Assignment 2:

Regional GDP Inequality in 4 Selected European Economies - Exploring Determinants

Kristoffer Tufta and Harald Bjarne Vika

Friday 12 Dec, 2025

1 Part A: Growth and Inequality

In this assignment we build on the foundation we made in assignment 1. Using the calculated regional inequality and gdp per capita, we estimate a simple cross-sectional Ordinary Least Squares (OLS) regression for the year 2017 in our selected countries, broken down by NUTS2-region.

1.0.1 Expected Relationships

Basing ourselves on the theory established by Lessmann & Seidel (2017), we predict that we can see a link between regional inequality and economic growth. They described the relationship as an "inverted U, or maybe even an N-shaped" relationship between inequality and development. In short, this implies that regional inequality is at it's lowest for low and high-income countries, but higher for middle-income countries. Based on this theory, we expect a negative slope coefficient: regions with stronger growth from 2016 to 2017 should, on average, exhibit slightly lower inequality.

1.1 Making the regression model

1.1.1 Preparing the data

First point of action: data preparation. In the following code chunk we read the data produced in document 1, and select our needed columns. We then sum up the totals of GDP and population for each NUTS2-region. Since we want to look at a cross-section of the year 2017 - including a change in gdp, we have to do a little bit of calculation. We filter our "Time"-column so that it only shows the years 2016 and 2017, and then we add a new column called "Lagged_capita". The "Lagged_capita"-column shows the GDP value for the previous year, and lets us add another column where we calculate the percentage change in GDP, on just one row. Since 2015 has been filtered away, we see some NA-values in all the rows showing GDP change and lagged capita for 2016, but we will filter these away in the next step anyways, so no harm done.

NUTS2	Country	Time	N2GDPSUM	N2POPSUM	N2GDPCAP
character	character	character	numeric	numeric	numeric
CH01	CH	2016	115,578.8	1,593,839	72,516.0
CH01	CH	2017	112,845.8	1,613,522	69,937.6
CH02	CH	2016	123,376.1	1,842,251	66,970.3
CH02	CH	2017	122,435.2	1,859,557	65,841.0
CH03	CH	2016	89,578.7	1,128,723	79,362.9
CH03	CH	2017	90,156.5	1,142,156	78,935.4
CH04	CH	2016	133,701.9	1,466,424	91,175.5
CH04	CH	2017	133,356.3	1,487,969	89,623.0
CH05	CH	2016	74,684.9	1,153,485	64,747.2
CH05	CH	2017	73,818.2	1,162,684	63,489.5
n: 106					

We then join the table showing our calculated Gini-coefficients from assignment 1 onto this newly prepared table, and filter down to the year 2017.

NUTS2	Country	Time	N2GDPSUM	N2POPSUM	N2GDPCAP
character	character	character	numeric	numeric	numeric
CH01	CH	2017	112,845.8	1,613,522	69,937.6
CH02	CH	2017	122,435.2	1,859,557	65,841.0
CH03	CH	2017	90,156.5	1,142,156	78,935.4
CH04	CH	2017	133,356.3	1,487,969	89,623.0
CH05	CH	2017	73,818.2	1,162,684	63,489.5
CH06	CH	2017	56,830.2	799,287	71,101.2
CH07	CH	2017	26,334.0	354,375	74,311.2
DE11	DE	2017	213,043.3	4,098,278	51,983.6
DE12	DE	2017	121,332.5	2,779,314	43,655.6
DE13	DE	2017	84,325.9	2,239,734	37,650.0
n: 53					

1.1.2 Model Specification

The model will examine if intraregional inequality - represented by a calculated Gini-coefficient of GDP per capita in each NUTS2 region - is associated with short-term economic development across NUTS2 regions:

$$\text{Gini}_{i,2017} = \alpha + \beta\, \Delta\text{GDPpc}_{i,2016\to2017} + u_i$$

Where:

- $\text{Gini}_{i,2017}$ — Calculated Gini coefficient of region i in 2017 (regional GDP per capita inequality)

- $\Delta\text{GDPpc}_{i,2016\to2017}$ — percent change in GDP per capita from 2016 to 2017 in region i

- α — intercept term (baseline inequality when GDP-per-capita change is zero)

- β — slope coefficient showing how inequality changes with a one-percentage-point increase in GDP-per-capita growth

- u_i — error term capturing unobserved regional factors

Variable selection

Dependent variable: Regional inequality (Gini) was chosen because it quantifies GDP inequality within each NUTS2 region and directly represents the outcome discussed in the literature on regional convergence.

Independent variable: Change in GDP per capita captures short-term economic development, making it suitable to test whether faster-growing regions are converging or diverging in terms of inequality.

Both variables are derived from Eurostat data prepared in assignment 1.

1.1.3 Model Diagnostics

If we take a look at the results of the model in Table **??**, we see that the intercept ($\hat{\alpha} = 0.1336$) reflects average inequality when GDP growth is zero, while the slope ($\hat{\beta} = -0.0024$) suggests a small, negative, and statistically insignificant relationship between GDP-per-capita growth and regional inequality ($R^2 = 0.012$). We can determine that the relationship is statistically insignificant due to the (quite large) P-value of 0,435.

In other words, faster-growing NUTS2 regions in 2017 tended to have slightly lower inequality, but the effect is small and not statistically meaningful. The R-squared indicates an explanatory power of around 1,2 %, meaning that other variables likely explain much more of the variation.

Table 3

statistic	p.value	parameter	method
1.1	0.2914	1.0	studentized Breusch-Pagan test

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

Testing for heteroskedasticity via Breusch-Pagan testing of our model.

In Table 3 we check for heteroskedasticity by running a Breusch-Pagan test on our regression model. With a P-value of 0,29 (more than <0,05), we conclude that we have homoskedasticity.

1.1.3.1 OLS Assumptions

The "ordinary least squares"-method includes a few classical assumptions that have to be validated for the model to produce the best possible estimates.

1. **Linearity:** The true relationship between our dependent variable (regional inequality) and explanatory variable (GDP per capita growth) is assumed to be linear in the parameters, plus an additive error term. In layman's terms: we assume the data can be described by a straight line, plus some random noise. If this assumption is wrong, our line will systematically miss the pattern, for instance, if the real relationship *curves* upward or *flattens* at higher GDP growth levels. Potential fix: Try transforming variables (log, square root) or adding non-linear terms.
2. **Sample Variation:** There must be actual variation in the explanatory variable(s). If all regions had identical GDP growth (the same X), we couldn't estimate a slope. In our data: The growth rates clearly vary between regions, so this assumption holds. If violated: OLS can't separate effects or compute a meaningful slope. Fix: Gather more varied data or remove redundant variables.
3. **Random Sampling**: The observations are assumed to be randomly drawn from the population. This ensures that each region represents the broader population of NUTS2 regions without systematic bias. Our dataset comes from Eurostat, which covers all regions of the selected countries, so it's not truly random but close enough to the full "population" to treat as representative. If violated: The sample might systematically over- or under-represent certain types of regions, leading to biased results.
4. **Exogeneity:** The most important assumption! It means that the error term — the "unexplained stuff" — has an average value of zero and is unrelated to our GDP growth. In other words, growth shouldn't be correlated with hidden factors that also influence (confound) inequality, like education levels, industry mix, or policy differences. If violated: The slope estimate becomes biased because it accidentally captures the influence of those hidden factors. Fix: Add instrumental variables that isolate exogenous variation and run a 2SLS-test.
5. **Homoskedasticity**: The error term has the same variance for each value of the independent variable. We tested for heteroskedasticity in Table 3

, so we can rest assured that this assumption holds for our data. If this assumption had been violated and we could conclude with heteroskedasticity, we would have had unreliable standard errors and P-values.

6. **Normality of the Error Term** The error term is assumed to follow a normal distribution with mean zero. This assumption mainly matters for small samples, as the assumption of normality can be replaced by a large sample size (as defined in the cental limit theorem). It ensures that t-tests and confidence intervals work as expected. In our data: With around 50 NUTS2-regions (aggregated from 476 NUTS3-regions), slight deviations from normality are not a big problem here.

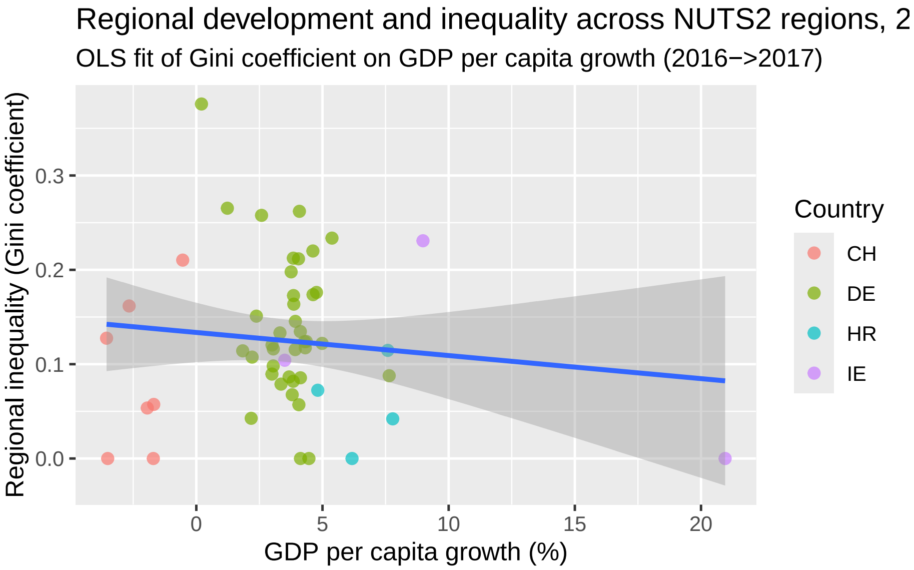


Figure 1: Regional GDP Inequality as a function of GDP

Figure 1 plots the relationship between GDP per capita growth and regional inequality across NUTS2 regions in 2017. Each point represents one NUTS2-region. The fitted regression line shows a slight negative slope, suggesting that regions with faster economic growth tended to have marginally lower inequality, although the relationship is weak and statistically insignificant.

2 Part B: Exploring Other Determinants of Inequity

In this part of the assignment we will gather data on three new variables that may have a influence on the regional inequity for the year 2017 and justify them. We will the use the Gini-coefficient as the dependent variable with the three new variables to create a multiple linear regression and document on the findings. The last part of B is to interpretate the estimated coefficients and to discuss their magnitude, direction ad statistical significant. Then reflect on the overall model and how the variables explaining the variations in regional inequality.

2.1 Our chosen variables

In part B we have chosen to test for the effect of three additional variables: available labour force, unemployment and population density. We will look into each of these variables, and see how they could be determinant factors for regional inequality. Like before, these datasets have been fetched from Eurostat.

- **Labour force** is the variable which includes all persons how are both employed and unemployed. It represents the total supply of labour that's available in the regions, represented in our dataset as "thousand workers".
- **Unemployment rate** can be seen as a sub sector from labour force, as these are persons without work, available for work and are seeking work. The variable is shown as a percentage of the total labour force per region. With this variable we may look into if the The unemployment rate has an effect on the income of those in the region, which affects GDP, which will then have a effect on the regional inequality.
- **Population density** is given as "persons per square kilometer", and indicates how populated a region is, divided by it's total area. A higher population density can indicate a higher degree of urbanization of a region.

```
New names:
* `TIME` -> `TIME...1`
* `TIME` -> `TIME...2`
```

Geo_Codes	Geo_Labels	2000	2001	2002
character	character	character	character	charac
DE11	Stuttgart	1931.5	1945.3	1972.2
DE12	Karlsruhe	1270.3	1293.9000000000001	1311.1
DE13	Freiburg	1011.8	1025.0999999999999	1054
DE14	Tübingen	869.7	870.6	881.6
DE21	Oberbayern	2091.6	2131.9	2127.1
DE22	Niederbayern	580.1	592.20000000000005	600.79
DE23	Oberpfalz	525.70000000000005	536.4	529.79
DE24	Oberfranken	536.5	538.29999999999995	543.6
DE25	Mittelfranken	844.5	842.8	832.2
DE26	Unterfranken	634.20000000000005	640.4	653.29

n: 55

```
New names:
* `TIME` -> `TIME...1`
* `TIME` -> `TIME...2`
```

Geo_Codes	Geo_Labels	2013	2014	2015	2016	201
character	character	character	character	character	character	cha
DE11	Stuttgart	3.7	3.1	3.4	3.3	3
DE12	Karlsruhe	3.6	3.5	3.3	3.1	3.3
DE13	Freiburg	2.9	3	2.5	3	2.9
DE14	Tübingen	2.9	2.7	3	2.6	2.2
DE21	Oberbayern	2.5	2.5	2.7	2.4	2.2
DE22	Niederbayern	3.2	2.9	2.5	2.1	2.1
DE23	Oberpfalz	3.4	2.7	2.8	2.9	2.2
DE24	Oberfranken	3.9	4	3.9	3.3	3
DE25	Mittelfranken	3.1	3.1	3	2.5	2.1
DE26	Unterfranken	3.3	2.9	3	2.5	2.2

n: 56

```
New names:
* `TIME` -> `TIME...1`
* `TIME` -> `TIME...2`
```

Geo_Codes	Geo_Labels	2012	2013	2014	2015	201
character	character	character	character	character	character	cha
DE11	Stuttgart	372.8	375.1	378.0	382.6	387
DE12	Karlsruhe	387.2	389.5	392.1	396.4	403
DE13	Freiburg	230.5	231.7	233.4	236.0	241
DE14	Tübingen	198.4	199.2	200.5	202.9	210
DE21	Oberbayern	250.4	253.4	256.4	259.8	268
DE22	Niederbayern	114.1	114.8	115.5	116.7	118
DE23	Oberpfalz	110.8	111.1	111.5	112.2	113
DE24	Oberfranken	146.6	146.2	146.0	146.2	147
DE25	Mittelfranken	233.7	235.1	236.3	238.4	241
DE26	Unterfranken	152.2	152.1	152.2	152.7	153

n: 51

Geo_Codes	Geo_Labels	2012	2013	2014	2015	20
character	character	character	character	character	character	cha

Table 7

NUTS2	Geo_Labels	Time	Pop_km2	Workforce	Unemp_prct
character	character	character	numeric	numeric	numeric
	Stuttgart	2012	372.8	2,056.0	
DE11	Stuttgart	2013	375.1	2,101.3	3.7
DE11	Stuttgart	2014	378.0	2,148.0	3.1
DE11	Stuttgart	2015	382.6	2,195.3	3.4
DE11	Stuttgart	2016	387.1	2,253.4	3.3
DE11	Stuttgart	2017	389.8	2,274.0	3.0
DE11	Stuttgart	2018	392.0	2,290.7	2.3
DE11	Stuttgart	2019	393.3	2,319.9	2.4
DE11	Stuttgart	2020	393.3	2,280.1	3.5
DE11	Stuttgart	2021	393.7	2,206.1	3.1
n: 612					

A cleaner dataset

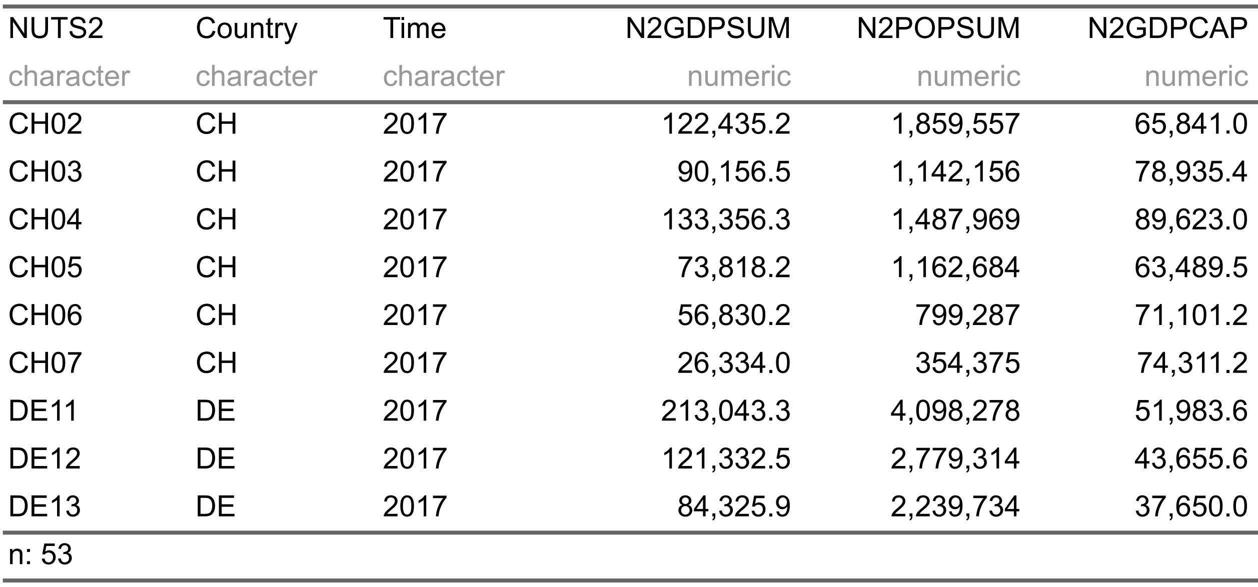
2.2 Multiple Regression Model

We will now test our new variables to see if they have more of an effect on regional inequality. To do this, we will be making a multiple regression model where labour force, unemployment and population density serve as explanatory variables. We have specified the following multiple linear regression model:

$$\text{Gini}_{i,2017} = \alpha + \beta_1\,\text{Workforce}_i + \beta_2\,\text{PopDensity}_i + \beta_3\,\text{Unemployment}_i + u_i$$

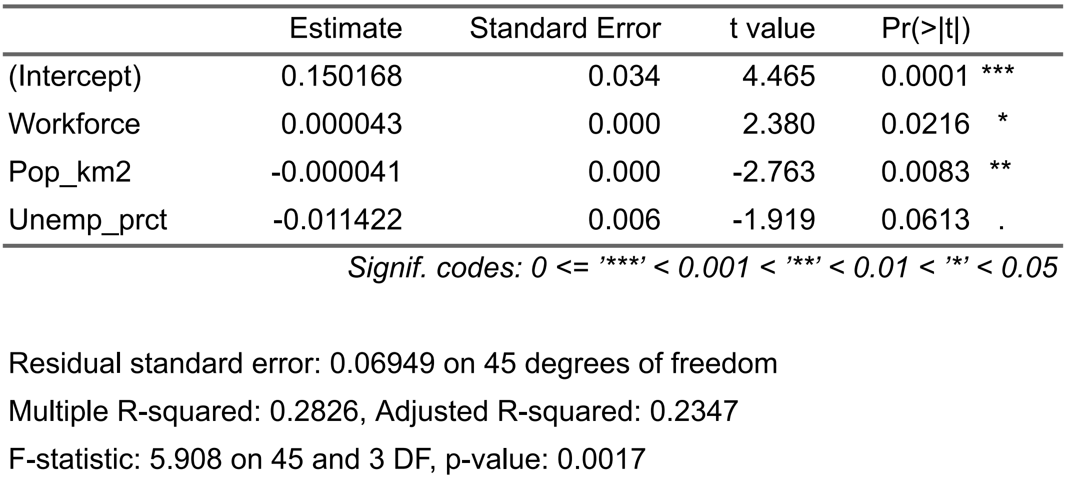
where α is the intercept, β_1, β_2, and β_3 are the slope coefficients for each explanatory variable, and u_i is the error term capturing unobserved factors affecting inequality.

NUTS2	Country	Time	N2GDPSUM	N2POPSUM	N2GDPCAP
character	character	character	numeric	numeric	numeric
CH01	CH	2017	112,845.8	1,613,522	69,937.6



NUTS2	Country	Time	N2GDPSUM	N2POPSUM	N2GDPCAP
character	character	character	numeric	numeric	numeric
CH02	CH	2017	122,435.2	1,859,557	65,841.0
CH03	CH	2017	90,156.5	1,142,156	78,935.4
CH04	CH	2017	133,356.3	1,487,969	89,623.0
CH05	CH	2017	73,818.2	1,162,684	63,489.5
CH06	CH	2017	56,830.2	799,287	71,101.2
CH07	CH	2017	26,334.0	354,375	74,311.2
DE11	DE	2017	213,043.3	4,098,278	51,983.6
DE12	DE	2017	121,332.5	2,779,314	43,655.6
DE13	DE	2017	84,325.9	2,239,734	37,650.0
n: 53					

Table 9



	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.150168	0.034	4.465	0.0001	***
Workforce	0.000043	0.000	2.380	0.0216	*
Pop_km2	-0.000041	0.000	-2.763	0.0083	**
Unemp_prct	-0.011422	0.006	-1.919	0.0613	.

*Signif. codes: 0 <= ’***’ < 0.001 < ’**’ < 0.01 < ’*’ < 0.05*

Residual standard error: 0.06949 on 45 degrees of freedom

Multiple R-squared: 0.2826, Adjusted R-squared: 0.2347

F-statistic: 5.908 on 45 and 3 DF, p-value: 0.0017

Multiple Regression - Regional GDP Inequality as a function of labour, unemploment and population density.

2.3 Interpreting our results

We can see in Table 9 that the model has an R^2 of 0,283 and an adjusted R^2 of 0,235, meaning that approximately 28 % of the variation in regional inequality can be explained by these three predictors. We will go over each of the variables below.

Workforce: The slope for Workforce (= 0,000043, p = 0,0216) is statistically significant at the 5 % level. Holding population density and unemployment constant, an increase of a thousand workers is associated with a 0,000043 rise in the regional Gini coefficient. Economically, regions with larger or more active labour markets tend to experience slightly higher income dispersion, consistent with the idea that economic agglomeration creates greater occupational and wage diversity.

Population density: Population density (= −0,000041, p = 0,0083) is statistically significant at the 1 % level and has a negative sign. This implies that more densely populated regions tend to have lower inequality. A possible interpretation is that highly urbanised regions benefit from better access to education, more specialized jobs or broader labour-market integration that reduce within-region disparities.

Unemployment: The coefficient for Unemployment (= −0,0114, p = 0,0613) is marginally significant at the 10 % level. The negative sign suggests that regions with higher unemployment rates have slightly lower Gini coefficient. However, given the weak statistical significance, this relationship should be interpreted with caution.

Model assessment: The improvement in explanatory power compared to the first regression model shows that labour-market and demographic characteristics are stronger predictors of regional inequality than just short-term GDP growth alone. The aligns with parts of the regional-development theory discussed by Lessmann & Seidel (2017): agglomeration and workforce expansion can increase income dispersion, while institutional factors within densely populated areas may counteract inequality through redistribution and accessibility.

3 Part C: Use of AI

In this assignment we have used AI to ask controlling questions and constructive judging of the text and the codes used, including interpreting the results from our models. The language models used were ChatGPT 3.5 and chatGPT 5.0. All code has been verified by humans, ethically sourced, and has not been tested on animals.

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