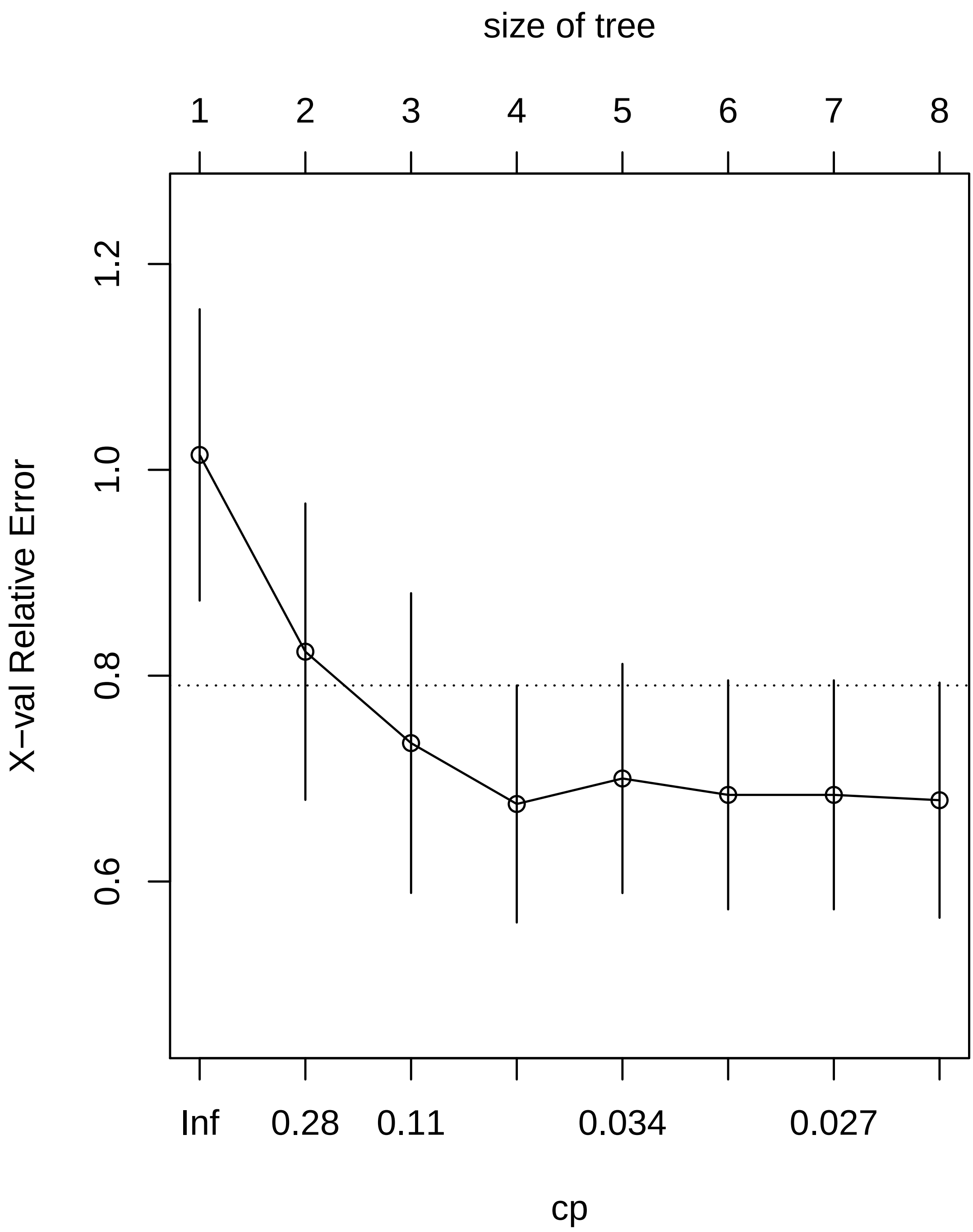


Figure 50: Crossvalidated error as a function of number of splits

Built tree:

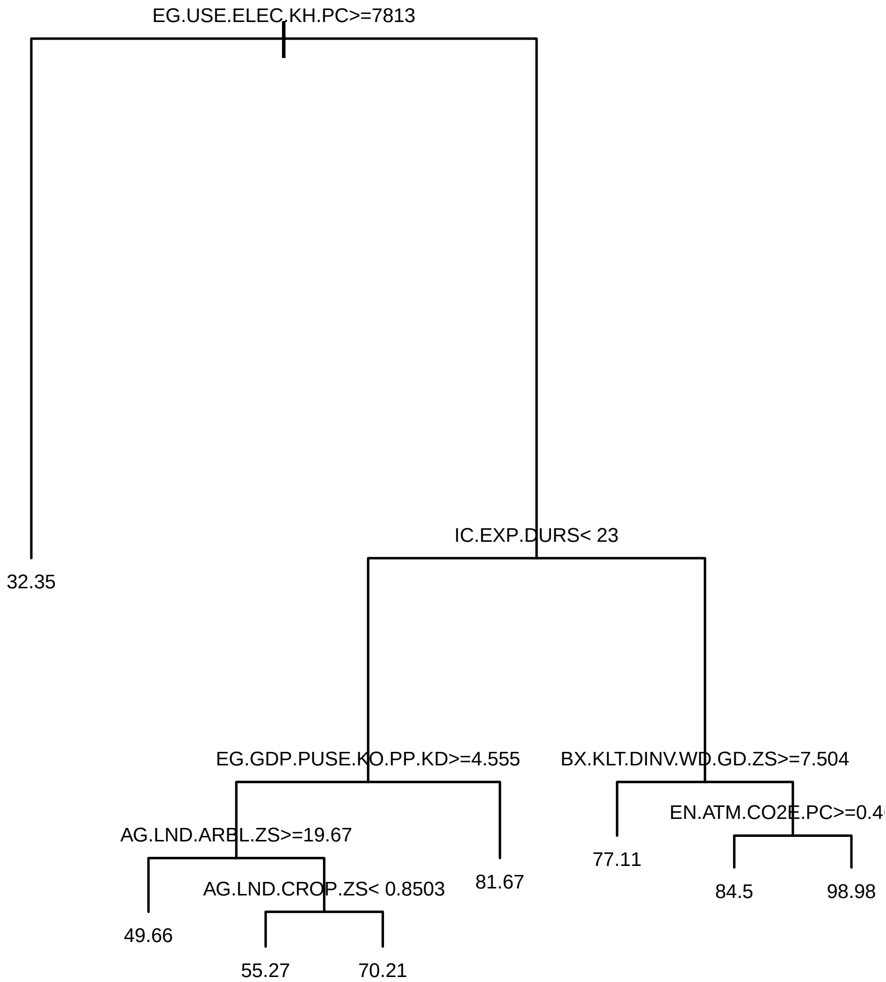


Figure 51: Regression tree

## 7.2 Moving average estimator of regression function

Fitting models with different values of span, because only 4 predictors are allowed:

It can be seen that small values of smoothing parameter results in the model that tries to fits the data too closely and is over-fitted.

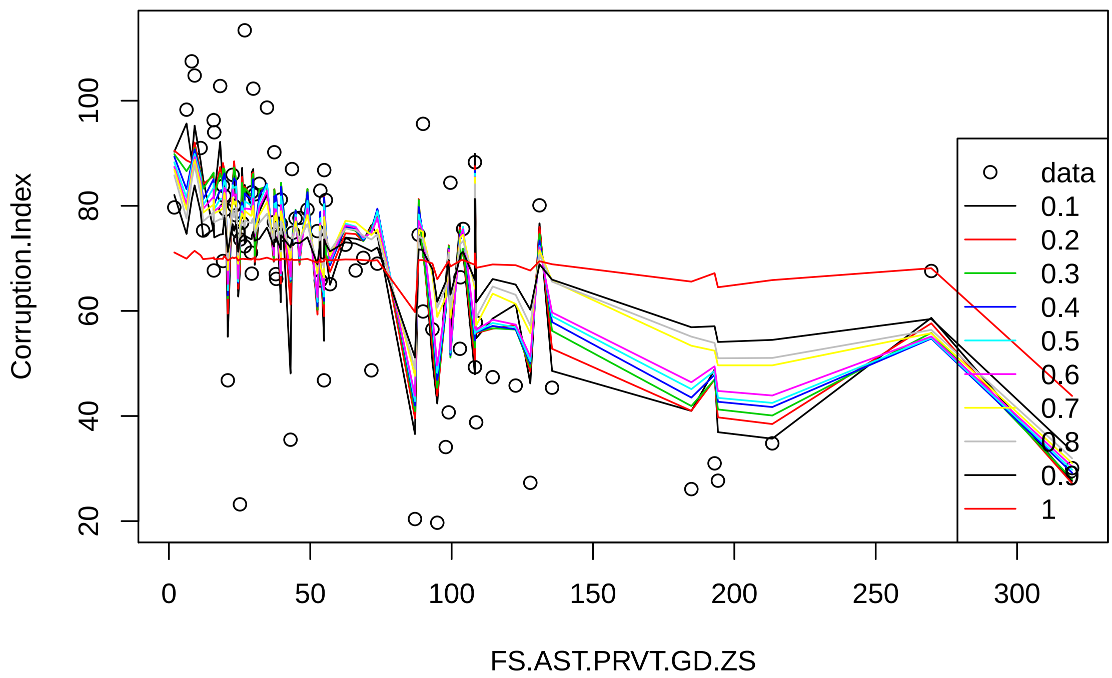


Figure 52: moving average model for different values of smoothing parameter

### 7.3 Local linear estimator of regression function

Fitting models with different values of span:

It can be seen that small values of smoothing parameter results in the model that tries to fits the data too closely and is over-fitted.

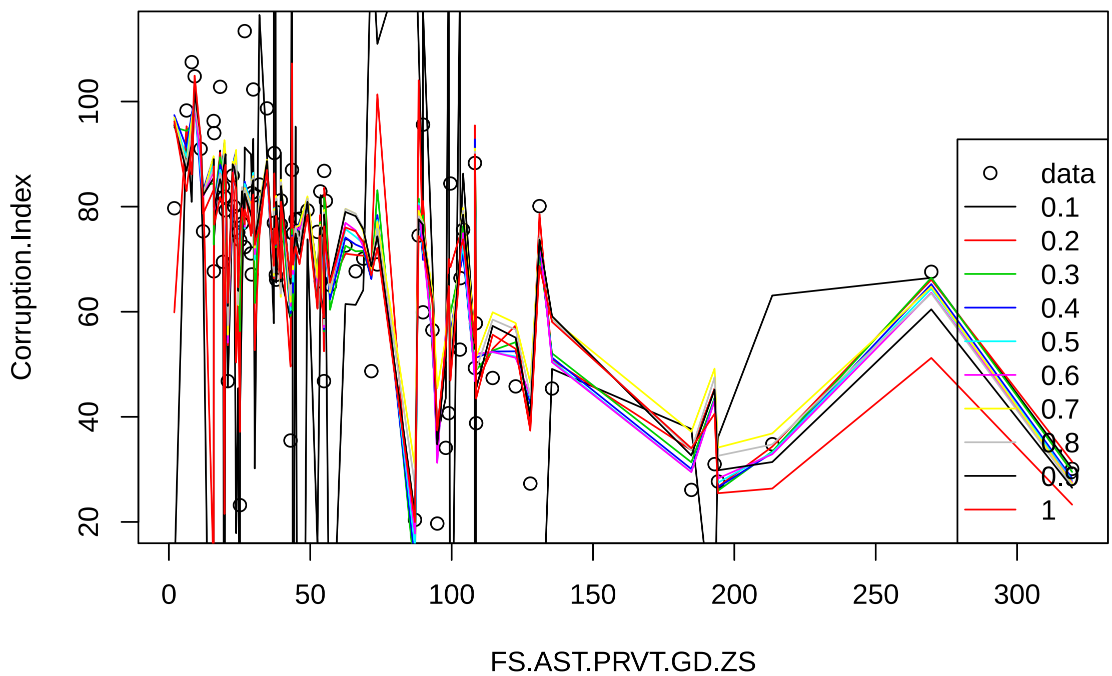


Figure 53: local quadratic estimator model for different values of smoothing parameter

## 7.4 Additive model

Selected additive model based on forward procedure with t-test:

```
> summary(model.additive)

Family: gaussian
Link function: identity

Formula:
Corruption.Index ~ s(EG.USE.ELEC.KH.PC) + s(IC.EXP.DURS) + s(EG.GDP.PUSE.KO.PP.KD) +
    s(AG.LND.ARBL.ZS) + s(IC.CRD.PRVT.ZS) + s(NE.TRD.GNFS.ZS)
<environment: 0x7fe2747d05f8>

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   68.913      1.307   52.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df      F  p-value
s(EG.USE.ELEC.KH.PC)    2.240  2.705 20.352 3.17e-09 ***
s(IC.EXP.DURS)          1.000  1.000  3.795  0.05534 .
s(EG.GDP.PUSE.KO.PP.KD) 5.723  6.712  3.075  0.00771 **
s(AG.LND.ARBL.ZS)       1.000  1.000  6.891  0.01058 *
s(IC.CRD.PRVT.ZS)       1.000  1.000  9.029  0.00366 **
s(NE.TRD.GNFS.ZS)       1.000  1.000  7.421  0.00809 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.712   Deviance explained = 75.4%
GCV score = 167.99  Scale est. = 141.75    n = 83
```

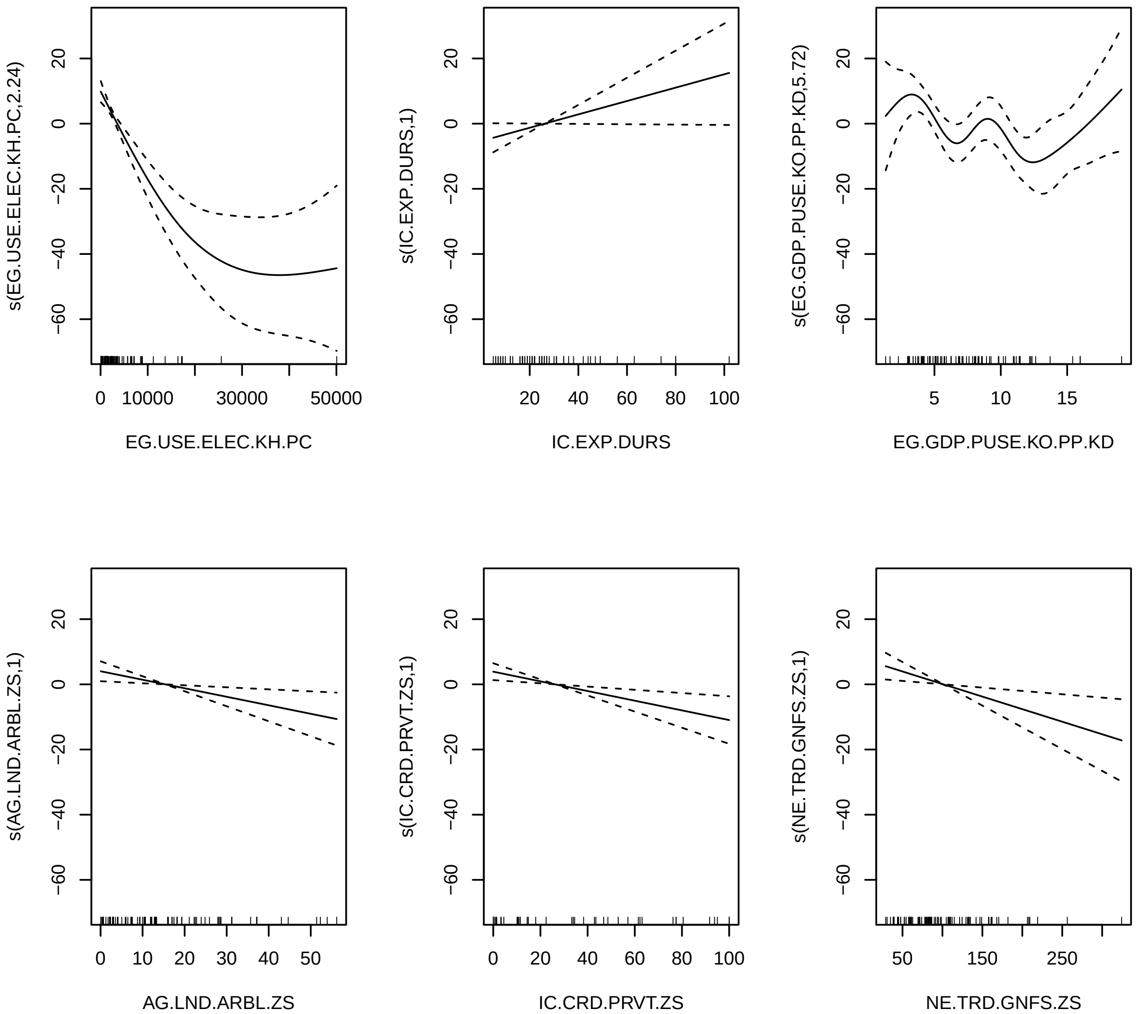


Figure 54: Plot of fuctions that constitute additive model

# 8 Prediction

The ultimate test for the model is the prediction of the out of sample data. I use United Kindom data from 2007 which is not included in the data set but its data can be found on Internet [2].Of course the model has been built from the data gathered in unknown year so we have to assume that the model is still valid in 2007 (Qualitative extrapolation [3]). Also checking model by predicting outcome variable that can be verified with real life data gives better feel of model quality than for example trying to explain the value of cooeficients [3]. The prediction quality also depends on the Mahalanobis distance of the new observation from the cases used to build the model which equals: 6.28. Because the mean value of observation distances from their center equals 5.35 and the standard deviation of those distances equals 1.43 we can predict the index for UK. The real value of Corruption.Index for UK is 34.1 [1].

| Model | Prediction |
|---|---|
| Full model | `fit lwr upr` <br> `49.35854 18.27891 80.43818` |
| VIF model | `fit lwr upr` <br> `44.32452 14.23575 74.41329` |
| Backward t-test | `fit lwr upr` <br> `43.74599 16.92364 70.56835` |
| Backward t-test with quadratic term | `fit lwr upr` <br> `45.77669 19.88974 71.66364` |
| Forward t-test | `fit lwr upr` <br> `47.08985 20.28054 73.89915` |
| Backward AIC | `fit lwr upr` <br> `47.92342 20.89829 74.94854` |
| Forward AIC | `fit lwr upr` <br> `41.28042 14.60942 67.95141` |
| Forward AIC with quadratic term | `fit lwr upr` <br> `41.93812 16.16607 67.71017` |
| Backward BIC | `fit lwr upr` <br> `39.8606 11.67301 68.04819` |
| Forward BIC | `fit lwr upr` <br> `48.88516 21.56853 76.20179` |
| Forward BIC with quadratic term | `fit lwr upr` <br> `45.87882 18.91644 72.84119` |
| Adjusted R2 | `fit lwr upr` <br> `49.04343 21.75061 76.33624` |
| Adjusted R2 with quadratic term | `fit lwr upr` <br> `49.81643 23.02251 76.61035` |

| | |
|---|---|
| Cp | `fit lwr upr`<br>`40.74901 13.93553 67.5625` |
| Cp with quadratic term | `fit lwr upr`<br>`42.23055 16.13695 68.32415` |
| PCR | `38.87232` |
| PLSR | `39.45651` |
| Lasso | 43.34699 |
| Ridge | 41.98209 |
| Least trimmed squares | 45.88068 |
| M-estimator(Huber) | `fit lwr upr`<br>`42.47002 20.80984 64.1302` |
| Regression tree | `49.6625` |
| Additive | 48.59347 |

Table 2: Prdicted index for United Kindom for different models

We can see that predcting new values gives very wide intervals which contain the true value.

# 9 Crossvalidation and final model selection

I decided to select model based on leave one out crossvalidation to choose model with the best potential to predict index for new data. Each model is recomputed with one observation left out (for each observation) and then RMSE (root mean square error is computed) of all models is calculated. I select model with the smallest crossvalidation RMSE.

| **Model** | **Prediction** |
|---|---|
| Full model | `17.43758` |
| VIF model | `18.06011` |
| Backward t-test | `14.3501` |
| Backward t-test with quadratic term | `16.57995` |
| Forward t-test | `14.18182` |
| Backward AIC | `13.93362` |
| Forward AIC | `14.37905` |

| | |
|---|---|
| Forward AIC with quadratic term | 13.69938 |
| Backward BIC | 14.17277 |
| Forward BIC | 14.32132 |
| Forward BIC with quadratic term | 14.17384 |
| Adjusted R2 | 14.45955 |
| Adjusted R2 with quadratic term | 14.10968 |
| Cp | 14.0058 |
| Cp with quadratic term | 13.54951 |
| Principal component regression | 17.4021 |
| Partial least squares regression | 17.36973 |
| Ridge regression | 22.605 |
| Regression tree | 18.38257 |

Table 3: Leave one out crossvalidation RMSE for all models

Besed on the Table-3 I select model obtainen by the Cp with quadratic term because of the smallest corss validation error. This model selection (based on Cp criterion) aims at choosing the best model for prediction tasks (selection procedure in this case minimizes average mean sqare error). The Cp model with quadratic term:

```
                         Estimate  Std. Error t value  Pr(>|t|)
(Intercept)             1.5606e+02  1.6991e+01  9.1847 3.589e-13 ***
AG.LND.ARBL.HA.PC      -1.6057e+01  5.5692e+00 -2.8831  0.005406 **
AG.LND.ARBL.ZS         -2.8621e-01  1.1041e-01 -2.5923  0.011875 *
AG.LND.CROP.ZS         -2.8151e-01  2.0406e-01 -1.3796  0.172679
AG.PRD.CROP.XD         -2.4866e-01  1.0277e-01 -2.4197  0.018480 *
AG.SRF.TOTL.K2          6.0943e-07  6.4591e-07  0.9435  0.349079
AG.YLD.CREL.KG         -2.7760e-03  1.0115e-03 -2.7446  0.007915 **
BM.GSR.INSF.ZS         -2.0882e-01  2.0494e-01 -1.0189  0.312196
BM.GSR.TRVL.ZS          4.0353e-01  1.3924e-01  2.8982  0.005182 **
BX.KLT.DINV.WD.GD.ZS    3.3210e-01  1.3861e-01  2.3959  0.019612 *
EG.GDP.PUSE.KO.PP      -5.6541e+00  2.1399e+00 -2.6422  0.010413 *
EG.USE.ELEC.KH.PC      -1.2730e-03  3.7490e-04 -3.3956  0.001200 **
EN.ATM.CO2E.KD.GD       1.0075e+00  1.0128e+00  0.9947  0.323753
EN.ATM.CO2E.PC         -1.2048e+00  4.3729e-01 -2.7553  0.007689 **
EN.ATM.PM10.MC.M3       7.3750e-02  4.5031e-02  1.6378  0.106535
```

```
FM.LBL.MQMY.GD.ZS      -1.1505e-01  6.3879e-02 -1.8010  0.076561 .
FS.AST.PRVT.GD.ZS       7.1091e-02  6.3217e-02  1.1246  0.265112
IC.CRD.PRVT.ZS         -1.0506e-01  5.6317e-02 -1.8655  0.066848 .
NE.TRD.GNFS.ZS         -1.1873e-01  4.0661e-02 -2.9199  0.004876 **
I(EG.GDP.PUSE.KO.PP^2)  2.2378e-01  1.0514e-01  2.1285  0.037278 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 10 Summary

Models also fulfill explanatory function for the data.

| Method | AG.LND.AGRI.K2 | AG.LND.ARBL.HA.PC | AG.LND.ARBL.ZS | AG.LND.CROP.ZS | AG.LND.TOTL.K2 | AG.PRD.CROP.XD | AG.PRD.FOOD.XD | AG.PRD.LVSK.XD | AG.SRF.TOTL.K2 | AG.YLD.CREL.KG | BM.GSR.INSF.ZS | BM.GSR.TRVL.ZS | BX.GSR.CMCP.ZS | BX.KLT.DINV.WD.GD.ZS | EG.GDP.PUSE.KO.PP | EG.GDP.PUSE.KO.PP.KD | EG.USE.COMM.KT.OE | EG.USE.COMM.GD.PP.KD | EG.USE.ELEC.KH.PC | EN.ATM.CO2E.KD.GD | EN.ATM.CO2E.PC | EN.ATM.PM10.MC.M3 | ER.H2O.INTR.K3 | ER.H2O.INTR.PC | FM.LBL.MQMY.GD.ZS | FS.AST.PRVT.GD.ZS | IC.CRD.PRVT.ZS | IC.EXP.DURS | IC.LGL.CRED.XQ | NE.RSB.GNFS.ZS | NE.TRD.GNFS.ZS | Systematic part exmplained | Sigma |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Full model | + | **-** | - | - | - | - | + | - | + | **-** | - | + | - | **+** | - | - | - | - | **-** | + | - | + | - | + | - | + | - | + | - | + | **-** | 0.639 | 13.333 |
| VIF Backward | + | - | - | - | | - | | - | | **-** | - | + | - | **+** | **-** | | - | | - | + | - | + | - | + | | + | - | + | - | + | **-** | 0.638 | 13.361 |
| Backward t-test | | **-** | **-** | | | - | | | | **-** | | **+** | | + | **-** | | | | **-** | + | **-** | + | | | - | | **-** | | | | **-** | 0.670 | 12.539 |
| Forward t-test | | | **-** | | | | | | | | | | | | | **-** | | | **-** | | - | | | | | | **-** | **+** | | | **-** | 0.646 | 13.041 |
| Backward AIC | | **-** | **-** | | | - | | | | **-** | - | **+** | | **+** | **-** | | | | **-** | + | **-** | **+** | | | **-** | + | **-** | | | | **-** | 0.689 | 12.389 |
| Forward AIC | | **-** | - | | | | | | | **-** | | **+** | | **+** | | **-** | | | **-** | | - | + | | + | - | | **-** | + | | | **-** | 0.680 | 12.554 |
| Backward BIC | | | | | | | | | | **-** | | | | | **-** | | | | **-** | | - | **+** | | | | | **-** | | | | | 0.619 | 13.519 |
| Forward BIC | | | | | | | | | | | | | | | | **-** | | | **-** | | - | | | | | | **-** | **+** | | | **-** | 0.631 | 13.319 |
| Adj R2 | | **-** | **-** | - | | **-** | | | + | **-** | - | **+** | | **+** | **-** | | | | **-** | + | **-** | **+** | | | **-** | + | **-** | | | | **-** | 0.691 | 12.349 |
| Cp | | **-** | | | | | | | | **-** | | + | | **+** | | **-** | | | **-** | | | **+** | | **+** | **-** | | **-** | | | | **-** | 0.674 | 12.675 |

Table 4: Models and selected predictors. "+" means that cooeficient in the model is positive and "-" means the cooeficient is negative. Lack of sign means that predictor is not used in the model. Red color means that coefficient is significant.

We can see from the table how particular predictors influence the Corruption.Index (holding all other predictors constant). Of course there is a danger of lurking variables and we cannot make strong conclusions but we can at least try to discover some trends:

- AG.LND.AGRI.K2 positively influences Corruption.Index so we can assume that more agricultural land then country could be more corrupted.
- AG.LND.ARBL.HA.PC negativel influences Corruption.Index so we can assume that more arable land per person then country could be less corrupted
- AG.YLD.CREL.KG negatively influences Corruption.Index so we can assume that more Cereal yield then country could be less corrupted (better efficiency)
- BM.GSR.TRVL.ZS positively influences Corruption.Index so we can assume that if country has more turism as share of its economy then country could be more corrupted. (southern european countries)
- BX.KLT.DINV.WD.GD.ZS positively influences Corruption.Index so we can assume that bigger foreign direct investment as share of its economy then country could be more corrupted. (looks like coruption helps direct investment?)
- EG.GDP.PUSE.KO.PP negatively influences Corruption.Index so we can assume that more GDP per unit of energy use then country could be less corrupted (more corrupted countries have less energy hungry economies).
- EG.USE.ELEC.KH.PC negatively influences Corruption.Index so we can assume that more electric power consumption per capita then country could be less corrupted (more corrupted countries have less energy hungry economies)

- EN.ATM.CO2E.PC CO2 negatively influences Corruption.Index so we can assume that bigger emissions per capita the country could be less corrupted (more corrupted countries have less energy hungry economies)
- EN.ATM.PM10.MC.M3 positively influences Corruption.Index so we can assume the bigger polution the country could be more corrupted.
- IC.CRD.PRVT.ZS negatively influences Corruption.Index so we can assume the more state information on business the country could be less corrupted
- IC.EXP.DURS positively influences Corruption.Index so we can assume the more time to export the country could be more corrupted
- NE.TRD.GNFS.ZS negatively influences Corruption.Index so we can assume the more trade as proportion of GDP the country could be less corrupted

We can see that all those assosiations make logical sense.

## 11 Methods that failed

Here I note the methods that I tried but haven't imroved the model

- Taking log of predictor which is right skewed

## References


[1] `http://ffp.statesindex.org/rankings-2007-sortable` Corruption index 2007

[2] `http://data.worldbank.org/` Worldbank data

[3] Julian J. Faraway "Linear Models with R"

[4] `http://en.wikipedia.org/wiki/List_of_countries_by_Failed_States_Index`