Advanced Statistical Methods - Project.

Pawel Chilinski

December 29, 2013

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	ordinal		

Describing the data:

```
> options(width=400)
> describe(kaggle.data)
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
country*	1	87	44.00	25.26	44.00	44.00	32.62	1.00	87.00	86.00	0.00	-1.24	2.71
AG.LND.AGRI.K2	2	87	381134.93	950246.05	47930.00	132379.06	70082.50	50.00	5344172.00	5344122.00	3.63	13.39	101877.07
AG.LND.ARBL.HA.PC	3	87	0.25	0.34	0.17	0.19	0.16	0.00	2.57	2.56	4.38	24.77	0.04
AG.LND.ARBL.ZS	4	87	14.83	13.75	10.67	12.83	11.40	0.06	56.17	56.12	1.22	0.94	1.47
AG.LND.CROP.ZS	5	87	4.81	8.12	1.49	3.02	1.83	0.02	48.96	48.94	2.97	10.37	0.87
AG.LND.TOTL.K2	6	87	1038205.52	2637135.94	156000.00	345147.35	221604.22	28.00	16381390.00	16381362.00	3.70	14.53	282730.65
AG.PRD.CROP.XD	7	87	124.30	19.95	120.00	122.23	16.31	90.00	195.00	105.00	1.05	1.03	2.14
AG.PRD.FOOD.XD	8	87	123.89	18.45	122.00	122.23	17.79	98.00	191.00	93.00	1.02	1.41	1.98
AG.PRD.LVSK.XD	9	87	125.61	19.68	126.00	124.82	25.20	95.00	177.00	82.00	0.29	-0.94	2.11
AG.SRF.TOTL.K2	10	87	1080836.41	2758017.22	163820.00	354817.89	234324.93	28.00	17098240.00	17098212.00	3.71	14.55	295690.48
AG.YLD.CREL.KG	11	87	3229.53	1761.32	2807.70	3022.96	1675.63	812.60	9631.90	8819.30	1.08	1.14	188.83
BM.GSR.INSF.ZS	12	87	11.02	9.79	9.29	9.70	5.78	-2.55	61.87	64.42	2.49	9.15	1.05
BM.GSR.TRVL.ZS	13	87	28.58	14.06	27.70	27.48	12.37	2.51	71.02	68.51	0.87	1.08	1.51
BX.GSR.CMCP.ZS	14	87	39.46	21.88	35.46	38.13	22.80	3.04	100.00	96.96	0.59	-0.38	2.35
BX.KLT.DINV.WD.GD.ZS	15	87	17.02	56.76	6.31	8.66	4.89	0.11	524.88	524.77	8.26	70.72	6.09
EG.GDP.PUSE.KO.PP	16	87	7.90	4.18	7.02	7.52	4.20	1.44	18.48	17.04	0.74	-0.19	0.45
EG.GDP.PUSE.KO.PP.KD	17	87	7.34	4.03	6.24	6.86	3.27	1.33	19.10	17.77	1.09	0.77	0.43
EG.USE.COMM.KT.OE	18	87	111238.03	348862.76	13578.00	37238.44	19751.20	42.00	2336546.00	2336504.00	5.26	28.74	37402.01
EG.USE.COMM.GD.PP.KD	19	87	234.18	193.76	185.99	197.67	105.76	61.29	1219.64	1158.35	3.18	12.53	20.77
EG.USE.ELEC.KH.PC	20	87	4431.40	7196.19	2017.49	2834.37	1786.62	49.15	50067.10	50017.95	3.74	17.82	771.51
EN.ATM.CO2E.KD.GD	21	87	1.58	1.95	0.83	1.13	0.56	0.20	11.33	11.13	2.91	9.12	0.21
EN.ATM.CO2E.PC	22	87	5.80	8.66	3.64	4.28	4.50	0.02	69.15	69.13	4.75	30.53	0.93
EN.ATM.PM10.MC.M3	23	87	53.14	38.43	42.92	47.01	25.28	7.44	212.39	204.95	1.97	4.62	4.12
ER.H2O.INTR.K3	24	87	362.35	919.27	35.20	117.01	51.03	0.02	5418.00	5417.98	3.58	13.44	98.56
ER.H2O.INTR.PC	25	87	27325.44	76742.66	2438.18	10789.31	3072.69	25.65	590277.78	590252.13	5.37	33.24	8227.68
FM.LBL.MQMY.GD.ZS	26	87	78.13	77.37	58.99	65.75	40.59	14.52	636.51	621.99	4.62	29.02	8.29
FS.AST.PRVT.GD.ZS	27	87	65.27	57.88	44.78	55.74	36.89	1.90	319.47	317.57	1.94	4.58	6.21
IC.CRD.PRVT.ZS	28	87	24.97	35.65	1.50	19.34	2.22	0.00	100.00	100.00	1.13	-0.31	3.82
IC.EXP.DURS	29	87	27.02	18.19	21.00	24.01	7.41	5.00	102.00	97.00	1.84	3.60	1.95
IC.LGL.CRED.XQ	30	87	5.24	2.37	5.00	5.23	2.97	1.00	10.00	9.00	0.08	-1.18	0.25
NE.RSB.GNFS.ZS	31	87	3.36	15.73	1.37	1.89	11.86	-31.04	47.49	78.53	0.82	0.78	1.69
NE.TRD.GNFS.ZS	32	87	101.60	53.80	86.49	95.33	41.76	28.97	324.33	295.35	1.36	2.51	5.77
Corruption.Index	33	87	69.20	22.02	74.50	70.30	14.68	19.70	113.40	93.70	-0.52	-0.44	2.36

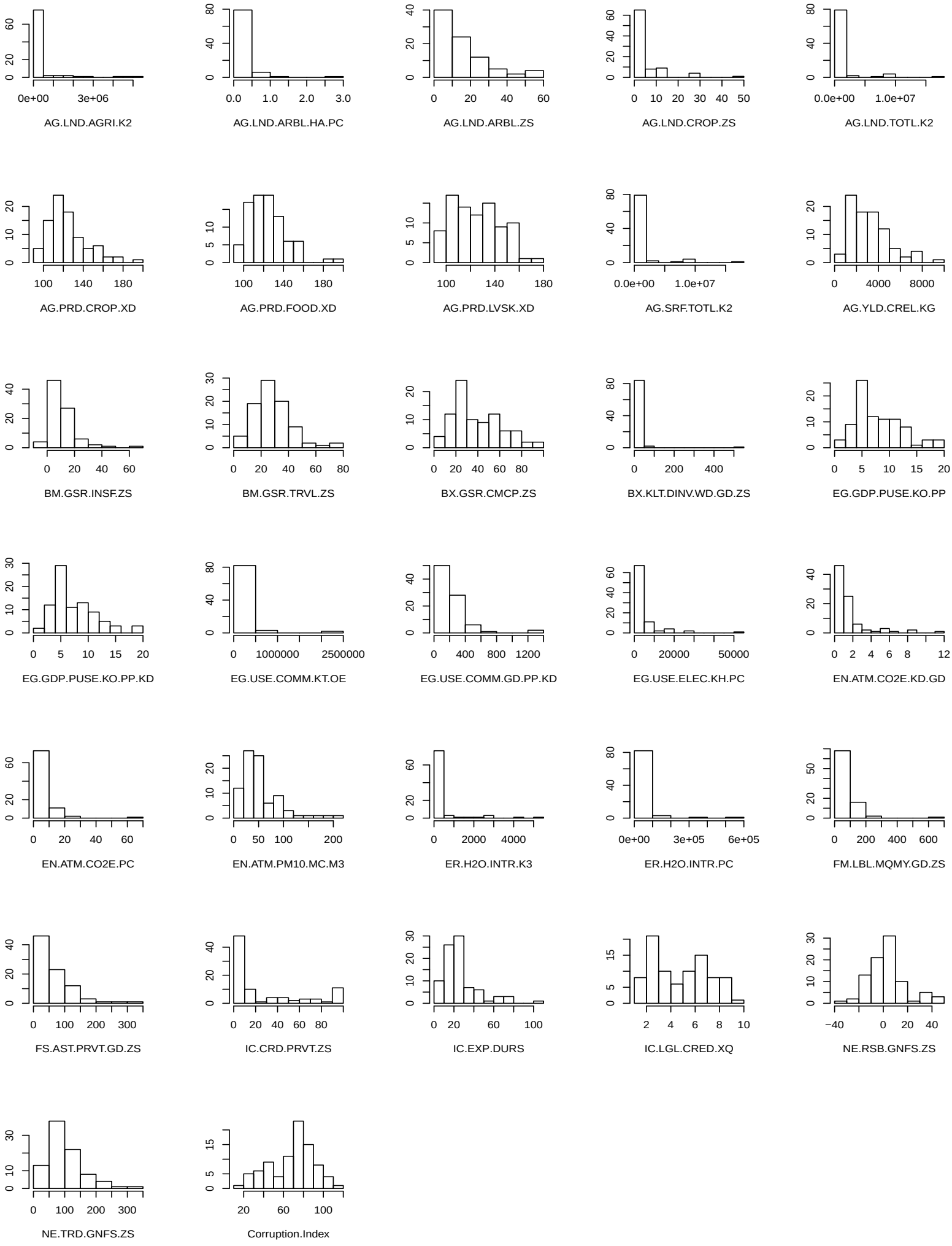


Figure 1: Histograms of variables

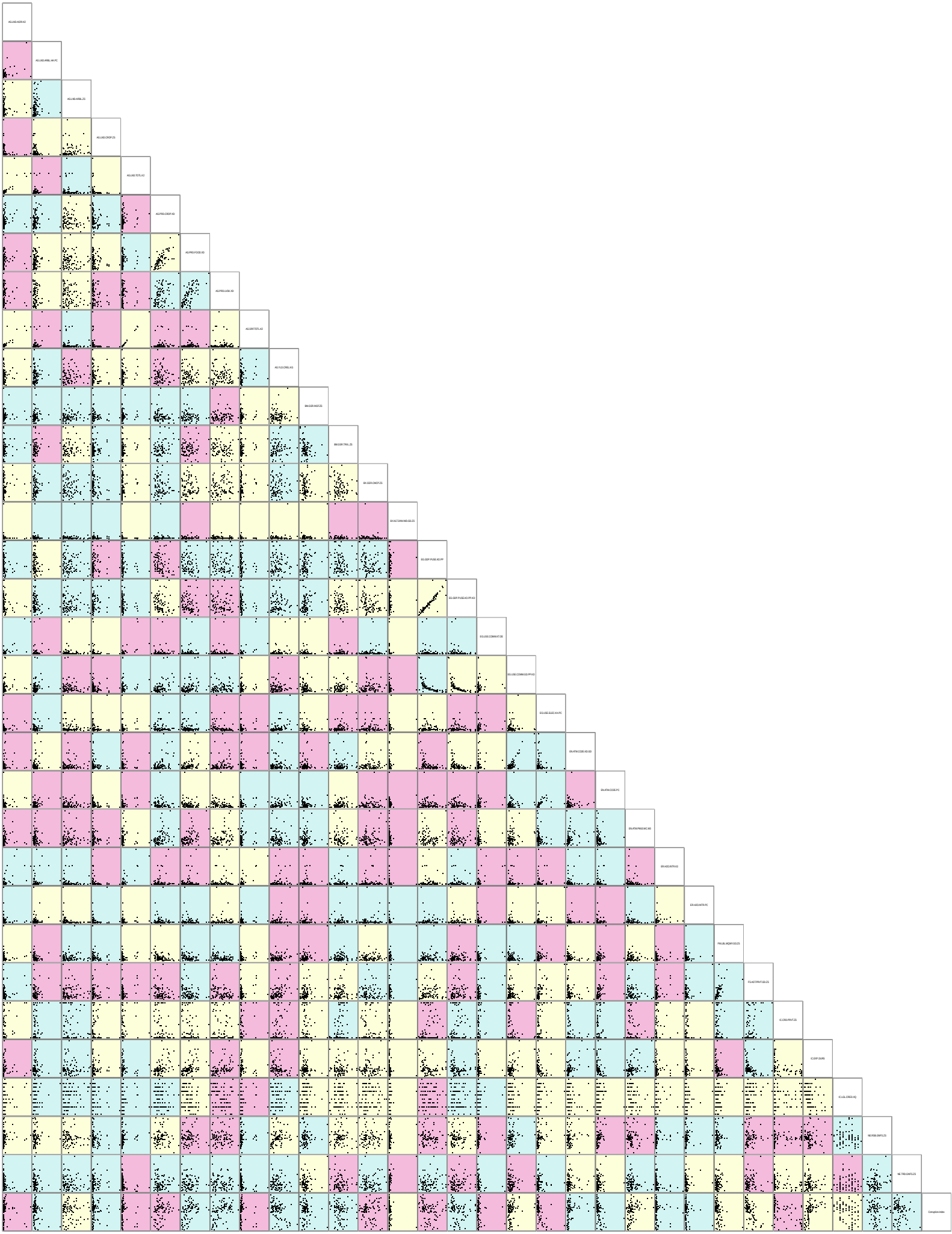


Figure 2: Scatterplots between variables

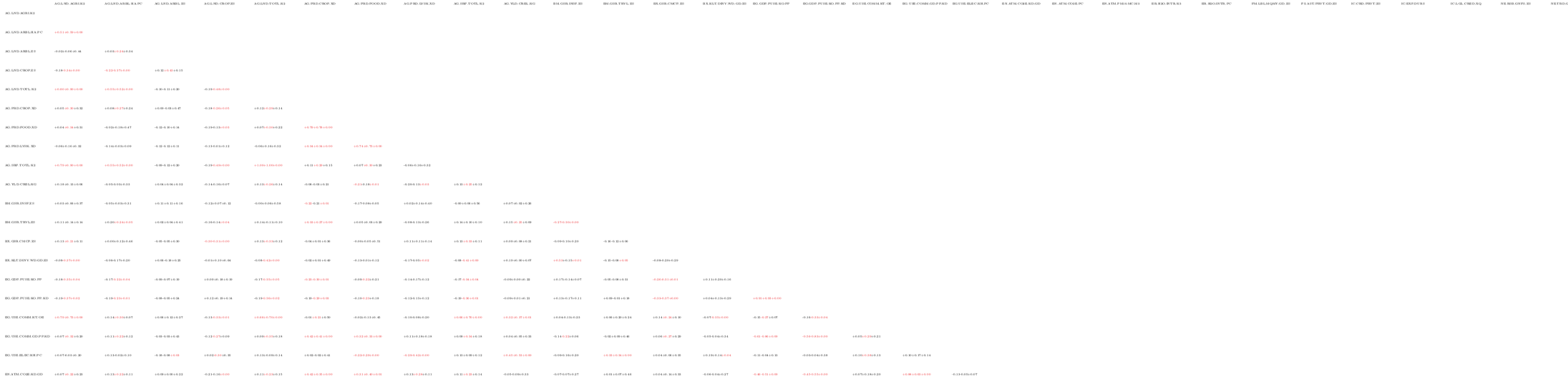
Fitting all predictors as the linear model:

```
> model.full<-lm(Corruption.Index~.,kaggle.data[variable.names])
> model.full<-lm(Corruption.Index~.,kaggle.data[variable.names],weights = 1/model.full$residuals^2)
> (model.full.sum<-summary(model.full))

Call:
lm(formula = Corruption.Index ~ ., data = kaggle.data[variable.names],
    weights = 1/model.full$residuals^2)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-1.30437 -0.75235 -0.01084  0.74554  1.57653

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.388e+02  1.075e+01  12.916  < 2e-16 ***
AG.LND.AGRI.K2        1.138e-05  2.150e-06   5.292 2.17e-06 ***
AG.LND.ARBL.HA.PC    -2.904e+01  3.908e+00  -7.431 7.42e-10 ***
AG.LND.ARBL.ZS       -1.879e-01  4.589e-02  -4.095 0.000140 ***
AG.LND.CROP.ZS       -4.482e-01  9.753e-02  -4.595 2.57e-05 ***
AG.LND.TOTL.K2       -5.956e-05  1.167e-05  -5.105 4.26e-06 ***
AG.PRD.CROP.XD       -2.364e-01  9.045e-02  -2.614 0.011535 *
AG.PRD.FOOD.XD       -7.791e-02  1.362e-01  -0.572 0.569613
AG.PRD.LVSK.XD        6.427e-02  8.478e-02   0.758 0.451628
AG.SRF.TOTL.K2        5.779e-05  1.103e-05   5.240 2.62e-06 ***
AG.YLD.CREL.KG       -2.479e-03  4.542e-04  -5.459 1.18e-06 ***
BM.GSR.INSF.ZS       -5.537e-01  9.371e-02  -5.909 2.26e-07 ***
BM.GSR.TRVL.ZS        3.669e-01  6.039e-02   6.075 1.22e-07 ***
BX.GSR.CMCP.ZS       -4.041e-02  3.640e-02  -1.110 0.271746
BX.KLT.DINV.WD.GD.ZS  2.615e-01  2.671e-02   9.791 1.18e-13 ***
EG.GDP.PUSE.KO.PP    -1.951e+00  5.955e-01  -3.276 0.001825 **
EG.GDP.PUSE.KO.PP.KD  1.458e+00  6.144e-01   2.374 0.021129 *
EG.USE.COMM.KT.OE    -2.250e-05  5.017e-06  -4.485 3.76e-05 ***
EG.USE.COMM.GD.PP.KD -1.154e-02  1.121e-02  -1.030 0.307525
EG.USE.ELEC.KH.PC    -1.489e-03  2.312e-04  -6.441 3.09e-08 ***
EN.ATM.CO2E.KD.GD     1.763e+00  8.857e-01   1.990 0.051530 .
EN.ATM.CO2E.PC       -6.030e-01  1.542e-01  -3.909 0.000256 ***
EN.ATM.PM10.MC.M3     9.950e-02  1.913e-02   5.202 3.01e-06 ***
ER.H2O.INTR.K3       -1.852e-03  1.314e-03  -1.410 0.164117
ER.H2O.INTR.PC        4.377e-05  1.651e-05   2.651 0.010455 *
FM.LBL.MQMY.GD.ZS    -1.512e-01  2.327e-02  -6.497 2.51e-08 ***
FS.AST.PRVT.GD.ZS     1.064e-01  2.860e-02   3.722 0.000467 ***
IC.CRD.PRVT.ZS       -1.272e-01  3.261e-02  -3.901 0.000263 ***
IC.EXP.DURS           1.642e-01  6.724e-02   2.442 0.017859 *
IC.LGL.CRED.XQ       -8.525e-01  3.062e-01  -2.784 0.007339 **
NE.RSB.GNFS.ZS       -3.508e-02  6.488e-02  -0.541 0.590909
NE.TRD.GNFS.ZS       -1.103e-01  1.426e-02  -7.737 2.34e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 55 degrees of freedom
Multiple R-squared:  0.9974,        Adjusted R-squared:  0.996
F-statistic: 684.9 on 31 and 55 DF,  p-value: < 2.2e-16
```

We can see that the following coeffcients are statistically significant (5% critical value), they are significant taking into consideration all other predictors (significant when all other predictors inclueded in the model):

- Intercept
- AG.LND.ARBL.HA.PC
- AG.YLD.CREL.KG
- BM.GSR.INSF.ZS
- BM.GSR.TRVL.ZS
- BX.KLT.DINV.WD.GD.ZS

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.



- EG.USE.ELEC.KH.PC
- EN.ATM.CO2E.PC
- FM.LBL.MQMY.GD.ZS
- IC.CRD.PRVT.ZS
- NE.TRD.GNFS.ZS

Also the entire model (H_0 all cooefcients are 0 against some of them are not zero) is significant.

Veryfying model assumptions

Assumptions of multiple linear regression model:

- quantitative response variable (Corruption.Index)
- p-1 explanatory independent quantitative variables
- values of the predictors are deterministic Corruption.Index are values of random variables satyfing linear equations
- ϵ_i mutually independent random variables with mean 0 and variance σ^2. For testing purposes we assume that $\epsilon_i \sim N(0, \sigma^2)$ (mean 0 and constant variance). The distribution assumption are needed for testing purposes.

  Checking normality assumption for errors:

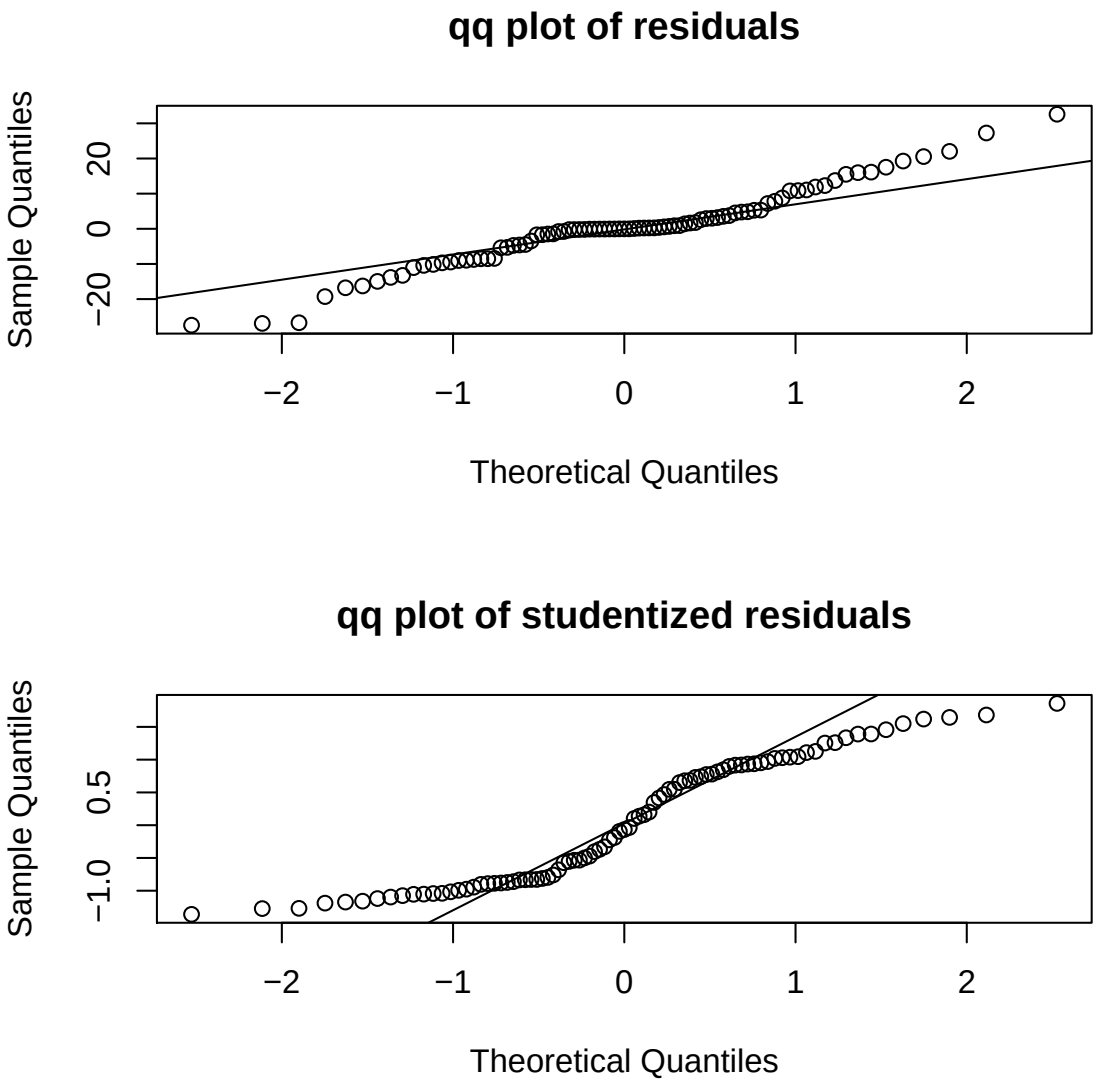


Figure 3: QQ plot

QQ plots depicts quite good linear trend so we can assume residuals have normal distribution.

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

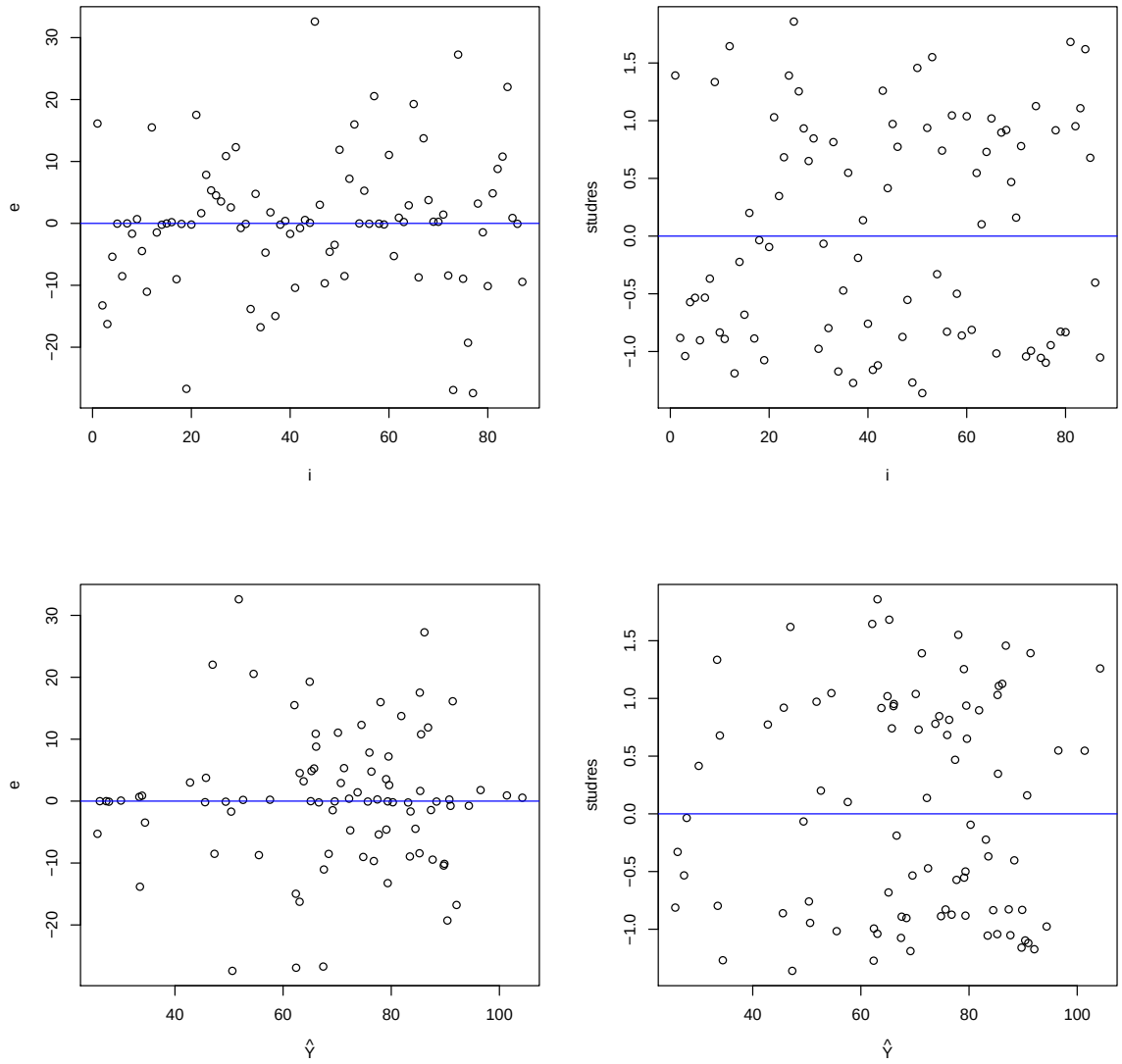


Figure 4: Residual plots

Residual plots show that errors have aproximately constant variance i.e. there is no visible pattern.

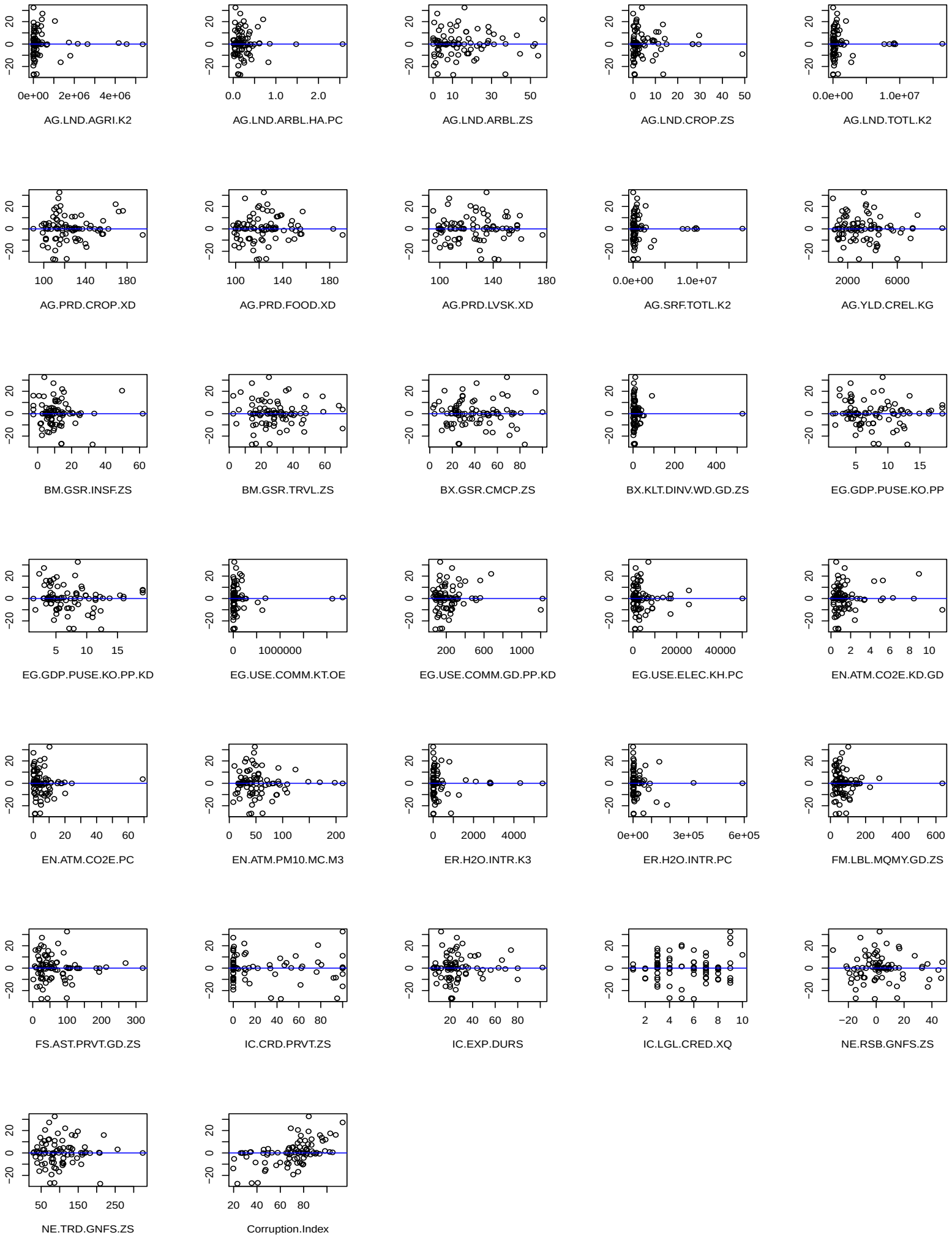


Figure 5: Residual plots for predictors

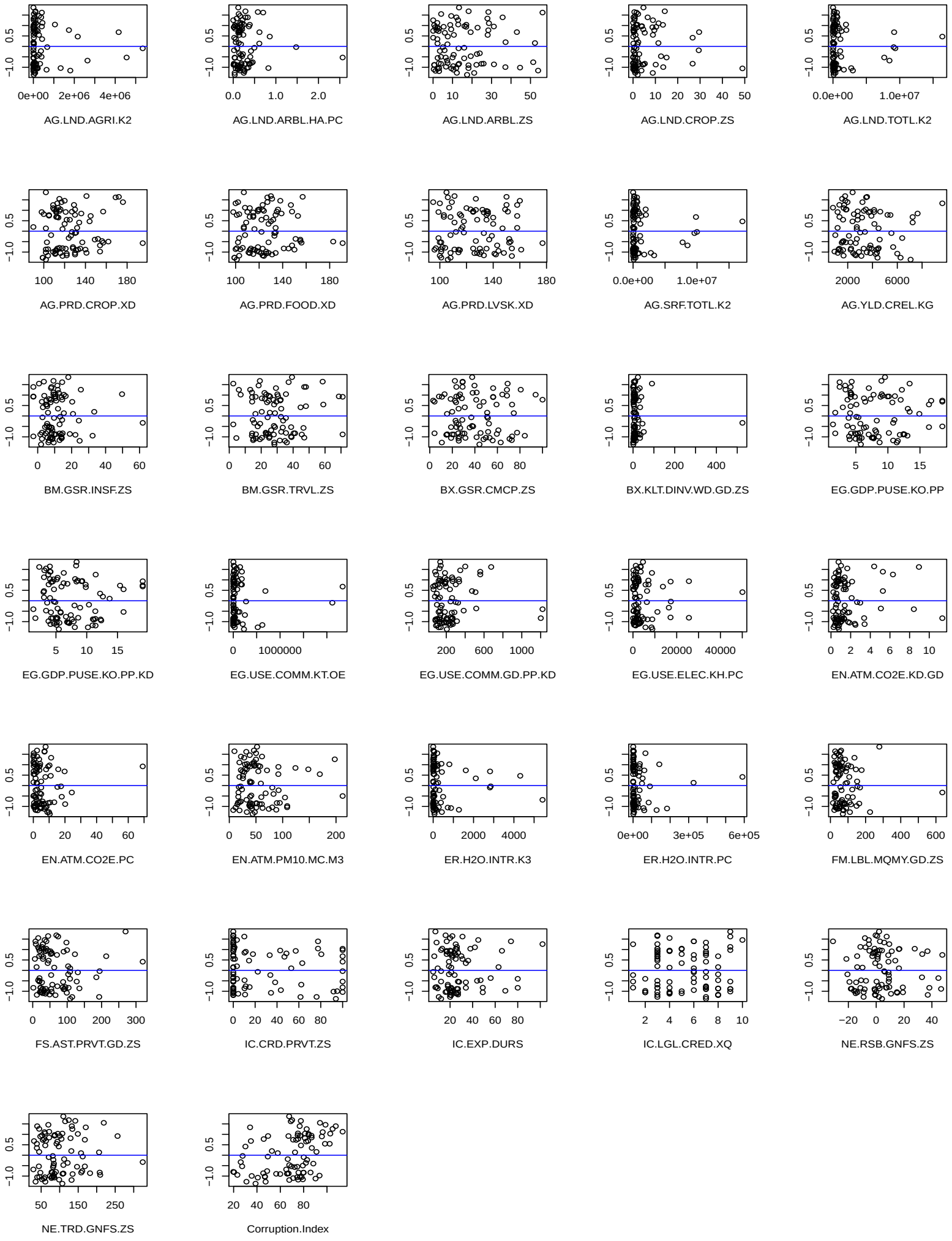


Figure 6: Studentized residual plots for predictors

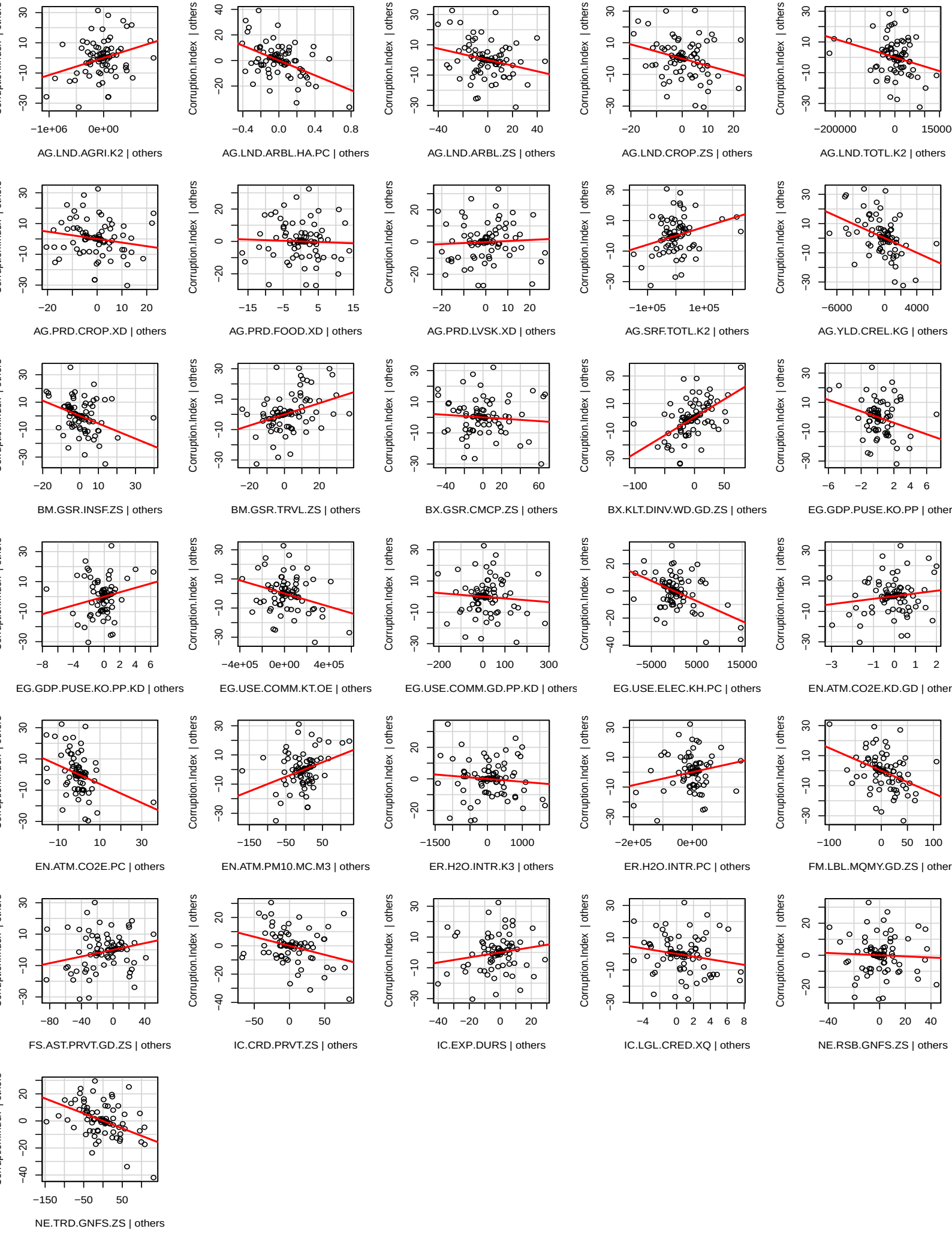


Figure 7: Partial regression plots

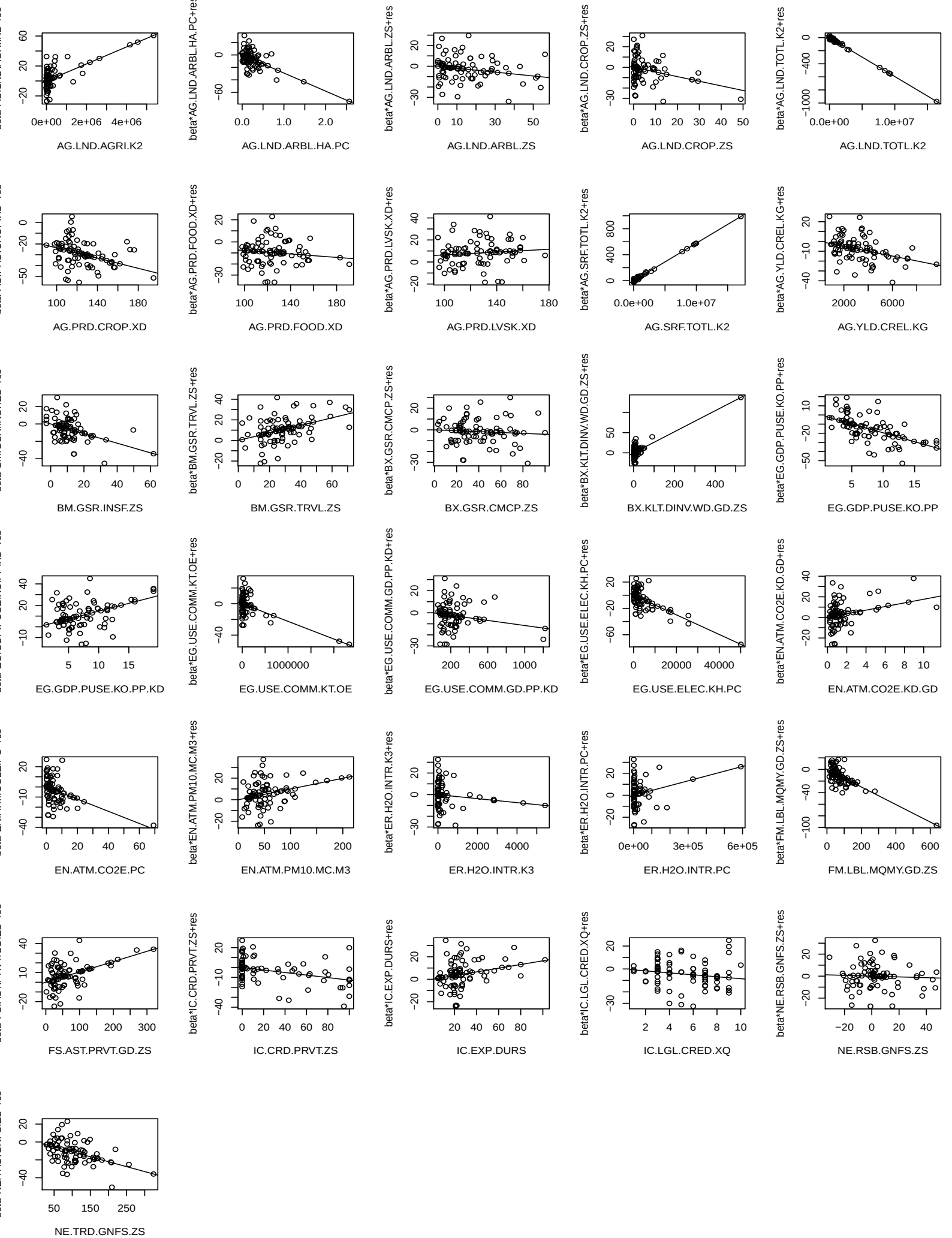


Figure 8: Partial residual plots

- additive impact of each predictor variable on exmplained variable.
- columns of experiment matrix are alegbraically independent (predictors cannot be collinear) ■= X<-cbind(replicate(nrow(kaggl
  ncol(X) as.integer(rankMatrix(X))
- $p \leq n$

  ```
  > length(all.predictor.names)+1
  ```

  ```
  [1] 32
  ```

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

  ```
  [1] 87
  ```

- is specification of the structural equation of the model correct?
- dependence of errors
- occurance of outliers or influential observations
- all significant predictors are included