ECONOMETRICS

Assignment 1

Niccolò Chiaverini - 3232465
Filippo Grandoni - 3187864
Ludovico Panariello - 3192212
Lorenzo Ravera - 3200788

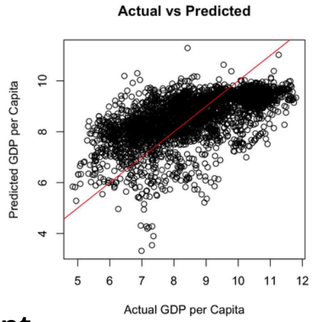
Data Preparation

At the beginning of our analysis there is a rigorous data processing and preparation segment to clean and prepare the dataset for analysis. Initially, it retrieves the relevant economic indicators for the Solow-Swan model (GDP per capita - “NY.GDP.PCAP.CD”, savings rate - “NY.GNS.ICTR.ZS”, and population growth rate - “SP.POP.GROW”) by using the WDI function from the WDI package, which gives access to the WDI database for all countries. To ensure data integrity, complete cases—observations without missing values—are retained for further analysis. The datasets are then merged based on the years and the ISO 3 country code. Finally, a log transformation is applied to the variables, with an offset added to handle zero values, thereby preparing the dataset for the linear regression.

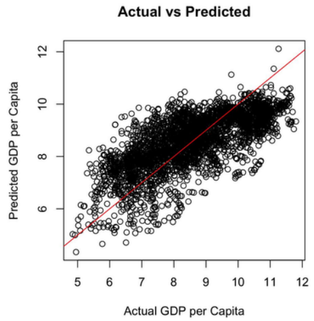
Cleaning Datasets

To perform the regression we decided to focus solely on individual country-level observations and not on aggregates. Avoiding medians helped us to ensured a linear regression model based on granular and diverse observations, leading to more accurate and reliable results. We further cleaned the data by **excluding observations** that are found to be **less influential** in regression analysis. To do so, we compute the **Cook's distance**, which helped in quantifying the effect of removing each observation on the regression coefficients, allowing us to identify these potential outliers and thus avoid potential bias in our model and ensure that our variance is not inflated due to this data.

Regression with outliers



Regression without outliers

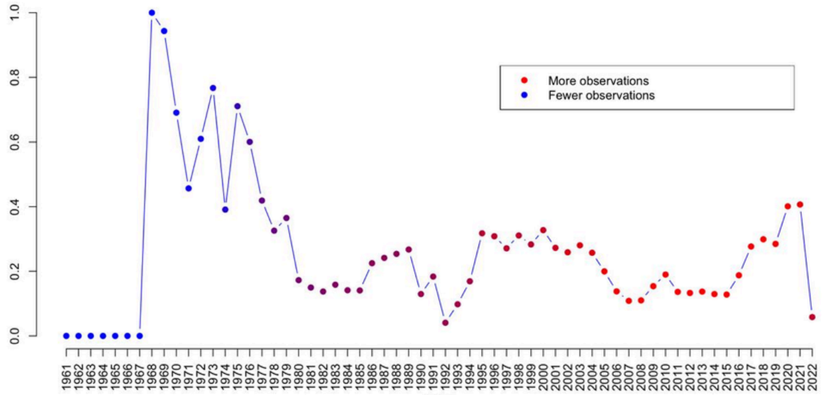


Interpretation of the Results

As it can be seen from the graphs and from the computation of the R-squared coefficient (0.12 versus 0.3), there is a **significant difference in performing a regression on the model with and without outliers**. The data fit better the model having lower variance and not being biased by outliers. In both cases the model suggests that both *savings_rate* and *population_growth* are statistically significant predictors of *gdp_per_capita*, with *savings_rate* having a positive impact and *population_growth* a negative one. However, an R-squared value of approximately 0.3 indicates that while there is still a significant relationship, the model doesn't capture all the variability in the data even if we remove the outliers, and there may be other factors influencing *gdp_per_capita* that are not included in the model.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        16.8428     2.4422   6.897 5.79e-12 ***
savings_rate        9.3964     0.3596  26.128  < 2e-16 ***
population_growth -17.8846     0.3820 -46.819  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.278 on 7012 degrees of freedom
Multiple R-squared:  0.3006,    Adjusted R-squared:  0.3004
F-statistic:  1507 on 2 and 7012 DF,  p-value: < 2.2e-16
```



Analysis of the R-Squared values over time

Initially, when data is limited to fewer observations per year, the model effectively captures the variation in GDP per capita based on the available explanatory variables. This leads to higher R-squared values. However, as the number of observations per year increases over time, the model's explanatory power diminishes, reflected in lower R-squared values, as it struggles to capture a strong relationship if subjected to a significantly larger number of data. We can see though that during the last 10 years there has been an improvement that suggests that savings and population growth could possibly influence more gdppc than before.

Results

```
        RESET test

data:  model
RESET = 99.516, df1 = 2, df2 = 5757, p-value < 2.2e-16
```

The high RESET statistic and the very low p-value indicate that the model may have omitted important variables or the relationship between the predictors and the outcome is not correctly specified

```
        Goldfeld-Quandt test

data:  model
GQ = 0.99868, df1 = 2878, df2 = 2878, p-value = 0.5141
alternative hypothesis: variance increases from segment 1 to 2
```

The p-value is strictly above the typical alpha level of 0.5, suggesting that there is strong evidence on H0 supporting constant variance

```
        studentized Breusch-Pagan test

data:  model
BP = 11.496, df = 2, p-value = 0.003189
```

```
        studentized Breusch-Pagan test

data:  model
BP = 208.36, df = 2, p-value < 2.2e-16
```

The first BP statistic is 11.496 with a p-value of 0.003189, indicating strong evidence of heteroscedasticity. The second BP statistic is significantly larger at 208.36 with a p-value < 2.2e-16, which further supports the presence of heteroscedasticity

```
 lag Autocorrelation D-W Statistic p-value
   1       0.9188926     0.1617793       0
 Alternative hypothesis: rho != 0
```

```
        Breusch-Godfrey test for serial correlation of order up to 5

data:  model
LM test = 4888.2, df = 5, p-value < 2.2e-16
```

This test provides a strong evidence of serial correlation in the residuals up to order 5

```
        Jarque Bera Test

data:  model$residuals
X-squared = 29.23, df = 2, p-value = 4.496e-07
```

The Jarque-Bera statistic is 29.23 with a p-value of 4.496e-07, suggesting that the residuals do not follow a normal distribution, which is an important assumption in many regression models

```
Call:
lm(formula = gdp_per_capita ~ savings_rate + population_growth +
    savings_rate_2 + population_growth_2, data = transformed_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3835 -0.8675 -0.0877  0.8649  4.1304

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          27.708    102.325   0.271 0.786563
savings_rate        -84.725     29.071  -2.914 0.003575 **
population_growth   129.305     38.452   3.363 0.000776 ***
savings_rate_2        8.481      2.622   3.235 0.001224 **
population_growth_2 -21.640      5.648  -3.831 0.000129 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.276 on 7010 degrees of freedom
Multiple R-squared:  0.303,     Adjusted R-squared:  0.3026
F-statistic: 761.9 on 4 and 7010 DF,  p-value: < 2.2e-16
```

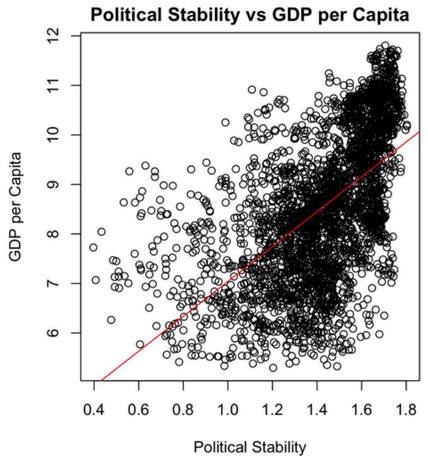
The inclusion of transformed variables, savings_rate_2 and population_growth_2, alters the coefficients of the original variables, suggesting a possible non-linear relationship with GDP per capita. Savings_rate now has a large negative coefficient, while population_growth has a large positive one, indicating a more complex relationship.

Both transformed variables are statistically significant, capturing additional aspects of the relationship not explained by the original variables alone. The adjusted R-squared of 0.3026 suggests a slightly better fit than the original model, but still relatively low.

Political Stability

We have considered several additional variables that could be influential, among which political stability is the one that convinced us the most. The **economic intuitions** for which we expected a positive correlation between political stability and GDP per capita are many:

- **Investment Confidence**: Political stability creates a predictable and secure environment for both domestic and foreign investors. Higher levels of investment in turn can lead to more robust economic growth and thus higher GDP per capita.
- **Attracting Talent and Skills**: Politically stable countries are more attractive destinations for talented individuals and skilled workers from around the world. The influx of skills and talent can enhance innovation, productivity, and economic growth, leading to higher GDP per capita.
- **Economic Policies**: Politically stable environments tend to facilitate the development and implementation of coherent and consistent economic policies. Governments in stable countries are more likely to be able to plan for the long term, implement necessary reforms, and ensure policy continuity. This consistency helps in creating a conducive environment for economic growth.
- **Risk of Conflict**: Politically unstable regions are at a higher risk of conflicts, which can have devastating effects on the economy.



When choosing the additional variable we also considered the correlation between the latter and GDP per Capita. As it can be perceived from the graph, the **relationship between our additional variable and GDP is pretty linear**. Political stability is a good proxy especially for higher GDP, meaning that when there is a high GDP the political stability index is high too, while the opposite is not necessarily true.

We hence ran a regression on **Political Stability and Absence of Violence/Terrorism** which is an index that evaluates the perceived stability of a country's political environment and the absence of violence or terrorism. It considers factors such as government effectiveness, rule of law, control of corruption, and the likelihood of political unrest or violent conflicts. To handle the data we adopted the same procedure we followed in the first regression model.

Results

Linear Model with Political Stability

In the refined econometric model, the R-squared value robustly captures **50.87%** of GDP per capita's variability, significantly improving upon the previous iteration by taking into account political stability. Economically, the model suggests a unitary increase in the savings rate correlates with a 6.98 increment in GDP per capita, reflective of the classical economic theory positing that **savings fuel investments** leading to growth. The negative coefficient for population growth aligns with the **Malthusian perspective** where unchecked growth potentially dilutes per capita wealth creation. The inclusion of **political stability**, with its substantial coefficient of 3.49, underscores the decisive role governance plays in economic development. This particular variable's elasticity indicates a considerable sway over GDP, likely due to its effect on attracting investments, securing property rights, and ensuring a stable economic trajectory, fundamental elements conducive to robust economic growth.

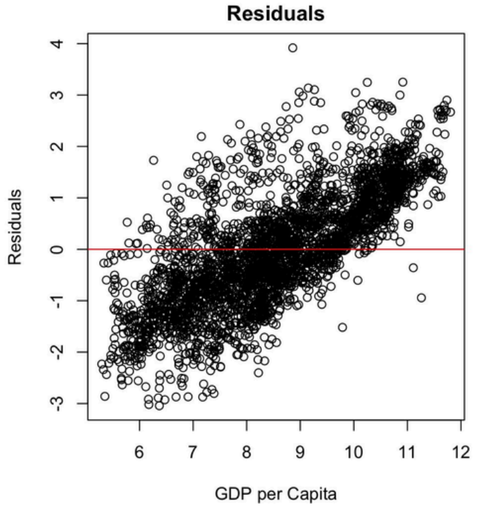
```
Call:
lm(formula = gdp_per_capita ~ savings_rate + population_growth +
    Political_stability, data = merged_data_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-3.13426 -0.72789 -0.02009  0.75390  3.07639

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.95027    2.85916  -0.682    0.495
savings_rate         6.98449    0.44245  15.786   <2e-16 ***
population_growth   -9.80825    0.48159 -20.367   <2e-16 ***
Political_stability  3.49028    0.08821  39.569   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 3301 degrees of freedom
Multiple R-squared:  0.5087,    Adjusted R-squared:  0.5083
F-statistic:  1139 on 3 and 3301 DF,  p-value: < 2.2e-16
```

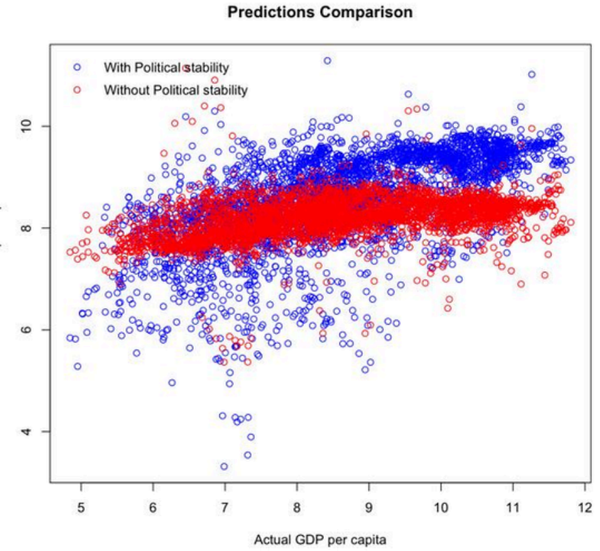
Similarly to the previous results, after removing outliers the regression model captures better the underlying patterns in the data and provides more accurate estimates of the true relationships between variables. This enhances the reliability of the regression results and improves the model's predictive power.



Analysis of Residuals

The diagnostic residual plot elucidates the presence of **heteroscedasticity**, a common variance anomaly in cross-country analyses, where the scatter of residuals broadens with an increase in GDP per capita. This breach of homoscedasticity suggests the error **variance is not uniform** across all levels of GDP and our model is flawed: lower GDP nations are consistently underpredicted while the opposite holds for higher GDP nations. This may be a reflection of **latent non-linearities** not captured by the current model or omitted variables that exert disproportionate influence at different income levels. Economically, this implies that uniform policy prescriptions across diverse economic landscapes could be ineffectual, as the model's residuals indicate varying sensitivities to the explanatory variables. This pattern warrants a transformation of variables or the adoption of a heteroscedasticity-consistent estimation technique to rectify the non-constant variance and achieve unbiased standard errors, thereby ensuring the model's coefficients are truly representative of the underlying economic phenomena.

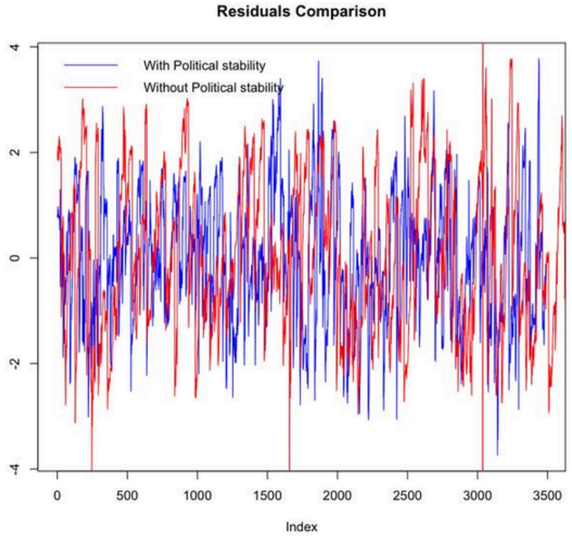
Comparison of the results




```
Summary of Differences:

Model Performance Metrics Difference:
R-squared difference: 0.2238419
 Adjusted R-squared difference: 0.223792
 F-statistic difference: 388.6773
```

The significant increase in the adjusted R-squared value upon including the political stability index parameter in the Solow model regression highlights the substantial explanatory power of political stability we expected on variations in economic growth, as captured by GDP per capita.



The similarity in the average pattern of residuals between the two regressions suggests underlying economic forces or data characteristics that transcend the inclusion of political stability. This similarity implies that the addition of the political stability index to the regression model does not substantially alter the residual patterns, indicating that other factors not captured by the political stability index may be driving the observed residuals.

Reasoning about the change in regression parameters due to Inclusion of Political Stability

The revised model's coefficients, adjusted for political stability, provide a nuanced understanding of its interplay with economic indicators. Political stability, with its notable coefficient, possibly absorbs some **predictive power** from savings rate and population growth, refining their estimated effects on GDP per capita. This could be due to political stability's potential **correlation** with these factors, suggesting that it functions as a control variable that helps to disentangle their individual contributions to economic growth.

Econometrically, the reduction in the magnitude of the coefficients for savings rate and population growth upon including political stability suggests a **re-allocation of their explained variance**. It highlights how the absence of a key variable—political stability—might have led to over or underestimation of the others' influence. The kind of errors this adjustment addresses includes **specification errors** due to omitted variable bias, where political stability acts as a latent factor influencing the relationship between savings, population growth, and GDP per capita. Additionally, **multicollinearity** concerns are partially alleviated, as the shared variance between savings rate, population growth, and political stability is now more accurately attributed, leading to more stable coefficient estimates and a clearer picture of the underlying economic structure.