World development indicators and Fragile States Index

Semester Project on course "Statistical methods in economics"
Teacher: PhDr. Mgr. Sherzod Tashpulatov

Galina Alperovich, shchegal@fel.cvut.cz

June, 2017

Contents

1 Introduction

In the current project, we consider macroeconomics World development indicators, which presents current and accurate global development data, and includes national, regional and global estimates [4]. The World Development Indicators (WDI) might be categorized under different topics such as Economy Growth, Social Development, Environment and so one. We will choose 17 the most important indicators from each category, perform their analysis and use them as explanatory variables for the Fragile States Index (FSI).

FSI is another global indicator which was developed by Fund for Peace organization, and aim to reflect the state vulnerability to conflict. This index is based on comprehensive social science approach named Conflict Assessment System Tool (CAST). This approach includes analysis of publicly available data, news, and experts opinion [2].

All visualization and calculation will be made in statistical program Gretl [3].

2 Data description

In this section, we list analyzed variables, their description and perform visualization. All listed indicators are taken for the 2010 year for the list of 177 countries with available information on indicators.

The list of variables is the following:

- *Fragile States Index* reflects the state vulnerability to conflict or collapse. We will use FSI as a dependent variable in our analysis. FSI equals to the sum of its 12 components. See the list of components on the table 2. We will not use them in the analysis, but it is important to understand from what factors FSI consists.
- We selected 20 the most valuable WDIs from every category based on Investopedia article [1]. We use them as explanatory variables.

See the detailed list of variables in the table 1.

By definition, the higher FSI for a country, the more vulnerability to conflict it has. To understand the shape of FSI distribution, let's draw the histogram and boxplot graphs. On the figure 1a we see that the FSI distribution is asymmetric and has negative skewness. We can see the same pattern on the boxplot 1b. The mass of the distribution is concentrated on the right of the figure, what means the majority of the countries has relatively high FSI. On the left side of the distribution, we see the small peak around 30, developed countries such as the USA, Canada, Austria, France, Germany, etc. Northern countries such as Norway, Finland, Sweden, Ireland have the lowest FSI. African countries such as Somalia, Sudan, Zimbabwe have the highest FSI.

On the table 3 you can find the summary statistics for FSI. The coefficient of skewness is -0.56726. Originally for several indicators, we had missing values, so we performed missing values imputations with the corresponding mean of the variable. So now the number of missing values is zero.

Also, let's look at the relationship between some explanatory variable and the target variable FSI. After we had calculated correlations between FSI and all other variables, we found out that The number of Internet users indicator has the lowest negative correlation, what means, the more Internet users the country has, the lower FSI index. It's quite interesting observations, let's draw the scatter plot between these two variables, see figure 2. On the picture, we clearly observe negative linear dependence between these two variables. It is important to mention, that correlation does not mean causality, and thus we can state only that these two variables are

Explanatory variables: World development indicators
GDP growth
Employment rate
Access to the water
Real interest rate
Energy use
Death rate
Internet users
Net flows on external debt
Total tax rate (% of commercial profits)
Number of infants death
Access to electricity (% of population)
Imports of goods and services (% of GDP)
Urban population (% of total)
School enrollment, primary (% gross)
Improved sanitation facilities (% of population with access)
Taxes on income, profits and capital gains
Number of scientific articles
Dependent variable: Fragile States Index

Table 1: List of variables considered in the project. The data is taken for 177 countries in 2010.

changing linearly at the same time, but we don't know what variable influences on this change. The correlation matrix visualization can be found on the figure 3.

Security Apparatus	External Intervention
Factionalized Elites	Refugees and IDPs
Group Grievance	Demographic Pressures
Economic Decline	Human Rights and Rule of Law
Uneven Economic Development	Public Services
Human Flight and Brain Drain	State Legitimacy

Table 2: The list of 12 components of Fragile State Index which is calculated as their sum.

Mean	71.874
Median	77.100
Minimum	18.700
Maximum	114.30
Standard deviation	23.151
C.V.	0.32210
Skewness	-0.56726
Ex. kurtosis	-0.46150
5% percentile	27.290
95% percentile	105.14
Interquartile range	30.400
Missing obs.	0

Table 3: Summary statistics for Fragile States Index (177 observations)

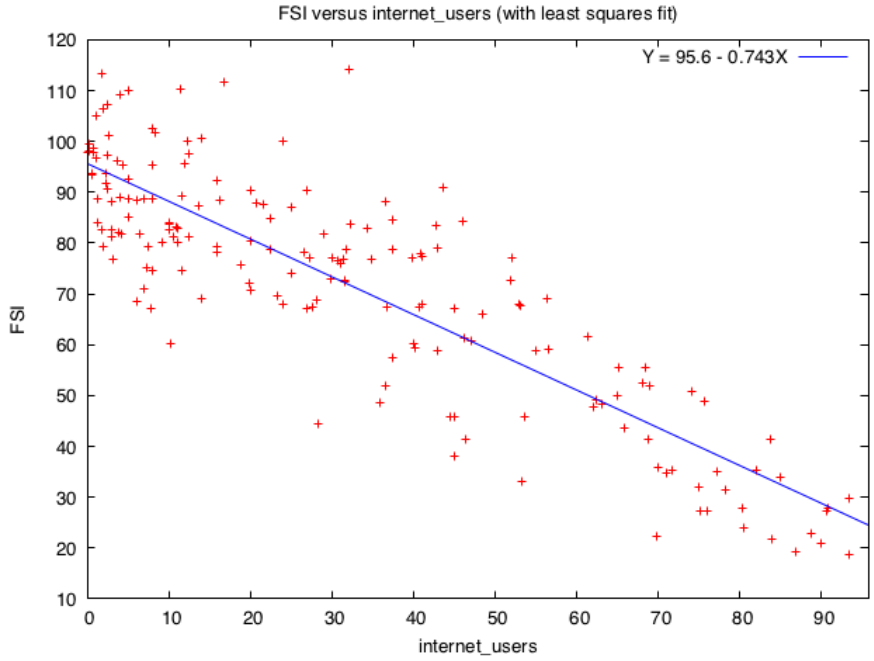
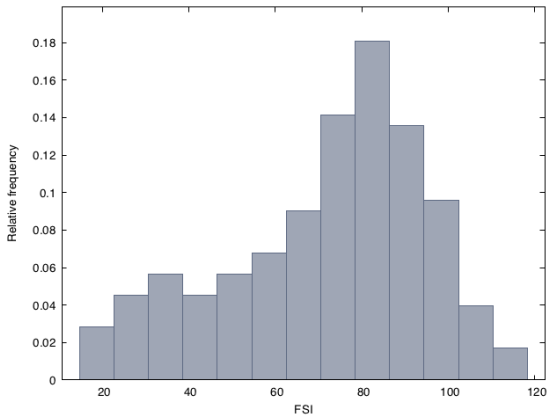
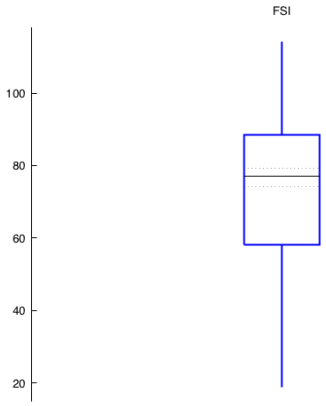


Figure 2: Linear relationship between the number of Internet users and FSI. We see negative linear dependence.



(a) Histogram of the FSI. The higher FSI, the more state vulnerability to conflict.



(b) Boxplot of the FSI

Figure 1: Fragile States Index distribution

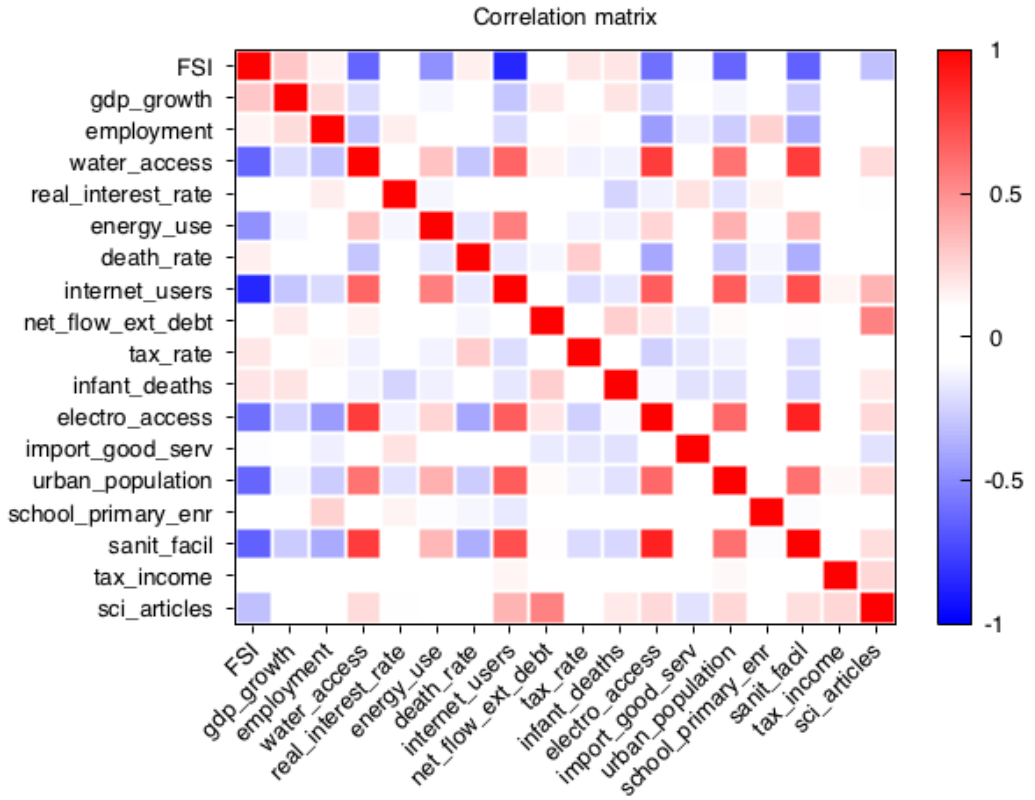


Figure 3: Visualization of the correlation matrix between all variables. 5% critical value (two-tailed test) = 0.1476

3 Economic relationship

On the figure 3 we see the visualization of the correlation matrix. Red color means positive correlation; blue color means negative correlation. FSI is our target variable, so let's firstly analyze

the apparent relationship between it and other variables. As we already mentioned, it has the lowest negative correlation with the number of Internet users. Also, FSI has such relationship with the following indicators: Water access, Energy use, Electricity access, Urban population and Sanitation facilities. This negative linear relationship reflects the main problems of the least developed countries: few people have access to the pure water, electricity and surely Internet. People mostly live not in the cities, and with the poor sanitation conditions. The FSI index can be considered also as a composite measure of country development. Also, we can see the high positive correlation between pairs of listed indicators.

Let's calculate the Variance Inflation Factor (VIF) for every explanatory variable. VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity. The values greater than 10.0 may indicate a collinearity problem. With Gretl, we calculated all VIFs and found out that all values are less than 7, what means we don't have multicollinearity problem and don't need to exclude variables. See the table 4.

gdp_growth	1.342
employment	1.650
water_access	3.309
real_interest_rate	1.321
energy_use	1.737
death_rate	1.485
internet_users	4.431
net_flow_ext_debt	1.752
tax_rate	1.227
infant_deaths	1.403
electro_access	6.904
import_good_serv	1.298
urban_population	2.364
school_primary_enr	1.235
sanit_facil	6.364
tax_income	1.115
sci_articles	2.038

Table 4: Variance Inflation Factors for all explanatory variables. All values are less than 10 what means we don't have multicolleniarity problem.

4 Methodology

In this section, we will build the models for FSI based on selected WDIs.

4.1 OLS

Let's build the first Ordinary Least Squares model, and include all variables. See the table 5. The quality of the model can be judged by R^2 coefficient, in our case it equals to 0.79, what is the relatively good but not he best quality. The p-value for each term in the table tests the null hypothesis that the regression coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In such way, we have four significant variables: Employment rate, Water access, Internet users and Urban population.

The standard assumption in linear regression is that the theoretical residuals are independent and normally distributed. Observed residuals are an estimate of the theoretical residuals, so let's test them on normality with QQ plot, see the figure 4. We see it is close to normal distribution

(diagonal line), hence normality assumption holds. Also normality test with gives us the p-value = 0.740798, what means we can't reject null hypothesis about normality.

$$H_0 : \text{errors are normally distributed}$$

$$H_1 : \text{errors are not normally distributed}$$

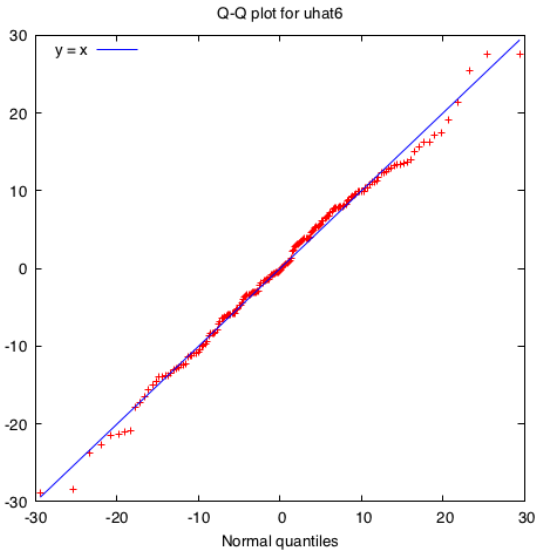


Figure 4: QQ plot for residuals of the OLS model. We see it is close to normal distribution, hence normality assumption holds.

	Coefficient	Std. Error	t-ratio	p-value
const	147.089	12.4407	11.82	0.0000 ***
gdp_growth	0.294608	0.266989	1.103	0.2715
employment	−0.202418	0.0923576	−2.192	0.0299 **
water_access	−0.222474	0.0968161	−2.298	0.0229 **
real_interest_rate	−0.0680424	0.135074	−0.5037	0.6151
energy_use	−4.10924e–05	0.000427028	−0.09623	0.9235
death_rate	−0.258066	0.303401	−0.8506	0.3963
internet_users	−0.592710	0.0658440	−9.002	0.0000 ***
net_flow_ext_debt	1.44676e–10	8.88111e–11	1.629	0.1053
tax_rate	0.00134957	0.0265030	0.05092	0.9595
infant_deaths	−4.19536e–06	9.43917e–06	−0.4445	0.6573
electro_access	0.0289993	0.0694400	0.4176	0.6768
import_good_serv	−0.0446160	0.0424610	−1.051	0.2950
urban_population	−0.127797	0.0569667	−2.243	0.0263 **
school_primary_enr	−0.138846	0.0862127	−1.611	0.1093
sanit_facil	−0.0347277	0.0712599	−0.4873	0.6267
tax_income	1.56924e–13	1.47843e–13	1.061	0.2901
sci_articles	−0.000163556	0.000126628	−1.292	0.1984

Mean dependent var	71.87401	S.D. dependent var	23.15073
Sum squared resid	19901.63	S.E. of regression	11.18783
R^2	0.789017	Adjusted R^2	0.766460
$F(17, 159)$	34.97747	P-value(F)	6.18e–45
Log-likelihood	−669.0852	Akaike criterion	1374.170
Schwarz criterion	1431.341	Hannan–Quinn	1397.357

Table 5: First OLS model for FSI, all linear terms are included. We can see a lot of insignificant variables.

4.2 RESET test

Now let's run the Ramsey's RESET test for the first model specification. RESET test considers the following hypothesis:

$$H_0 : \text{model is correctly specified}$$

$$H_1 : \text{model is incorrectly specified}$$

For this test, we have p-value = 0.00133 what means we should reject the null hypothesis about the correctness of the model. It shows us that some components of the model might be included not in the linear form but quadratic or even cubic form. Let's build Auxiliary regression for RESET specification test to find quadratic components. In this regression, coefficient next to the quadratic terms is tested for equality to zero. See results on the table 6. After this test, we will select the most significant quadratic variables to add them to the next model. The list of quadratic terms includes GDP growth, Employment rate, Water access, Death rate, Internet users, Net flows on external debt, Imported goods and services, Urban population and School enrollment. The list of the most significant linear terms from the original OLS model includes Employment rate, Water access, Internet users, Urban population. We run the RESET test for cubic terms as well and found out that the Water access and Internet users are also can be added to the model in cubic form.

	coefficient	std. error	t-ratio	p-value
const	315.013	50.9679	6.181	5.22e-09 ***
gdp_growth	0.747051	0.290985	2.567	0.0112 **
employment	0.449082	0.115297	3.895	0.0001 ***
water_access	0.621257	0.150413	4.130	5.85e-05 ***
real_interest_ra	0.167452	0.134069	1.249	0.2135
energy_use	0.000222429	0.000420831	0.5285	0.5979
death_rate	0.671893	0.318191	2.112	0.0363 **
internet_users	1.28433	0.213706	6.010	1.24e-08 ***
net_flow_ext_debt	2.51920e-10	9.16473e-11	2.749	0.0067 ***
tax_rate	0.00842815	0.0257539	0.3273	0.7439
infant_deaths	6.88892e-06	9.17664e-06	0.7507	0.4539
electro_access	0.0412992	0.0673531	0.6132	0.5406
import_good_serv	0.0824425	0.0426113	1.935	0.0548 *
urban_population	0.295141	0.0740259	3.987	0.0001 ***
school_primary_e	0.390419	0.111700	3.495	0.0006 ***
sanit_facil	0.111289	0.0726174	1.533	0.1274
tax_income	2.84887e-13	1.48081e-13	1.924	0.0562 *
sci_articles	0.000217790	0.000123682	1.761	0.0802 *

Table 6: Auxiliary regression for RESET specification test, quadratic terms.

4.3 Final OLS model

Let's build a new OLS model with the linear, quadratic and cubic terms. See the results for the model on the table 7. As we can see we included the following variables: squared School enrollment, squared Internet users, cubic Water access and Urban population. We reduced the number of variables from 17 to 4 without loss of quality (we got R^2 even a bit higher). After collinearity check, we could observe that all VIF values are less than 3, what means we don't have collinearity problem.

	Coefficient	Std. Error	t-ratio	p-value
const	111.153	4.77485	23.28	0.0000 ***
school_primary_enr_2	−0.000747504	0.000347218	−2.153	0.0327 **
internet_users_2	−0.00665839	0.000475151	−14.01	0.0000 ***
water_access_3	−2.05824e–05	3.77392e–06	−5.454	0.0000 ***
urban_population	−0.0852354	0.0482812	−1.765	0.0793 *

Mean dependent var	71.87401	S.D. dependent var	23.15073
Sum squared resid	19178.76	S.E. of regression	10.55956
R^2	0.796681	Adjusted R^2	0.791953
$F(4, 172)$	168.4902	P-value(F)	2.22e–58
Log-likelihood	−665.8108	Akaike criterion	1341.622
Schwarz criterion	1357.502	Hannan–Quinn	1348.062

Table 7: Final OLS model for FSI. Linear, quadratic and cubic terms are included. Number of variables is reduced from 17 to 4.

4.4 Heteroskedasticity test

Let's run White's test for heteroskedasticity in order to check whether the variance of the errors in a regression model is constant. White's test considers the following hypothesis:

$$H_0 : \text{homoskedasticity is presented}$$

$$H_1 : \text{heteroskedasticity is presented}$$

We got the p-value = 0.223339, what means we don't reject null hypothesis and we deal with homoskedasticity case.

5 Estimation

The final regression model is specified on the table 7. The equation of the model is the following:

$$\begin{aligned}\widehat{\text{FSI}} = &\underset{(23.279)}{111.153} - \underset{(-2.153)}{0.000747504}\,\text{school\_primary\_enr\_2} - \underset{(-14.013)}{0.00665839}\,\text{internet\_users\_2} \\ &- \underset{(-5.454)}{2.05824\text{e}{-05}}\,\text{water\_access\_3} - \underset{(-1.765)}{0.0852354}\,\text{urban\_population} \\ &T = 177 \quad \bar{R}^2 = 0.7920 \quad F(4, 172) = 168.49 \quad \hat{\sigma} = 10.560 \\ &(t\text{-statistics in parentheses})\end{aligned}$$

Let's run the Ramsey's RESET test for the final specification. RESET test considers the following hypothesis:

$$H_0 : \text{model is correctly specified}$$

$$H_1 : \text{model is incorrectly specified}$$

We see the following result:

RESET test for specification (squares and cubes)
Test statistic: F = 1.988712,
with p-value = P(F(2,167) > 1.98871) = 0.14

RESET test for specification (squares only)
Test statistic: F = 0.257972,
with p-value = P(F(1,168) > 0.257972) = 0.612

RESET test for specification (cubes only)
Test statistic: F = 0.009258,
with p-value = P(F(1,168) > 0.00925784) = 0.923

We observe that all p-values are greater than 0.05, that means we don't reject the hypothesis about the correct specifications of the model. Hence the specification is correct.

Let's compare real target values and fitted values from the model. Firstly, look at the scatter plot composed of FSI and $\widehat{FSI}$. On the figure 5 we can see that predictions are not perfect (all points are not on the same line) but it is close to the line.

Also we can calculate the following quality measures between FSI and $\widehat{FSI}$:

Mean Error: -1.6619e-14
Root Mean Squared Error: 10.31

Mean Absolute Error: 8.1059
Mean Percentage Error: -2.793
Mean Absolute Percentage Error: 13.424

As for a coefficient of determination R^2, we see the following result:

R^2 : 0.797
Adjusted R^2 : 0.792
p-value: 2.22e-58 refers to the testing of R^2 on equality to zero, it will be discussed later in the section 7.

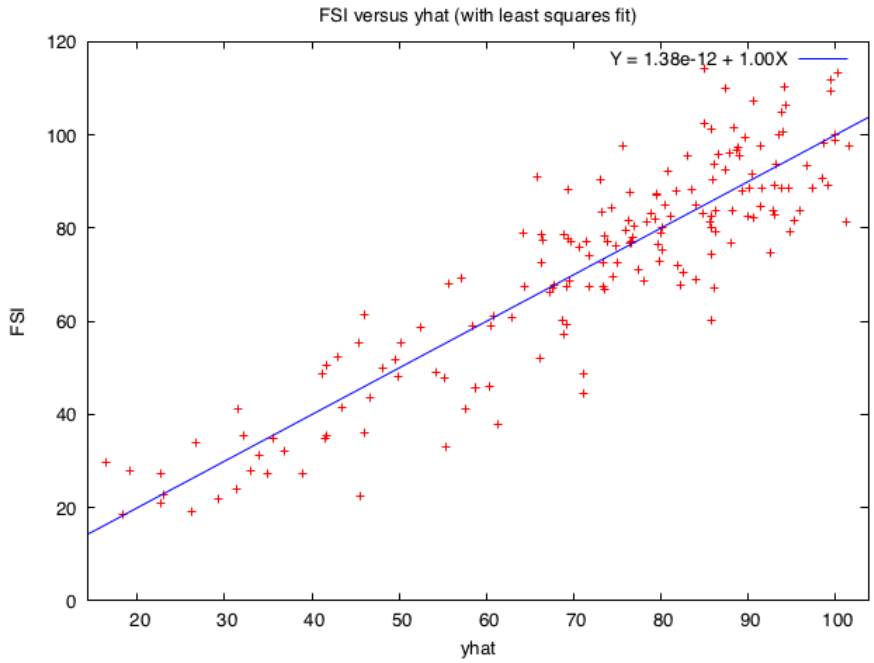


Figure 5: Scatter plot between FSI and $\widehat{FSI}$

6 Statistical inference

To test the significance of every variable in the model we consider the following hypothesis for every coefficient corresponding to the explanatory variable:

$$H_0 : \beta_j = 0, \text{for all } j$$

$$H_1 : \beta_j > 0$$

All variables have p_value < 0.08 and hence significant result (reject null hypothesis).

7 Goodness of fit test

To test the significance of the overall model, we test the coefficient of determination R^2 which shows how much our model explains the changes in the dependent variable, by using independent variables [5]. The test uses F-statistic and considers the following hypothesis:

$$H_0 : R^2 = 0$$
$$H_1 : R^2 > 0$$

In our case p-value = 2.22e-58 what means F-statistic is in the tail and our decision is rejecting H_0, hence R^2 is significant.

8 Economic verification

All four included variables have a negative sign, what means we have negative linear relationship between these variables and the target FSI variable. We discussed this in the first section, where we looked at the correlation matrix. In that section, we also observed negative correlation coefficient.

In our case, the higher proportion of the children enrolled in the primary school, the lower FSI index (we remember, that Finland, Norway, and other north countries have the lowest FSI). The higher number of Internet users, Number of people with water access, Urban population, the lower FSI.

9 Interpretation

The R^2 coefficient of determination can be interpreted as follows: our four explanatory variables (plus the constant) explain 80% of the variance of our dependent (target) variable FSI.

Interpretation of the coefficient can be as follows:

$$\frac{d(FSI)}{d(school\_primary\_enr)} = -2{\cdot}0.000747{\cdot}mean(school\_primary\_enr) = -0,001495{\cdot}105.25 = -0,157$$

That means that if in average the proportion of children enrolled to the primary school increases by 1, the FSI index will decrease by 0.15. The same is actual for the other three variables:

$$\frac{d(FSI)}{d(internet\_users)} = -2 \cdot 0.00665839 \cdot mean(internet\_users) = -0.00665839 \cdot 32 = -0,21$$

$$\frac{d(FSI)}{d(water\_access)} = -3 \cdot 2.05824e{-}05 \cdot mean(water\_access)^2 = -2.05824e{-}05 \cdot 86.476 = -0,17$$

$$\frac{d(FSI)}{d(urban\_population)} = -0.0852354$$

Hence, if in the average proportion of the Internet users increases by 1, the FSI index will decrease by 0.21. If in the average proportion of the people who have access to the water increases by 1, the FSI index will decrease by 0.17. If the share of Urban population increases by 1, the FSI index will decrease by 0.08.

10 Conclusion

In this project, we analyzed various global World development indicators for 177 countries for the 2010 year. We considered 17 the most important indicators from different categories and analyzed how they influence on the Fragile State Index. We discovered that there are four the most significant indicators which explain 80% of FSI variance: Proportion of the children enrolled in

the primary school, Number of Internet users, Proportion of the people with access to the water and Urban population. The interesting fact is that the Number of Internet users and FSI has the lowest negative correlation among other variables, and hence the negative linear relationship. The simplest linear regression model with only this one variable explains 74% of FSI variance, and thus this indicator is the most significant for FSI prediction.

We performed missing values imputation, correlation analysis, we built a linear regression model for FSI based on various WDIs, checked collinearity and heteroskedasticity and also run RESET test to find quadratic and cubic components for the model.

FSI is calculated by the independent company Fund for Peace and requires a lot of expertise and comprehensive analysis. In this project, we showed that it is possible to built statistical model for FSI based on only four publicly available WDI variables which explain 80% of FSI changes. In further works, it would be a good direction to consider other 970 different WDIs and build a more complicated model to improve the prediction of FSI.

References

[1] The world bank's all-important world development indicators (wdi), 2014. `http://www.investopedia.com/articles/investing/100614/world-banks-allimportant-world-development-indicators-wdi.asp`.

[2] Fragile state index methodology, 2017. `http://fundforpeace.org/fsi/methodology/`.

[3] Gretl: Gnu regression, econometrics and time-series library, 2017. `http://gretl.sourceforge.net/`.

[4] World bank open data, 2017. `http://data.worldbank.org/`.

[5] S. Tashpulatov. Lecture notes on statistical methods in economics, 2017. *Czech Technical University.*