Regional Inequalities and economic growth; cross section and times series analysis

Anine Therese Karlsen & Mona Lisa Jones

This assignment aims to acquire, process, and analyze sub-national GDP and population data for a subset of European countries. Firstly (assignment 1), Calculateing GDP per capita and exploreing regional inequity using various descriptive statistics and visualizations.Secondly (assignment 2), examining the relationship between regional development and inequality, employing cross sectional estimation techniques for the year 2010.Lastly exploring alternative functional forms and implementing panel estimation to analzise the relationship between regional development inequality

1 Introduction

While national GDP and GDP per capita are vital indicators of a country's collective economic health, they do not shed light on how wealth or income is distributed among its residents. A high national GDP can, unexpectedly, coexist with pockets of regional deprivation (Lessmann & Seidel, 2017). The truth of this statement becomes more evident when taking a closer look at sub-national data. Regional wealth disparities are of prime concern, especially when crafting policies for equitable growth (Lessmann & Seidel, 2017). Furthermore, a country's macro-level prosperity does not automatically guarantee, that all its regions partake equally in this wealth. By studying smaller regions within a country, it is possible to get a healthier narrative about the state of regional economic disparities (Lessmann & Seidel, 2017).

The aims are to acquire, process, and analyse sub-national GDP, GINI and population data for a subset of European countries. Using regression models with time series data, cross-sectional and panel data observations we investigate economic growth and inequality trends in; Portugal, France, Hungary, Slovakia and Denmark in the time period 2000 to 2020.

Moreover, the primary objective is to gain knowledge of econometric methods used for science in business through our given topic. Subsequently, econometric terms and definitions are explained in greater detail than it normally would in similar papers.

The research question ***"What are the regional economic growth and inequality measurements from 2000 to 2020, in selected European countries and how does other***

determinants affect the results?" is key in guiding our study. Setting a solid foundation for a comprehensive study that not only measure our given parameters, but also a critically examination of the role of determinants such as transportation infrastructure and education. Lessmann & Seidel (2017) states that transportation costs play an important role in agglomeration and income, and reduced transportation costs.

The structure of our study is methodically designed to facilitate this exploration. In part 1A and B we calculate GDP per capita and explore regional inequality using various descriptive statistics and visualizations such as panel data. In part 2A and B we use cross sectional method to examine the relationship between regional development and inequality, employing cross sectional estimation techniques for the year 2010. Furthermore, exploring alternative functional forms and implementing panel estimation to analyse the relationship between regional development inequality.

Lastly, it is worth noting potential limitations may arise due to the use of ChatGPT and ChatGPT PDF, which have been utilized as tools to gain understanding and extract knowledge from academic papers, as well as to explain econometric terms and definitions. The text generated by these tools has been cross-checked to the best of our knowledge with our textbook and Andre Seidel, lecturer, and author of 'Regional Inequality, Convergence, and Its Determinants – A View from Outer Space' (2017). Despite these precautions, there may be instances where incorrect information was provided, and some errors could have been overlooked.

2 Literature Review: Regional Economic Growth, Inequality and determines (2000-2020)

Why some places grow and prosper is a fundamental question in the field of social science. This question motivated Adam Smith's [1863] work and has been a major influence of the *new economic geography* (Feldman, 2014). *Furthermore,* regional growth and inequality have been important topics of research in Europe for many years. Despite the European commission's efforts to decrease income gaps, significant disparities in income and economic development persist across regions in Europe (Lessmann & Seidel, 2017).

Economic Growth and Regional Disparities

Regional growth in Europe reveals a complex pattern, influenced significantly by factors such as geography, human capital, and economic policies. Wealthier nations like Denmark tend to experience quicker regional convergence, a phenomenon supported by [Gennaioli et al. (2014)] findings.

Financial crisis, and EU membership for Slovakia and Hungary

The 2008 financial crisis disrupted regional growth and worsened income inequalities across Europe. Nguyen et al. (2022) highlight that financial instability typically leads to increased income disparity; a trend observed globally. Focusing on Europe, we expect varied regional economic performances, especially when comparing Eastern European countries such as

Hungary and Slovakia to their Western counterparts. Given the role of capital market regulations in promoting regional convergence, outlined by Gennaioli et al (2014), we anticipate a potential growth acceleration after submission. EU membership provided Slovakia and Hungary with resources for economic recovery, while Western European countries, including Denmark, Portugal, and France, likely experienced significant economic declines due to the financial crisis. This pattern possibly mirrored in the recent COVID-19 pandemic. These complexities highlight the importance of a detailed analysis to understand regional economic growth and inequality in challenging times.

Regional inequality

The effect that surprised us most was that the economic crisis led to increase inequality[Nguyen (2022)]. indicate that any type of financial crisis results in higher income inequality. Furthermore, Lessmann & Seidel (2017) address the importance of studying these aspects. Moreover, its potential consequences, such as political tensions that can undermine social and political stability (Lessmann & Seidel, 2017).

Determinants of regional inequality

Geographic characteristics stand out as fundamental, with the natural and topographical elements of a region shifting income distributions (Lessmann & Seidel, 2017). Urbanization emerges as a crucial determinant, where active, populations around urban centres usually enjoy raised income levels, contrasting with the disparities witnessed in more rural and secluded areas (Lessmann & Seidel, 2017).

Transport infrastructure plays a central role in shaping regional economic landscapes. Improved connectivity fosters economic activities, reduces travel time, and enhances accessibility to markets and resources. Studies indicates that regions with robust transport infrastructure tend to attract more investments, generate employment, and ultimately contribute to regional prosperity.

(Lessmann & Seidel, 2017) also discusses the determinant of Education in the context of human capital. The study mentions that human capital is the most important determinant of differences in regional development within countries. Moreover, the quality of human capital within countries, as measured by the secondary-school enrolment rate, can decrease regional inequality by facilitating regional spillovers and convergence. Lastly, promoting internal migration.

Expected Inverted U- and N shaped curves

Lessmann & Seidel (2017) provides insights into the inverted U-shaped relationship between regional inequality and the level of economic development in different country groups (Lessmann & Seidel, 2017). The inverted U-shape also referred to as Kuznets Curve suggests that as an economy develops, regional inequality initially rises, reaches a peak, and then starts to decline. Additionally, there is an N-shaped relationship between regional inequality and economic development according to @lessmann2017. Subsequently, regional inequality increases again after the inverted U-shaped pattern has been completed (Lessmann & Seidel, 2017).

Lastly, (Iammarino et al., 2019) argue that there are some countries in the EU that are more evenly developed than others, and that a map of underperformance or over-performance means less in a high-income but evenly developed country, such as in Denmark in our case (Iammarino et al., 2019). On the other hand, France has been a subject of discussion and policy making for many years, regarding its well-known pockets of extreme wealth in regions such as Cap d'Antibes.

Overall, it is evident that regional economic growth and inequality in Europe from 2000 to 2020 have been influenced by several factors, with transport infrastructure and education emerging as significant determinants. Earlier studies provide us with a fundamental understanding of these dynamics, while also highlighting the impacts of economic crises on regional disparities. The integration of transport infrastructure and education adds an additional layer of complexity, underscoring its potential as a catalyst for regional development and equality.

3 Data and Methodology

3.1 Data collection

All dataset used in this paper is from Eurostat and based on the Nomenclature of Territorial Units for Statistics (NUTS) classification. The NUTS system provides a hierarchical classification of the EU's economic territory, facilitating the collection and harmonization of European statistics, and ranges from the national level (NUTS 0) to smaller regional divisions (NUTS 3).

The collected data for GDP per capita includes values in current prices and those adjusted for purchasing power parity (PPP), enabling comparisons by accounting for price level differences. GDP at market prices measures the total economic output within regions, adjusted by taxes, subsidies, and not delineated by specific sectors.

Furthermore, population data includes annual figures on births, deaths, net migration, along with demographic details. Eurostat update these datasets annually. Transportation infrastructure data encompasses the extent of the inland transport network, including motorways, railways, and waterways, within the EU, EFTA, and candidate countries. The data is reported annually by various national authorities and used to explore the link between infrastructure and regional development. Educational attainment data reflects the highest level of education completed by individuals within regions, with additional metrics employment and education status, particularly among young people. This data is derived from annual averages of the EU Labor Force Survey.

In the next section, we explore the data using descriptive statistics, crucial for understanding regional income inequality, a concept that presents challenges due to the heterogeneity of regions.

3.2 Descriptive statistics

In this section, we explore a variety of descriptive statistics, crucial for understanding regional income inequality, a concept that presents challenges due to the heterogeneity of regions. Our dataset encompasses a diverse range of regions, each varying significantly in size and population.

This paper focuses on two distinct types of statistic measures: static and dynamic. Static measures offer a cross-sectional snapshot, providing a perspective on these inequalities at a specific point in time. In contrast, dynamic measures track historical trends, offering insights into how these disparities evolve over time, a concept well-articulated in by Wooldridge (2020).

Moreover, to effectively convey our findings, we utilize graphical representations, particularly focusing on GDP per capita across various regions. These visualizations not only aid in understanding the current state of regional disparities but also in tracing their progression over time.

Mean

We calculate the mean to provide a representative value for the dataset, facilitating understanding of its central tendency and serving as a benchmark against which deviations and anomalies can be assessed, in later steps when building and interpreting regression models (Wooldridge, 2020).

MMR

Comparing the GDP per capita of the region with the highest income to the region with the lowest income (minimum GDP per capita) provides a measure of the range of disparities. If this measure is small (close to 1), the different regions have relatively equal incomes (Wooldridge, 2020). If this measure is large, then the interpretation is more problematic, as it does not tell if the high ratio is due to substantial variation in the distribution or the presence of outliers. Nevertheless, maximum to minimum ratio (MMR) provides a quick, comprehensible, and powerful measure of regional income inequality.

Standard diviation (SD)

Calculating the standard deviation (SD) quantifies the dispersion or variability of a dataset around its mean, thereby aiding in assessing the degree of uncertainty, variability, or risk associated with an economic variable or parameter. This is crucial for evaluating the reliability of estimations and predictions in economic analysis, as highlighted by Wooldridge (2020).

Median

The median serves as a robust measure of central tendency, especially when a dataset may have outliers or is skewed. Unlike the mean, the median is not influenced by extreme values and, thus, can provide a clearer picture of the "typical" value in situations where the data distribution is not symmetrical (Wooldridge, 2020).

3.3 Part 1A: Sub-national GDP and GDP per Capita

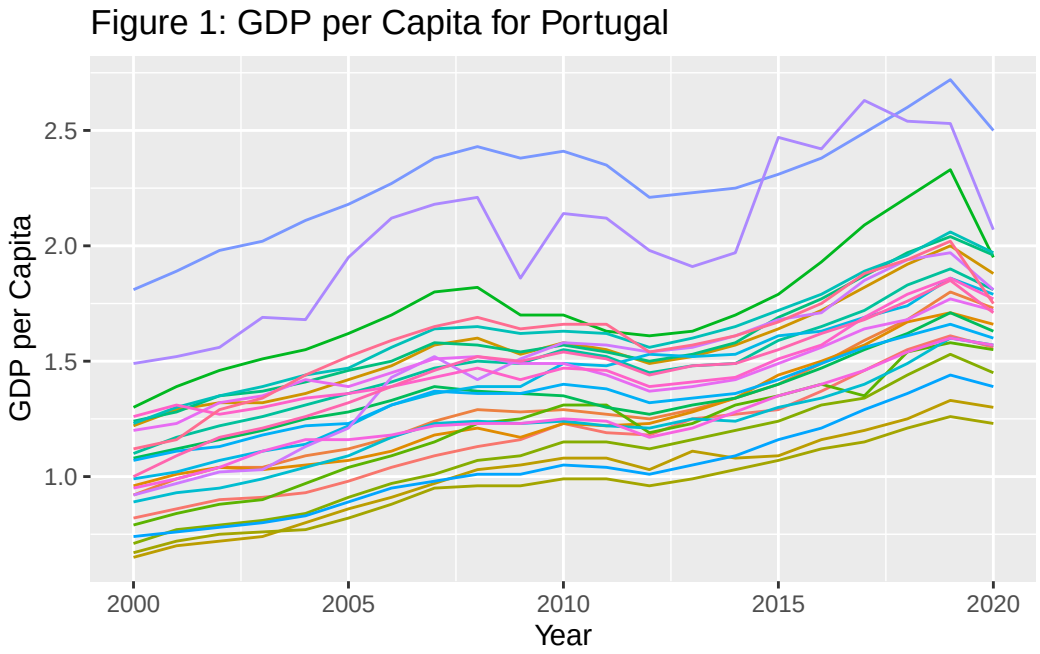
3.3.1 GDP per Capita Calculation

The formula for calculating GDP per Capita below:

$y_i = GDP_i / population_i$

In the following sections we will investigate our data at various NUTS levels and consider countries both collectively and individually. This approach allows us to discern broader regional trends, while also capturing unique national characteristics. Such a multifaceted examination is essential for a nuanced understanding of the economic dynamics within our chosen countries.

3.3.2 GDP per capita Portugal



The graph depicts a 20-year upward trend in GDP per capita across Portugal's regions, illustrating resilience and economic recovery despite various setbacks. Three regions have noticeable higher GDP per capita than the rest among those Alentejo Central (top, green) and Algarve (top, purple) as well as areas around the capital Lisbon (top, blue). The first two regions experienced significant growth, bolstered by factors like human capital development and a thriving tourism industry. This growth, however, was followed by a notable dip around 2008, reflecting these regions' sensitivity to global economic shifts and their reliance on tourism. The Metropolitan Area of Lisbon (Top Blue), while also impacted, showed less fluctuation during this period, indicative of its more diverse economic base.

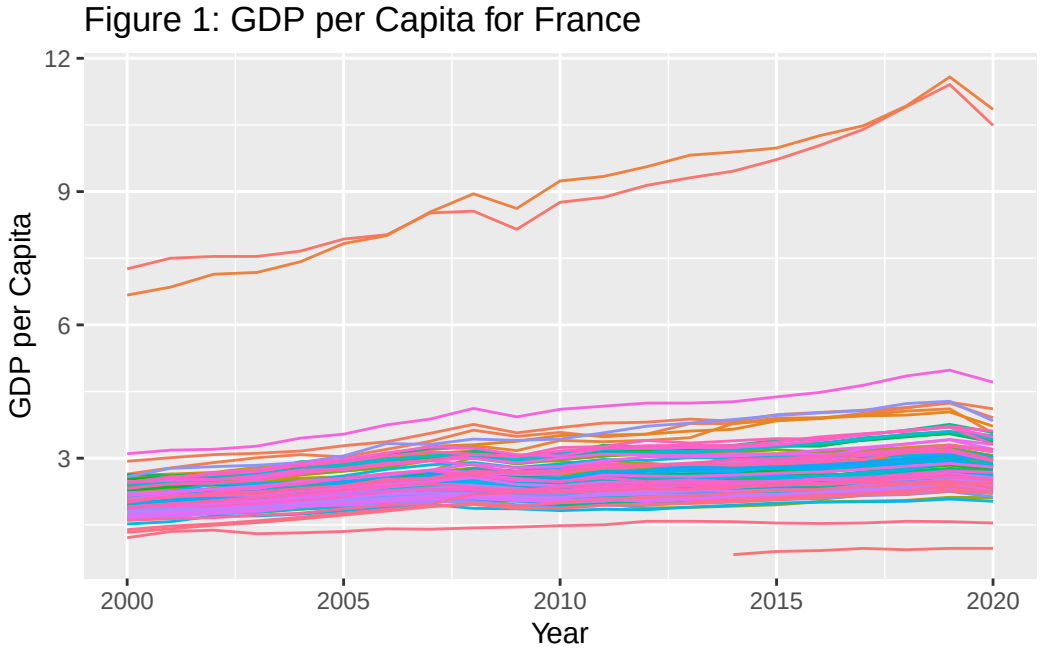
Post-financial crisis, there was a clear trend of recovery and return to growth across the regions, demonstrating their ability to adapt and rebound from economic challenges. However, the onset of the COVID-19 pandemic introduced a universal decline in all regions, underscoring the extensive impact of global health crises on regional economies. This pandemic-induced downturn mirrors global economic trends, where financial instability, as noted by Nguyen et al. (2022), often exacerbates regional economic hardships.

Portugal's regional GDP per capita is characterized by relatively consistent values across the country. The descriptive statistics of these values show a mean of approximately 1.42 and a median of 1.39. This close alignment between the mean and median suggests a balanced distribution of economic value across regions.

Furthermore, the standard deviation is 0.37, indicating a modest variation in GDP per capita across different regions. This is comparatively low when viewed against similar statistics from other countries, suggesting a more uniform economic landscape within Portugal.

Examining the range, the minimum GDP per capita is 0.65, and the maximum is 2.72. The narrow gap between these extremes further supports the observation of limited economic disparity across the regions. While urban areas like Lisbon have higher GDP per capita, indicative of economic concentration, the overall disparity across the country remains relatively low.

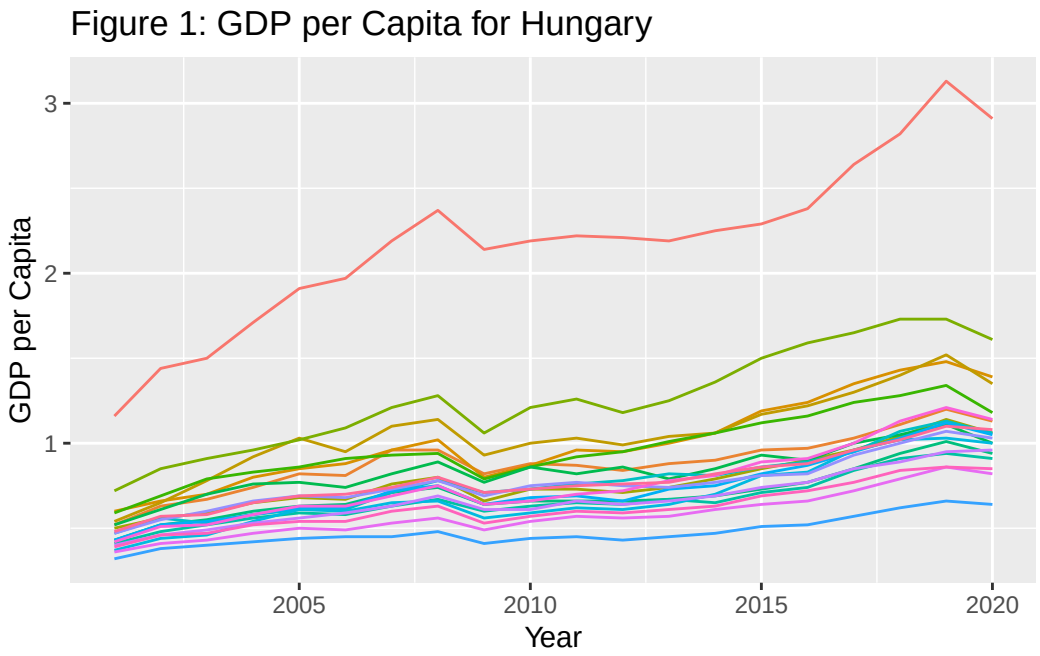
3.3.3 GDP per capita France



GDP per capita statistics from France show that the mean is 2.63, indicating an overall high economic status across regions. The median, at 2.44, is lower than the mean, suggesting a distribution skewed towards wealthier areas. The standard deviation is 1.05, highlighting

considerable regional economic variability. The range is notably wide, with the minimum GDP per capita at 0.83 and the maximum at 11.58, underscoring stark contrasts. As observed in our time series, regions like Île-de-France, encompassing Paris, and affluent areas like the southern Riviera, likely contribute to the high maximum value, illustrating a significant economic divide between urban centres and rural areas.

3.3.4 GDP per capita Hungary



HVA SKJER HER!

In Hungary, the GDP per capita data shows notable regional disparities. With a mean of 0.8598 and a median of 0.765, the higher mean suggests Budapest's significant economic influence. The standard deviation of 0.406 and the wide range from a minimum of 0.32 to a maximum of 3.13 further highlight Budapest's economic dominance compared to other regions.

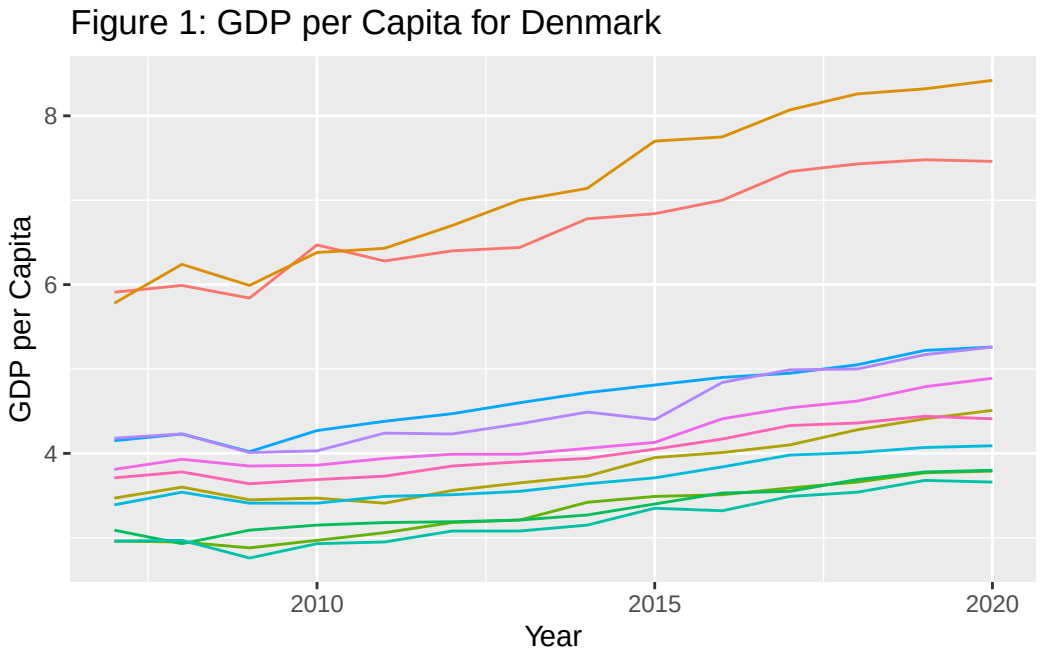
3.3.5 GDP per capita Slovakia

In Slovakia, the GDP per capita statistics reveals significant regional economic differences. The mean GDP per capita is 1.2502, higher than the median of 1.095, indicating that Bratislava, the capital and largest city, may substantially boost the national average due to its economic prominence.

The large standard deviation of 0.8018 and the extensive range from a minimum of 0.30 to a maximum of 4.02 highlighting urbanization in areas such as Bratislava's distinct economic status. These figures indicate that while most regions have moderate economic levels,

Bratislava stand out with a much higher GDP per capita, with its role as the economic centre of Slovakia.

3.3.6 GDP per capita Denmark



Beskrivelse kommer! :D

Lastly, in Denmark we observe a higher mean of 4.4192 compared to the median of 4.0000 suggesting regions like Copenhagen, with its significant economic activity, are likely elevating the national average. The standard deviation of 1.3439 indicates moderate regional economic disparities, while the range from 2.76 to 8.42 reflects Copenhagen's economic prominence relative to other regions.

3.4 Part 1B: Regional Inequity

In this part we compute the population-weighted GDP Gini coefficient for each European NUTS2 region in our assigned countries. The GINI coefficient measures inequality in our distribution, a useful tool when looking at regional inequality. The closer the GINI coefficient is to 1, the bigger the inequality; a number closer to 0 equals equality. Furthermore, looking at the GINI coefficient for NUTS 2 regions, will provide a better overview over differences in income between different regions, making it easier to address difference between the regions Hasell & Roser (2023).

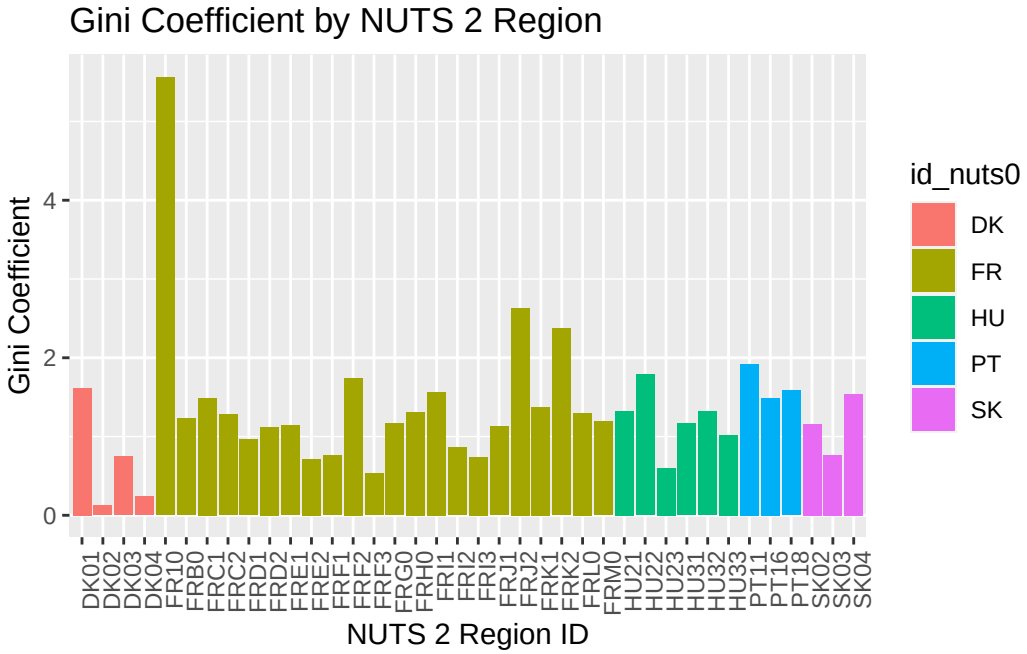
3.4.1 Gini Coefficient Calculation

Calculation the NUTS3 GDP per capita data with the following formula:

$GINW_j = \frac{1}{2\bar{y_j}} \sum_i^{n_j} \sum_l^{n_j} \frac{p_i}{P_j} \frac{p_l}{P_j} |y_i - y_l|$

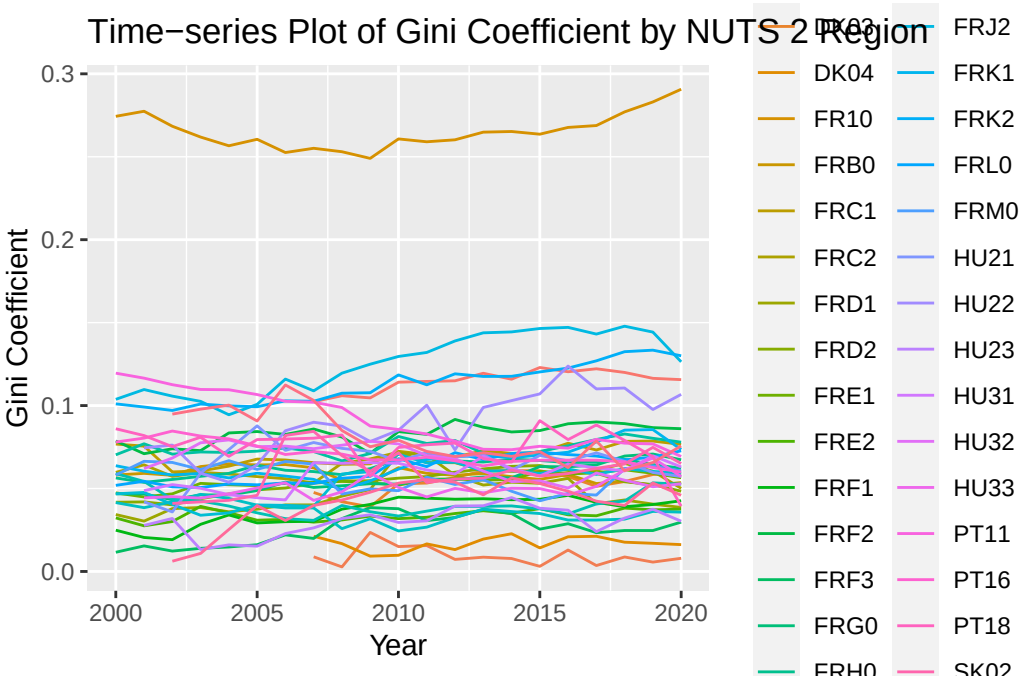
The GINI serves as the dependent variable Y, reflecting the extent of income inequality within a specific region. Furthermore, the weights Pi / pi and Pj / pj adjust the calculation based on the proportion of individual's income, to the total population. The equation considers the population size of each area to give more accurate importance to areas with larger populations. This way, income disparities in more populated areas have a proportionally greater impact on the Gini coefficient than those in less populated areas, reflecting a more realistic picture of regional inequality.

After calculating the GINI, we observe similarities to the GDP per capita data for NUTS 3 regions. By adopting different approaches to visualization such as the one provided in the bar chart below, we get a better understanding of these similarities. Moreover, looking for other important aspects.



The chart underscores the varied landscape of income inequality within European NUTS 2 regions. France's Île-de-France (FR10), Champagne-Ardenne (FRF2), Midi-Pyrénées (FRJ2), and Auvergne (FRK2) show significant disparities, likely linked to urban wealth concentration and policy. Denmark differs, with Copenhagen (DK01) and Central Jutland (DK03) displaying lower inequality, and Zealand (DK02) and North Jutland (DK04) even lower, reflecting strong equal opportunities practices. Lorraine (FRF3) in France suggests similar equity. Hungary's inequality is uniformly high except for Southern Transdanubia (HU23). Portugal's Lisbon (PT11) stands out for high inequality, while Slovakia's regions show lower disparity, akin to Denmark's average but not its lowest. Denmark emerges as having the

least inequality, with France showing the greatest and Hungary and Slovakia presenting more uniform distributions.



The box plots (in the appendix) provide a distinct perspective on the data, independent of the assumptions typically associated with other statistical distributions. This method of visualization highlights Denmark as having the highest overall GDP per capita across its regions. France, while not having the overall highest GDP per capita, demonstrates significant variability, especially when considering its outliers. Hungary, in contrast, is still presenting the lowest GDP per capita. These insights, derived from the box plot analysis, augment our understanding from previous descriptive statistics, bar plots, and time series analyses, offering a more nuanced view of the GDP per capita distribution among these countries.

4 Empirical Findings

4.1 Cross sectional estimates

4.2 Part 2A: Growth and Inequity

4.2.1 Cross Sectional Analysis

With cross-sectional data analysis we create a snapshot of the year 2010. Cross-sectional data is simpler to manage and interpret than time-series or panel data. With data from only one time point, we avoid complications arising from temporal dynamics. Cross-sectional

data allows for the comparison of different regions at the same time, which can be crucial for identifying disparities or differences between the regions (Wooldridge, 2020).

4.2.2 Simple linear regression model

In this part of the paper, we will carry out a simple regression model and explore the effect of regional economic development, represented by (GDP per capita), on regional inequality, represented by . We will do this to gain an understanding of the connection between economic growth indicators such as GDP per capita and inequality might have. We will explore if higher GDP per capita may lead to less or greater inequality, and gain an understanding to what extent these variables are related.

Unlike the traditional Gini coefficient, which treats all individuals equally regardless of the population size of the region they reside in, the weighted Gini considers the population size of each region, assigning more weight to regions with larger populations.

In the context of regional inequality, this is particularly important because it ensures that the income disparities in more populous regions have a proportionally larger impact on the overall measure of inequality. For instance, if a country has one region with a very high level of income per capita but a small population, and another region with a lower level of income per capita but a large population, the weighted Gini coefficient would reflect the inequality experienced by a larger portion of the country's population, providing a more accurate picture of the national income distribution.

We use simple linear regression to model the relationship with the GINI as the dependent variable and the natural logarithm of GDP per capita as the independent variable. Capturing the relationship between regional development and regional inequality for all regions in 2010.

Model assumptions

"The relationship between our dependent and independent variables is linear, ensuring a clear and direct connection between them. Each observation operates independently of the others, emphasizing the unique contribution of every data point. Additionally, we expect homoscedasticity, implying that the variance of the residuals remains consistent regardless of the independent variable's level. It's also crucial that, for any specified value of X, Y maintains a normal distribution. And although more pertinent to multiple regression, it's worth noting the absence of multicollinearity, ensuring that no two predictors are closely correlated."

Model specification

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Y_i represent the dependent/ explained variable

X_i represent the independent/ explanatory variable

β_1 represent the intercept/ constant

β_2 represent the slope coefficient

ε_i represents the residuals or error in the prediction.

Intercept (0): represents the value of Y when X is 0.

Slope (1): Indicates the change in Y for a one-unit change in X.

For our model we have:

$GINI = \beta_0 + \beta_1 \cdot \text{GDP per capita} + \epsilon$

4.2.2.1 Goodness of fit

The goodness of fit in simple linear regression, evaluated by R^2, gauges how well the regression line fits the data. This line is determined using the least squares criterion, minimizing squared differences between data points and the line. R^2 varies from 0 to 1, where 0 indicates no explanatory power and 1 signifies perfect fit. A higher R^2 suggests a better model fit. For an effective model, data should linearly align with minimal patterns in residuals, assuming all linear regression prerequisites are met.

Residuals vs. Fitted Values Plot is used to check for homoscedasticity and linearity. Ideally, this plot shows no pattern; the residuals are randomly scattered around the horizontal line at zero. If there's a pattern (like a curve or systematic spread of residuals), it suggests non-linearity or heteroscedasticity.

Normal Q-Q Plot to check if the residuals are approximately normally distributed. The points should fall roughly along a straight line. Deviations from a straight line suggest deviations from normality.

Model Diagnostics

We'll now look at some of the numbers we got from the linear regression model (for all countries combined):

- Coefficients:
  - Intercept is 0.0551 (expected value of gini when GDP per capita is 0). Statistically significant (indicated by p-value).
  - Estimated coefficient for GDP per capita is 0.0193, if the natural logarithm of GDP per capita increase by one, then the gini coefficient will increase by 0.0193. The p-value associated with the coefficent is however not statistically significant at 5% level.
- Goodnes of fit:
  - Multiple R-squared (0.08028) indicate that around 8% of the variability in the GINI coefficient is explained by the model. This is low, which can suggest that the model dosen't explain the variation in gini.

  - Adjusted R-squared is even lower (0.05474), is therefore expected that the model dosen't really explain the variance in the dependent variable (gini).

- Model significance:

  - The F-statistic (3.142) indicate the significance of the regression model. We can see here, that with the p-value of 0.08474, the model isn't significant at a 5% level.

By looking at these numbers, we can see that the model may not reliably predict the gini coefficient. It also suggest that GDP per capita may not be a valid predictor of the gini coefficient.

We also examined our selected countries seperately in order to see how the reliability and validity of the model might vary between countries. However, since there are too few observations for most of the countries, it makes it hard to make a conclusion of the reliability. What we can see from this examination, is that the models for Denmark, Portugal and Slovakia are not statistically significant. France seem however to have significant coefficiants, and Hungary have a moderate R-squared (but lacks significance in the slope).

Ordinary Least Squares (OLS) Estimation

The OLS method is according to Wooldridge (2020) employed to identify the best-fitting linear relationship between the dependent and independent variables, aiming to minimize the sum of squared residuals. Furthermore, this technique ensures that the estimations of the intercept (0) and slope (1) yield the least possible cumulative discrepancy between the actual and predicted values. The strength of OLS lies in its closed-form solution, providing a straightforward computation of coefficients directly from the dataset, without necessitating iterative procedures. According to Wooldridge (2020) the following assumtions need to be met to achieve BLUE.

1. ***Linear in coefficients and error term***: The model is linear in its terms and the error term.

2. ***Error term has zero mean***: The error term averages zero, ensuring unbiasedness.

3. ***No correlation between independent variables and error term***: Independent variables should not predict the error term.

4. ***Observations of error term are uncorrelated*:** Each error term observation doesn't predict the next.

5. ***Homoscedasticity*:** The error term's variance is consistent across observations.

6. ***No perfect multicollinearity***: No independent variable is a perfect linear function of others.

7. ***Normally distributed error term (Optional)***: This aids in statistical testing, although not required for unbiased OLS estimates

The residuals vs fitted values and the Q-plot, combined with the regression statistics may help us find out if classical OLS assumptions are met. The first help us check for homoscedasticity assumption of a linear regression model. The residuals should be randomly scattered around the horizontal 0 line, something that can indicate that the variances of the error terms are constant. In our plot, the residuals are in some extent randomly distributed, and there is also no clear pattern; this suggest that there is likely no significant issues with heteroscedasticity or non-linearity. The latter is used to assess if the residuals of the linear model are normally distributed. In our plot, most of the points follow the line closely, suggesting that the residuals are normally distributed. There are however some outliers in the tails, that suggest some variation from normality. Both plots are presented in appendix.

In this plot we are visualizing the relationship between GDP per capita and gini by using a mix of the two previous plots. We can here as well, see extreme outliers.

4.2.3 Part 2B: Exploring Other Determinants of Inequity

4.2.3.1 I. Data Acquisition

Se data section for more specifications on the added population nut2, transport infrastructure and education.

In order to conduct a Multiple Linear Regression model, we need to have some independent variables to use in the model and compare them with the dependent variable. The first variable, education, can explain income inequality, since it can influence income distribution in a region. If access to education is unequal, then higher education levels might incease income disparities (Rodriguez-Pose & Tselios, 2008). The population density, our second variable, can explain inequality since regions with a higher population density might have different economic behaviours. Our last variable, rail network (infrastructure), can influence economic development and accessibility, which also can affect income inequality in a region (Chatterjee & Turnovsky, 2012).

4.2.3.2 II. Multiple Linear Regression Model

Multiple Linear Regression (MLR) extends simple linear regression to incorporate multiple explanatory variables, allowing us to examine how multiple factors impact a dependent variable. Choosing a data set from the year 2010 that consists of various regions, with data on each region's economic indicators, demographic variables, and other factors. Our aim is to understand how these variables collectively affect regional inequality.

We will in this part do a Multiple Linear Regression model by using the variables education (in percentage of pupils and students in education, % of total population), population density and rail network in km. This model will tell us if these variables can help explain change in the gini coefficient.

In both our simple linear regression model, and now in our multiple linear regression model, we use the logarithmic function which makes it easier to linearize the relationship between the variables. By using it for GDP per capita, we can reflect changes more effectively. For

rail network and population density, the logarithm function ensure that the model capture proportional changes and deals better with the wide range of values.

4.2.3.3 II. Model specification

Understanding the Coefficients

Intercept beta_0 Represents the expected value of the dependent variable when all independent variables are set to zero. Interpretation is often nonsensical in multiple regression if there is no meaningful condition where all predictors are zero.

Slope Coefficients beta_1, beta_2, ..., \beta_k \))**: Represent the expected change in the dependent variable for a one-unit change in the respective independent variable, holding all other variables constant.

4.2.3.4 III. Model Interpretation

Our first model in the Multiple Linear Regression model examine how education in addition to GDP per capita can help explain the gini coefficient. The second model look at both education, GDP per capita, and population density, while the third model examine them all and also add rail network.

For model 1, the adjusted R-squared is 0.033, indicating that the model explains around 3% of the variability in the gini coefficient. Since this is relatively low, it suggests that model 1 is not suitable in explaining the variance in gini. Both variables have a p-value above 0.05, indicating that they are not statistically significant.

For model 2, the adjusted R-squared is 0.300, which indicate that the model explains around 30% of the variability in the gini coefficient. This is pretty high, suggesting that model 2 is suitable in explaining the variance in gini, meaning that adding population density improves the model's explanatory power. The p-value for population density is less than 0.01, which means that its statistically significant.

The adjusted R-squared for model 3 is 0.370 ~ 37%; this suggest that this model has the greatest fit of all 3 models. Population density is significant at a 0.05 level. The p-value for rail is not below 0.05, and is therefore not statistically significant.

To summarize, we can see that the population density is the variable that affects the gini coeffcient the most. The more people that live in an area, the bigger the economic inequality is.

4.3 Alternative Functional Forms and Panel Estimates

4.3.1 Testing Development Effects Across Subsets

4.3.1.1 Subset Analysis

When exploring whether the effect of development and economic inequality is significantly different between different subsets. We look at the variantions between our biggest country (France) compared to the rest of the countries.

Choosing France as a dummy variable amongst our group of countries could be justified based on several criteria. Firstly, France stands out as one of the largest economies in this group, both in terms of GDP and global economic influence. This distinction makes France a unique case compared to the smaller economies of Denmark, Portugal, Hungary, and Slovakia. Secondly, France's larger geographical area and population size provide different economic dynamics and complexities, which are not as pronounced in smaller and more homogenous countries. Thirdly, France's economy is more diversified compared to the other countries listed, which might have more specialized or less varied economic structures. This diversity can lead to different patterns in wealth distribution and economic development. Lastly, France's policies, particularly in social welfare and economic regulation, and its larger domestic market, might influence economic outcomes differently than in the smaller economies, where external trade and different policy contexts play a more significant role.

	France	All other countries
(Intercept)	-0.006 *	0.060 ***
	(0.003)	(0.002)
GDP_per_capita	0.028 ***	0.001
	(0.001)	(0.001)
N	462	295
R2	0.637	0.004
FStat	805.761	1.109
PValue	0.000	0.293

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ T statistics in brackets.

4.3.1.2 Subset Analysis Discussion

For France, the regression model shows us a significant and positive relationship between GDP per capita and the Gini coefficient. The intercept is statistically significant, but negative. The positive coefficient of GDP per capita of 0.028 is highly significant with a p-value lower than 0.001; this suggests that when GDP per capita in France rises, then economic inequality rises correspondingly as well. We can also see this by looking at the R-squared of 0.637, indicating that around 63.7% of the variation in the economic inequality can be explained by changes in economic development (GDP per capita). To summarize our findings for France, we can see that economic growth in France can be associated with rising

economic inequality, something that could possibly be due to a concentration of wealth in some regions (like Paris for example).

When comparing France to the rest of the countries, we observe a different situation. Showing the relationship between GDP per capita and Gini is not statistically significant. This model has a positive, but non-significant coefficient for GDP per capita, which can suggest that variations in this variable does not have a significant impact on economic inequality. This is also shown by looking at the low R-squared of $0.004 \sim 0.4\%$, this may mean that GDP per capita only explains a rather small portion of the variance in inequality.

Overall we record major differences between France and the other countries when it comes to the relationship between economic development and economic inequality.

4.4 Exploring Alternative Functional Forms

4.4.1 Functional Form Exploration

In addition to the linear model, other models such as logarithmic transformation, the quadratic term and the cubic are useful to explain the relationship between regional development and economic inequality . By using these additional models, we can test different transformations of the variables in order to find the best representation of the data. The addditional models are explained in more details below.

Logarithmic transformation

The logarithmic transformation for all the variables allows the coefficients to be interpreted as elasticizes, which measure the percentage change in the dependent variable associated with a one percent change in the independent variable (Lessmann & Seidel, 2017). The use of logarithmic transformations is a common practice in econometrics to address issues such as heteroscedasticity and non-linearity in the data. By using logarithmic transformations, we are able to estimate the relationship between regional inequality and development in a more robust and accurate way (Lessmann & Seidel, 2017).

The quadratic term

The quadratic term in the regression model is used to investigate the relationship between regional inequality and development. The results suggest an inverted U-shaped relationship between income and inequality, as indicated by a negative coefficient for the quadratic term (Lessmann & Seidel, 2017). This implies that initially, as income rises, inequality increases, but after reaching a certain threshold, further increases in income lead to a decrease in inequality (Lessmann & Seidel, 2017).

The cubic function

The cubic function allows for the exploration of an N-shaped relationship between two variables. As suggested by Lessmann & Seidel (2017), this can be particularly relevant when studying the link between regional inequality and development. In such a scenario, inequality might initially decrease with development, then increase, and eventually decrease again at very high levels of development.

Including a cubic term in a regression model enables the analysis of intricate dynamics, such as changing marginal effects. This is crucial in understanding how the impact of a predictor variable on the outcome variable varies at different levels of that predictor.

4.4.2 Estimation and Visualization

	LinearFR	LinearC	LogFR	LogC	QuadraticFR	QuadraticC
(Intercept)	-0.006 *	0.060 ***	-0.034 ***	0.062 ***	0.073 ***	0.060 ***
	(0.003)	(0.002)	(0.005)	(0.002)	(0.008)	(0.002)
GDP_per_capita	0.028 ***	0.001				
	(0.001)	(0.001)				
log(GDP_per_capita)			0.110 ***	0.001	-0.035 ***	-0.012 **
			(0.005)	(0.002)	(0.010)	(0.004)
I(GDP_per_capita^2)					0.003 ***	0.001 ***
					(0.000)	(0.000)
N	462	295	462	295	462	295
R2	0.637	0.004	0.490	0.000	0.682	0.061
FStat	805.761	1.109	442.257	0.113	491.706	9.491
PValue	0.000	0.293	0.000	0.737	0.000	0.000

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ T statistics in brackets.

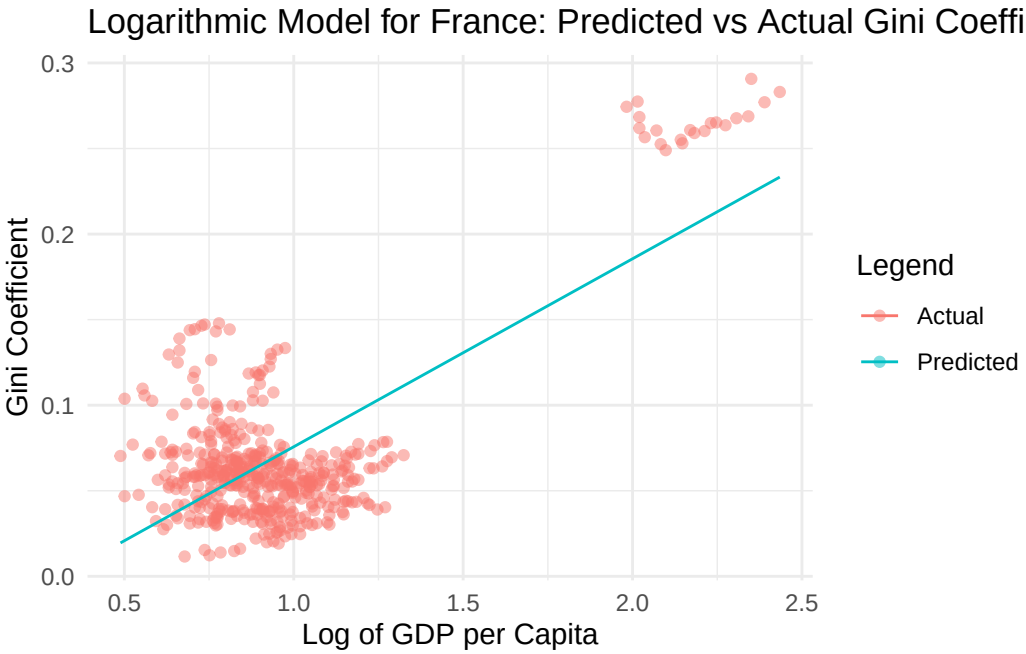
As earlier predicted there is a significant positive correlation between GDP per capita and Gini for France, while its non-significant for the rest of the countries. In the logarithmic model, R-squared of 0.490 in the French data set demonstrate this. Showing that the model does not explain much of the variance in the dependent variable. Other variables not included in the model might have a significant impact. While the model may still be statistically significant, the low R-squared value suggests that its predictive power is limited. However, the other countries still does not show a significant relationship, with results showing R-squared at zero. Indicates that the independent variables provide very little information about the variation in the dependent variable. This suggests that the model is not a good predictor of the dependent variable. It might be necessary to re-evaluate the model, consider additional variables, or review the underlying assumptions.

For the quadratic model, France show positive coefficient for squared term of GDP per capita, something that can suggest a U-shaped relationship between the variables. The R-squared for France is 0.682, higher than the R-squared for the other models. This implies

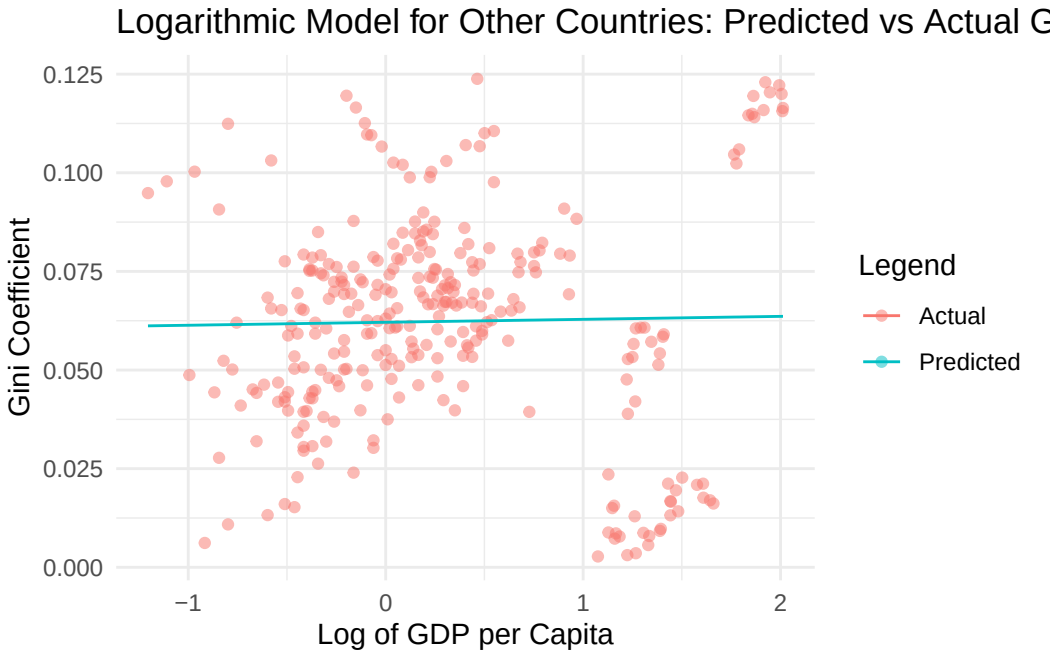
that the model explains a moderate amount of the variance in the dependent variable. It is a resanoble fit and such values are common in social sciences and economics where the behavior or phenomena being modeled are influenced by many unobserved factors.

The other countries also indicates a U-shaped relationship, despite R-squared levels as low as 0.061 in other words only about 6.1% of the variation in the dependent variable is explained by the independent variable(s) in our model. This indicates a low level of explanatory power. Overall, the quadratic model is a better fit than the other models to explain the variance in income inequality.

4.4.2.1 Logarithmic Model



Even with a logarithmic model, extreme outliers such as those spotted above, at the end of the line can indicate that the relationship captured by the model does not hold as well in these atypical circumstances. It suggests that the economic dynamics of these extremely wealthy regions differ significantly from the average. It's important to carefully assess how these outliers influence our model and consider additional or alternative approaches to capture the complex nature of the relationship between GDP per capita and the Gini coefficient accurately.

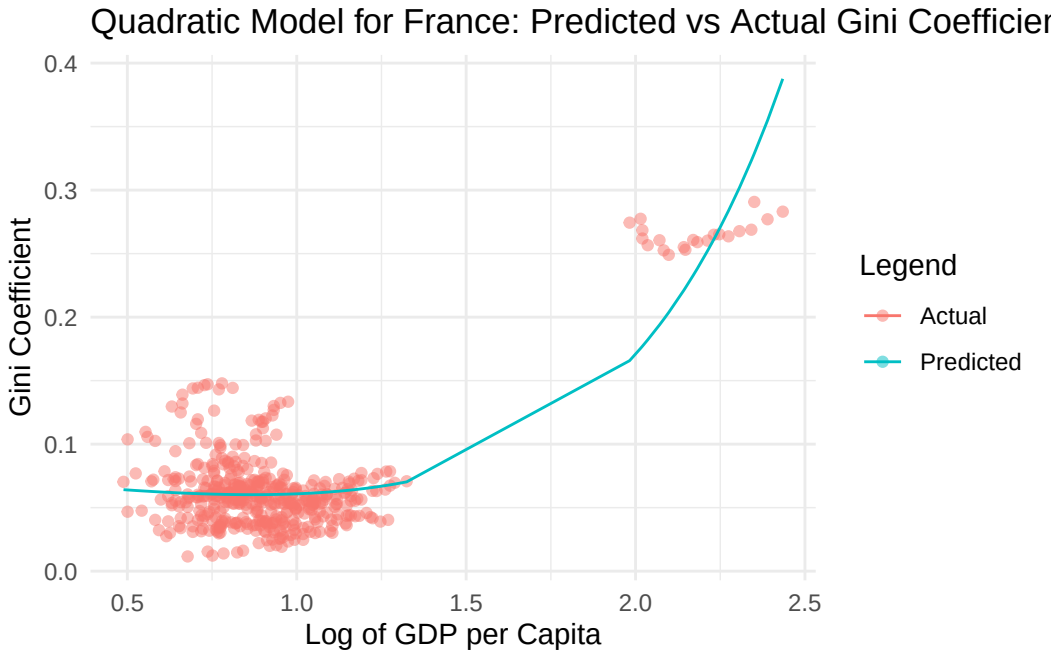


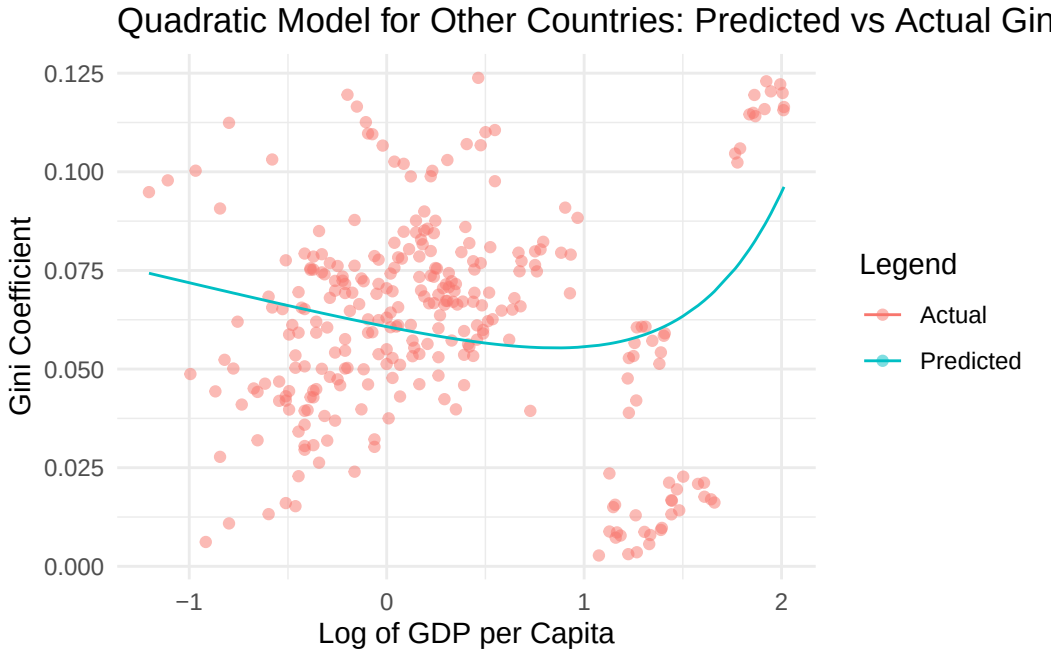
In the plot for France, we can see an upward trend, which indicated that as the log of GDP per capita increases, the gini coefficient also increases, something that suggests higher income inequality. The data points are scattered around the predicted regression line, which goes upward. This pattern also supports the regression model that shows a significant positive realtionship between the variables. There are some extreme outliers, suggesting some regions have a higher amount of economic activity (like Paris that is a economic center in France). This also shows that economic growth may lead to increased inequality because of the concentration of wealth in some regions.

The model for the other countries show that there is not a significant relationship between the variables. The data points also dosen't cluster around the predicted line, which shows that the log of GDP does not explain the variance in the Gini coefficient.

When analyzing both plots, it's clear that the relationship between economic development and income inequality behaves differently across these regions. The plot for France suggests that economic growth is associated with rising inequality, while the plot for the other countries indicates that economic growth has little to no effect on inequality

4.4.2.2 Quadratic Model





For France, the Gini coefficients are initially concentrated at lower levels of GDP per capita, presenting stability or a slight decrease in inequality, which then transitions into a sharp upward trend at higher GDP levels. This is indicative of an inverted U-shaped curve, a characteristic of the Kuznets curve hypothesis, suggesting that inequality escalates significantly with further economic growth (Ota, 2017). The quadratic model appears to fit the lower and middle range of the data well, yet it deviates from the actual data points at the higher end of GDP per capita, signaling that other factors may come into play as the economy

grows.

The quadratic model for the other countries shows a more subtle U-shaped relationship, with the initial part of the curve being relatively flat. This indicates that changes in GDP per capita have a limited effect on income inequality at first. However, a steep increase is observed at the higher end of GDP per capita, which may be attributed to outliers or specific country conditions not captured by the model.

Incorporating these observations, both plots still signify a non-linear relationship between GDP per capita and income inequality, but with varying patterns and intensities between France and the other countries. The pronounced curve for France supports a stronger Kuznets curve effect, while for the other countries, the effect is less pronounced and the connection between economic growth and inequality appears to be weaker and more influenced by external or unmodeled factors.

4.4.2.3 Results Interpretation

The linear model, while providing a significant fit for France, falls short for the other countries, suggesting that the economic inequality's relationship to GDP per capita may not be linear or might be influenced by factors not captured in a simple linear model. The logarithmic model's reduced explanatory power compared to the linear model for France suggests that the relationship between GDP and inequality may not be strictly proportional across all levels of GDP per capita.

The quadratic model's U-shaped curve implies that as countries develop, inequality may first decrease and then increase, supporting the Kuznets curve hypothesis.

In conclusion, while the linear model provides a baseline understanding of the relationship between GDP per capita and income inequality, the logarithmic and quadratic models offer more nuanced insights that can lead to a deeper understanding of the underlying economic mechanisms. These alternative models highlight the importance of considering non-linear dynamics when analyzing economic relationships.

4.4.3 Heteroskedasticity Testing and Causality Discussion

4.4.3.1 Heteroskedasticity Testing

Heteroskedasticity is according to Wooldridge (2020) an important factor to consider in econometrics and statistical modelling, especially in the context of regression analysis. Heteroskedasticity refers to the condition in regression analysis where the variance of the error terms (residuals) is not constant across all levels of the independent variables [Wooldridge (2020)]. In simpler terms, the spread of the residuals varies at different points in the regression model. Moreover, heteroskedasticity can lead to inefficient estimators and unreliable hypothesis tests. Leading to the assumption of homoscedastic in MRL not being met. In other words the variance of the error terms is not constant across all levels if the independent variables when heteroskedasticity is present.

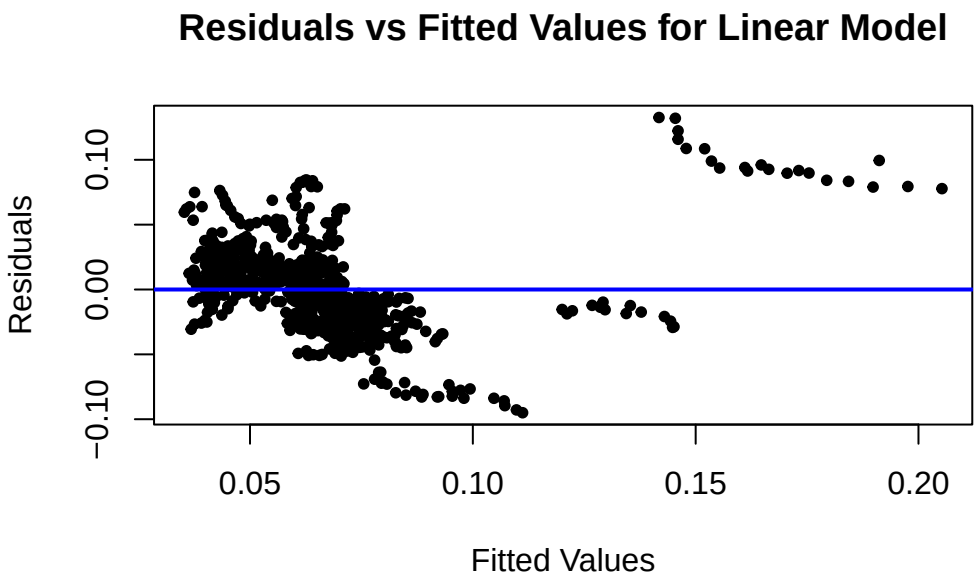
By conducting a Breusch-Pagan we can detect heteroskedasticity in our regression model. The test works by checking if the variances of the errors from the regression are dependent on the values of the independent variables. A higher BP value indicates a stronger presence of heteroskedasticity. Degrees of Freedom refers to the number of independent variables in the model. P-value determines the significance of the test result. A low p-value (typically less than 0.05) suggests that the null hypothesis of homoskedasticity (constant variance of errors) can be rejected, indicating heteroskedasticity. If heteroskedasticity is detected, the standard errors of the regression coefficients may be biased, which can lead to incorrect inferences in hypothesis testing.

In the case of the France, the BP value is 0.25123 with 1 degree of freedom, and a p-value of 0.6162, suggesting no significant evidence of heteroskedasticity, as the p-value is well above the typical threshold of 0.05. However, for the rest of the countries, the BP value is much higher at 83.806, with 1 degree of freedom, and the p-value is less than 2.2e-16, indicating a strong presence of heteroskedasticity. This is also the case for the log model for France and for the rest of the countries. The log model for France has a BP value of 22.971 and a p-value of 1.645e-06, while the log model for the rest shows a BP value of 60.814 and a p-value of 6.273e-15. Both indicate significant heteroskedasticity. The quadratic models exhibit similar trends. The quadratic model for France shows a BP value of 59.658 with 2 degrees of freedom and a p-value of 1.11e-13, which clearly indicates heteroskedasticity. Similarly, the quadratic model for the rest, with a BP value of 18.4, 2 degrees of freedom, and a p-value of 0.0001011, suggests notable heteroskedasticity.

Overall, these results, particularly the low p-values in most models, suggest that the null hypothesis of homoskedasticity (constant variance of errors) can be rejected, revealing heteroskedasticity. It is essential to correct for heteroskedasticity to ensure the reliability of the regression analysis. There are several methods to deal with heteroskedasticity, depending on the nature of the data and the specific requirements of the analysis. Residuals vs. Fitted is another way of detecting heteroskedasticity. However, it provides a deeper understanding of the heteroskedasticity's nature and extent

Residuals vs. Fitted

The residuals versus fitted values plot is another diagnostic tool that helps in visually identifying issues in a regression model, aiding in improving model accuracy and reliability (Wooldridge, 2020). Even after confirming heteroskedasticity. According to Wooldridge (2020) It provides a deeper understanding of the heteroskedasticity's nature and extent, which is crucial for effectively addressing it in our regression analysis. Furthermore, the interpretation of different patterns in the plot, like random dispersion indicating a good model fit or systematic patterns indicating issues. Subsiquently, ensuring robust and reliable modeling visual diagnostics, complement statistical tests [Wooldridge (2020)].



There appears to be a pattern in the residuals; they are not randomly dispersed around the horizontal line but show a curve-like trajectory as the fitted values increase. This pattern suggests non-random error variance, which is indicative of non-linearity in the relationship between the independent and dependent variables, or it could also suggest heteroskedasticity. Ideally, if the model was a perfect fit, the residuals would be randomly scattered around the zero line (the horizontal blue line in your plot) without any discernible pattern. This would indicate homoskedasticity and a good linear fit. The pattern we are seeing in our residuals plot suggests that our linear model may not be the best representation of the relationship between the variables. It could be that a non-linear model would provide a better fit or that there are other variables or transformations that should be considered. Understanding this plot is important because it directly relates to the assumptions underlying linear regression.

The observed patterns and their possible implications, suggests that our model might be missing a nonlinear relationship between the dependent and independent variables. This is a sign of model misspecification, confirming that the current linear model may not be the best fit for our data. The variance in the residuals first increasing, then decreasing, and increasing again, implies that the error terms have non-constant variance, which is a sign of heteroskedasticity.

4.4.4 Causality Discussion:

The statistical relationship between two variables does not imply that one causes the other. Correlation can be due to a causative relationship or due to both variables being influenced by a third factor. Establishing causation means demonstrating that changes in one variable directly result in changes in another. To claim causality, econometricians typically need to

demonstrate three things. Firstly, temporal Precedence meaning the cause must precede the effect in time. Secondly, covariation of the Cause and Effect. When the cause happens, the effect happens; when the cause doesn't happen, the effect doesn't happen. Lastly, no Plausible Alternative Explanations meaning there should be no other factors that can explain the relationship between the cause and effect.

4.5 Panel Estimates

Since panel data has a two-dimensional structure, with observations on multiple entities across several time periods. It enables us a more detailed and complex analyses compared to purely cross-sectional or time-series data. The data will then allow us to examine both cross-sectional (differences between subjects at a point in time) and longitudinal (changes within subjects over time) effects. For instance, this approach would be apt for studying how policy changes within a region affect its GDP and inequality levels. Fixed-Effects and Random-Effects Models in panel data analysis help in controlling for unobserved heterogeneity, thereby providing a clearer picture of whether there's a causal relationship.

Fixed effect models

Panel data estimation is a method used in econometrics to analyse data that involves observations over multiple time periods. In the context of panel data estimation specifics of fixed-effects and random effect are used. (Lessmann & Seidel, 2017) argues that the fixed-effects model is a reasonable approach when the differences between countries (or regions) can be viewed as parametric shifts of the regression function. However, the random-effects model allows for time-invariant unobserved heterogeneity across regions and is more appropriate when the fixed-effects model is too restrictive (Lessmann & Seidel, 2017).

Random-Effects Models

Lessmann & Seidel (2017) mention the use of a random-effects model to investigate the determinants of within-country changes in inequality. The random-effects model controls for several country-level fixed factors (national income, number of regions, and area) and fixed effects for various country groups (Lessmann & Seidel, 2017). The advantage of the random-effects model is that the expected value of the country-specific effect is zero, which means that there is no need to apply any arbitrary data imputation procedure for the missing intercepts (Lessmann & Seidel, 2017). However, this approach may come at the cost of founding the predictions on a slightly biased coefficient (Lessmann & Seidel, 2017). Lessmann & Seidel (2017) also notes that the major coefficient of interest is not sensitive to applying either a fixed-effects model or a random-effects model with additional country and region information.

4.5.1 Panel Estimation Task:

4.5.2 Panel Estimation Analysis

4.5.3 Panel Estimation Discussion

5 Discussion

6 Limitations and Future Research

7 Conclusion

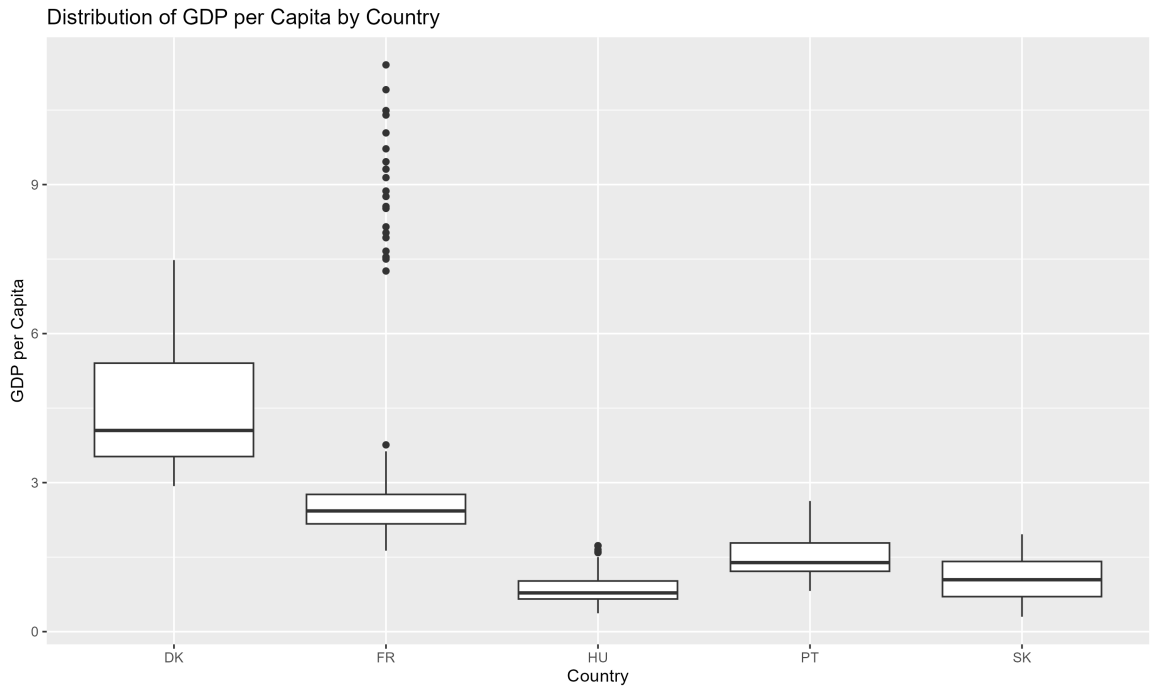
8 Appendix

1. **Descriptive statistics**

```
Summary Statistics for Portugal's GDP Per Capita:

        GDP_per_capita
mean         1.4185524
median       1.3900000
std_dev      0.3702905
minimum      0.6500000
maximum      2.7200000


Summary Statistics for France's GDP Per Capita:

        GDP_per_capita
mean          2.630444
median        2.440000
std_dev       1.053767
minimum       0.830000
maximum      11.580000


Summary Statistics for Hungary's GDP Per Capita:
```

```
        GDP_per_capita
mean         0.8598000
median       0.7650000
std_dev      0.4059723
minimum      0.3200000
maximum      3.1300000


Summary Statistics for Slovakia's GDP Per Capita:

        GDP_per_capita
mean         1.2501948
median       1.0950000
std_dev      0.8018259
minimum      0.3000000
maximum      4.0200000


Summary Statistics for Slovakia's GDP Per Capita:

        GDP_per_capita
mean          4.419221
median        4.000000
std_dev       1.343933
minimum       2.760000
maximum       8.420000
```

2. **Boxplot**



3. **Regression statistics of all countries separately for the year 2010**

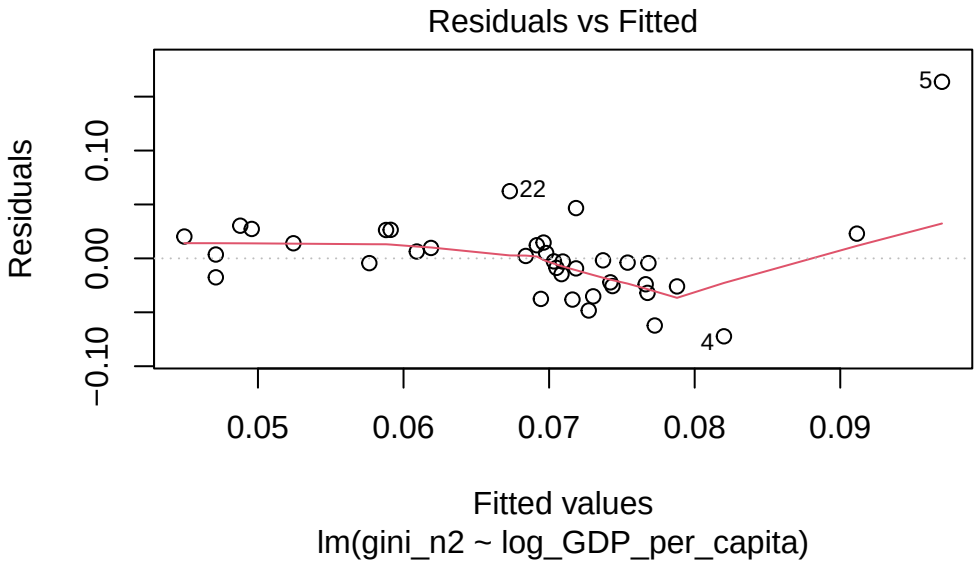
```
Intercept (Coefficient 1): 0.05512314

Slope of log GDP per capita (Coefficient 2): 0.01929064
```

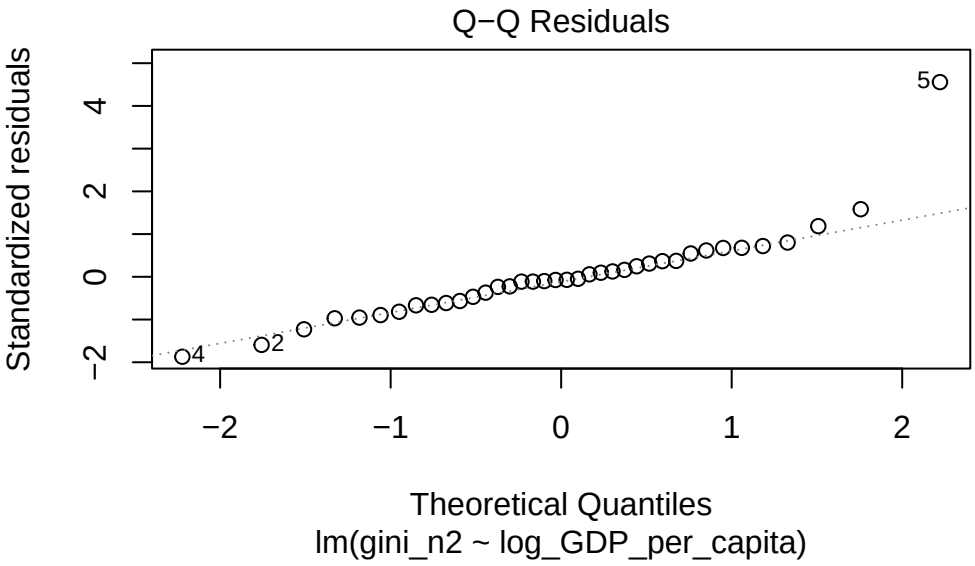
		SK	DK	HU	PT	FR
	(Intercept)	0.069	-0.127	0.075 **	0.079	-0.028
		(0.010)	(0.085)	(0.010)	(0.015)	(0.026)
	log_GDP_per_capita	-0.018	0.124	0.049	-0.008	0.107 ***
		(0.034)	(0.059)	(0.029)	(0.030)	(0.026)
	r.squared	0.221	0.689	0.410	0.067	0.457
	adj.r.squared	-0.558	0.534	0.262	-0.866	0.430
	statistic	0.284	4.432	2.778	0.072	16.820
	p.value	0.689	0.170	0.171	0.833	0.001

*** p < 0.001; ** p < 0.01; * p < 0.05.

Column names: names, SK, DK, HU, PT, FR

Testing for OLS assumtions





References

Chatterjee, S., & Turnovsky, S. J. (2012). Infrastructure and inequality. *European Economic Review*, *56*(8), 1730–1745.

Feldman, M. P. (2014). The character of innovative places: Entrepreneurial strategy, economic development, and prosperity. *Small Business Economics*, *43*(1), 9–20. https://doi.org/10.1007/s11187-014-9574-4

Gennaioli, N., La Porta, R., Lopez De Silanes, F., & Shleifer, A. (2014). Growth in regions. *Journal of Economic Growth*, *19*(3), 259–309. https://doi.org/10.1007/s10887-014-9105-9

Hasell, J., & Roser, M. (2023). Measuring inequality: What is the Gini coefficient? *Our World in Data*.

Iammarino, S., Rodriguez-Pose, A., & Storper, M. (2019). Regional inequality in europe: Evidence, theory and policy implications. *Journal of Economic Geography*, *19*(2), 273–298. https://doi.org/10.1093/jeg/lby021

Lessmann, C., & Seidel, A. (2017). Regional inequality, convergence, and its determinants – a view from outer space. *European Economic Review*, *92*, 110–132. https://doi.org/10.1016/j.euroecorev.2016.11.009

Nguyen, T. C. (2022). The effects of financial crisis on income inequality. *Development Policy Review*, *40*(6), e12600. https://doi.org/10.1111/dpr.12600

Ota, T. (2017). Economic growth, income inequality and environment: Assessing the applicability of the Kuznets hypotheses to Asia. *Palgrave Communications*, *3*(1), 1–23.

Rodriguez-Pose, A., & Tselios, V. (2008). *EDUCATION AND INCOME INEQUALITY IN THE REGIONS OF THE EUROPEAN UNION*. https://click.endnote.com/viewer?doi=10.1111%2Fj.1467-9787.2008.00602.x&token=WzM0MzQ0MzksIjEwLjExMTEvai4xNDY3LTk3ODcuMjAwOC4wMDYwMi5

Avuk84iH8myPb7qaaaJIQk.

Wooldridge, J. M. (2020). *Introductory Econometrics - A Modern Approach* (7th ed.). Cengage.