# Econometrics Project

## Early Leavers from Education and Their Impact on Productivity

### Professor Luca Stanca

#### Tutor Caterina Morelli

##### Student: Polyzoni Chrysanthi

##### Matriculation number: 911256

2025

# Preface

This study stems from my curiosity following a 180-hour internship at the “Istituto di Istruzione Superiore Luigi Galvani” in Milan, with the intention to pursue a career in education. During my time there, I observed a concerning lack of desire among young people to be part of a larger purpose, with teachers actively trying to find solutions to encourage this sense of belonging. Additionally, through sociology lectures by a professor from Bicocca University, I became increasingly aware of the crucial role education plays in the mental health and well-being of adolescents and young adults. However, I was also curious to explore the economic impact of young people leaving education early in my country. As I near the point where I will retain my Italian citizenship and voting rights—after nearly two decades in this country and raising two wonderful daughters—I feel it's important to better understand these issues from a broader perspective.

The school is inclusive, allowing access to all students and attracting those from distant areas, though it faces challenges like disciplinary issues related to managing diversity and social difficulties. Some behaviors, though not serious crimes, require supervision from the administration and teachers. Despite this, the school promotes cultural diversity and aims for cooperation and integration.

The internship took place across four classes in the applied sciences high school, primarily focusing on the first, fourth, and fifth-year students. Various dynamics were observed, with some classes encouraging competition, while others promoted cooperation. Many students struggle with issues such as substance abuse and express concerns about the shortage of teachers and inadequate laboratory facilities. Additional time was spent in the technical applied school, where some students pursue technical education to follow in their parents' footsteps, aiming for practical training that will lead to employment. However, they often feel demotivated due to the lack of modern facilities.

Students in technical schools often lack enough practical activities, which reduces motivation. Furthermore, with the rise of AI, there is concern about the decreasing value of individual effort, as answers can be easily obtained from external sources.

# Introduction

Very few observations (21) have been collected for the purpose of this research, primarily due to the limited scope and available data. This study is intended to provide an initial exploration into the relationship between early school leavers and labor productivity. While the sample size is small, it serves as a starting point for future research in this area. The intention and hope is that this preliminary analysis will inspire further investigation, leading to more robust findings and a deeper understanding of the impact of early education exits on economic performance.

In this study, we aim to test whether early school leavers from education and training is a significant variable for predicting labor productivity. The data, sourced from the Joint Research Centre (JRC), found at https://urban.jrc.ec.europa.eu/ardeco includes several important economic indicators, such as real labor productivity, total employment, and gross fixed capital formation, which can help explain labor market dynamics and economic performance.

Early school leavers, who exit education and training programs prematurely, often face challenges when entering the labor market without adequate skills. This can impact their productivity levels in the workforce. By analyzing this relationship, we seek to determine if the rate of early leavers from education is a statistically significant factor in predicting labor productivity.

Our hypothesis is that a higher percentage of early school leavers may negatively affect labor productivity, as individuals without sufficient educational qualifications may struggle to perform effectively in high-skill job roles. Conversely, a well-educated and adequately trained workforce is typically associated with higher productivity, innovation, and economic growth.

This analysis uses explores the correct regression to examine the significance of early school leavers as a predictor of labor productivity, while controlling for other potential economic variables. The findings of this study could provide valuable insights into the importance of education policies and training programs for improving productivity in the labor market.

Research Question:

Is the rate of early leavers from education and training a significant predictor of real labor productivity?

By testing this relationship, we can assess the role of education in enhancing workforce productivity and contribute to discussions on policy interventions aimed at reducing early school leaving and improving educational outcomes.

# Tools and Analytical approaches

To carry out the analysis, both Stata and Python were used, combining the strengths of both tools for statistical modeling and data analysis. Stata is a robust software designed for econometric analyses, offering a rich set of commands for fitting models. This allows us to model the relationship between early school leavers from education and training and labor productivity while controlling for other variables, ensuring that the findings are robust.

Python, on the other hand, is used for data manipulation, visualization, and additional analysis. Specifically, Jupyter Notebooks is employed for implementing Python scripts in an interactive environment, which enhances both the clarity and reproducibility of the analysis. Jupyter Notebooks provide an excellent platform for performing exploratory data analysis, visualizing relationships between variables, and running statistical tests. The ability to combine Python code, textual explanations, and visualizations in a single document makes Jupyter Notebooks an ideal tool for collaborative work and sharing insights.

In this project, Jupyter serves as the environment where we load the dataset, preprocess the data, perform statistical analysis, and generate meaningful visualizations, such as boxplots and distribution plots, to explore the data and check the model’s assumptions. By leveraging Python's powerful libraries (such as pandas for data manipulation, statsmodels for regression modeling, and matplotlib or seaborn for visualization), we complement Stata's econometric capabilities and ensure that the project is both thorough and reproducible. This integration of tools provides flexibility and efficiency in handling data, running models, and presenting findings.

# Variables and Exploratory Data Analysis

The ARDECO database, maintained by the European Commission's Directorate General for Regional and Urban Policy, collects long-term data on various economic and demographic indicators for EU regions and some regions in European Free Trade Association and candidate countries. The database includes variables in the following areas:

1. Population
2. Domestic Product
3. Employment
4. Labour Cost
5. Labour Productivity
6. Capital Formation
7. Capital Stock
8. Households

These indicators are categorized by statistical scales (NUTS1, NUTS2, NUTS3, and metropolitan regions) and are updated biannually. The data is sourced from Eurostat, as well as other national and international authorities. For each indicator, a single variable has been selected on an arbitrary basis. Below is a summary of each variable and its role in analyzing economic and labor trends.

All 9 variables selected for our modeling are shown in Figure 1 summarized in normalized boxplots from 0 to 1. We are rescaling the data so that all values fall within the 0, 1 range before plotting the boxplots, using min-max normalization.

A group of graphs showing different types of data

AI-generated content may be incorrect.

A graph showing different colored rectangular shapes

AI-generated content may be incorrect.

Figure

Figure 2 presents the Kernel Density Estimation (KDE) plots, which depict the distribution of variables in a smooth and continuous manner, similar to a smoothed histogram. For the normalized KDE plots, each dataset is first rescaled to a common range between 0 and 1 using min-max normalization. In this process, the minimum value is mapped to 0, the maximum to 1, and all other values are proportionally scaled within this range. It is important to note that the data does not follow a normal distribution.

A graph of colored lines

AI-generated content may be incorrect.

A group of blue lines

AI-generated content may be incorrect.

Figure

The correlation matrix in Figure 3 displays the pairwise linear relationships between the 9 variables in the dataset. Each cell in the matrix contains a correlation coefficient (typically Pearson’s r), which ranges from -1 to 1. A value close to 1 indicates a strong positive correlation, meaning that as one variable increases, the other tends to increase as well. A value close to -1 indicates a strong negative correlation, meaning that as one variable increases, the other tends to decrease. A value around 0 suggests no linear correlation between the two variables. The diagonal elements of the matrix are always 1, as each variable is perfectly correlated with itself.

A chart with different colored squares

AI-generated content may be incorrect.

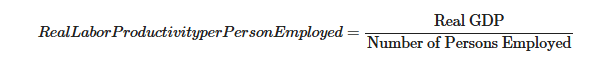
Figure

Several variables are strongly negatively correlated with Productivity (correlation < -0.6). For instance, Population (-0.84) shows that larger populations tend to be associated with lower productivity. Similarly, Early Leavers (-0.77) indicates that a higher proportion of students leaving education early correlates with reduced productivity. Compensation (-0.72) also shows a negative relationship, suggesting that higher compensation levels may be linked to lower productivity, potentially due to structural or labor market dynamics. Both GDP (-0.69) and Capital Stock (-0.66) further suggest that larger economies or higher accumulated capital do not necessarily translate into higher productivity per worker. The only variable with a moderately positive correlation is Taxes on Income and Wealth (0.51), where higher taxation appears to be mildly associated with increased productivity—possibly reflecting redistributive policies in more productive regions.

The variables most likely to influence Productivity appear to be demographic and structural economic factors, as indicated by the strong negative correlations observed with Population, Early School Leaving, GDP, and Capital Stock.

1. **SOVGDE (Productivity)**: Real Labor Productivity per Person Employed

SOVGDE variable represents the economic output (GDP) per worker employed. It is often used as a measure of economic efficiency and productivity. Higher productivity generally indicates that workers are producing more goods and services per unit of input, which is a key indicator of economic health and competitiveness. It reflects how effectively labor is being utilized in the production process, helping to assess overall economic performance.



1. **SNETD (Employment)**: Total Employment (Workplace Based, Employed Persons)

SNETD variable refers to the total number of people employed in the economy, measured at the workplace. This includes both full-time and part-time workers, reflecting the size of the employed labor force. This is an important indicator of the health of the labor market. A higher number of employed individuals typically signifies a stronger labor market and better economic conditions, whereas a lower number may indicate issues like high unemployment or underemployment.

1. **SNPTD (Population)**: Average Annual Population

SNPTD variable represents the average number of people in the population over the course of a year. This is typically used to calculate demographic trends and understand changes in population size over time. Population size is a critical factor in assessing labor supply, economic demand, and overall market size. Larger populations often require more jobs, services, and infrastructure, influencing economic planning and policies.

This dataset provides the annual average population based on Regional Accounts. It is used to compute indicators "per capita".

Methodology: The methods for the compilation of the annual average population differ between countries because of the varying availability of population data sources. The compilation of the annual average population may be based on any of the following methods:

12 monthly averages; 4 quarterly averages; The average of the population on 1st January of two consecutive years; A mid-year estimates. As the time-series for "total population on 1st January" (SNPTN) is rather accurate, the method applied in ARDECO to compute the annual average population is based on the average of the population on 1st January of two consecutive years, by applying the following formula:

A mathematical equation with a smiley face

AI-generated content may be incorrect.

1. **RUWCD (Compensation)**: Compensation of Employees at Current Prices

RUWCD refers to the total compensation paid to employees, including wages, salaries, and benefits, at current prices. It measures how much is spent on labor in the economy. It’s an indicator that helps assess the income distribution within the economy and provides insight into labor costs. A higher compensation of employees generally indicates better wages and employment conditions, while lower compensation could suggest lower labor costs or lower wage standards.

1. **RUYNH (Taxes on Wealth)**: Current Taxes on Income and Wealth

RUYNH variable represents the amount of taxes collected on income and wealth within the economy. It includes income taxes, wealth taxes, and taxes on property. Taxes on income and wealth play a critical role in government revenue, and this figure helps to assess the fiscal health of a nation. It also provides insights into the tax burden on individuals and businesses, which can affect economic decisions such as labor supply, savings, and investment.

Net property income in the context of households refers to the difference between the total revenue generated from property (such as rental income) and the associated operating expenses, which include maintenance costs, property taxes, and other property-related expenses. This income also encompasses, among other things, interest earned by households on their financial investments (after subtracting interest paid on loans), dividends, withdrawals from the income of "quasi-corporations," investment income from insurance or pension entitlements, and rental income from land. Net property income is part of primary income, alongside other components such as compensation of employees, net operating surplus, and mixed income.

1. **RUIGT (Capital Formation)**: Gross Fixed Capital Formation at Current Prices

RUIGT refers to the total value of investments in physical assets (like buildings, machinery, and equipment) within an economy, without adjusting for inflation. Gross fixed capital formation is an important measure of investment in an economy. It reflects the level of economic growth and development, as businesses and governments invest in infrastructure, production capabilities, and innovation. High levels of investment often signal a growing and expanding economy.

1. **ROKND (Capital Stock)**: Capital Stock at Constant Prices

ROKND measures the total value of physical capital (buildings, machinery, equipment, etc.) in the economy, adjusted for inflation (constant prices). Capital stock is crucial for understanding the capacity of an economy to produce goods and services. Over time, increases in capital stock indicate growth in production capabilities, which can lead to higher economic output and productivity.

Capital Stock at constant prices refers to the total value of all the physical assets (like machinery, buildings, equipment, etc.) that are used in the production of goods and services, adjusted for inflation to reflect the true value over time. This concept is important in economic analysis because it allows for the comparison of capital stock across different years without the distortion caused by changes in price levels.

When measuring capital stock over time, inflation can make it seem like the economy has more capital than it actually does. Capital stock at constant prices removes this effect, providing a more accurate picture of the actual changes in the amount of productive capital. Economists and policymakers use capital stock at constant prices to compare the growth or decline of capital stock in an economy over time, without the distortion caused by price changes. This helps to understand whether there’s an actual increase in the economy's productive capacity or if the apparent increase is just due to inflation. Companies and governments can use this data to understand how their investments in physical capital are evolving. For example, if the capital stock at constant prices is increasing, it may indicate that investment in new assets is outpacing inflation.

A white background with black text

AI-generated content may be incorrect.

1. **RPDNN (Early Leavers)**: Early Leavers from Education and Training

RPDNN variable represents the number of individuals who leave education or training programs before completion, typically measured as a percentage of the population of a certain age group.

Early school leavers are often considered a challenge in terms of labor market outcomes, as they may have lower qualifications, limiting their employability and earnings potential. A high rate of early leavers can be a sign of weaknesses in the education system or social barriers to completing education, which may lead to long-term economic challenges such as higher unemployment or lower productivity.

# Data cleaning

The dataset used for this project spans several decades, starting from 1980. However, when cleaning and preparing the data, we encountered a key challenge: the variable early leavers from education and training (RPDNN), which is critical to our analysis, only had reliable data starting from the year 2000. This issue arose because data on early leavers was either incomplete or unavailable for the years prior to 2000. As a result, to ensure the quality and accuracy of our analysis, we made the decision to focus on the period from 2000 to 2020.

We first identified the key variables for our analysis, including SOVGDE (real labor productivity), RPDNN (early leavers from education and training), and other economic indicators such as RUWCD (compensation of employees) and RUIGT (gross fixed capital formation). It was important to retain these variables to ensure a comprehensive model.

We then examined the data for missing or incomplete values. For the years before 2000, there were significant gaps in the data for RPDNN, making it unsuitable for analysis. As a result, these years were excluded from the dataset, as using incomplete data could potentially distort the results of the model.

To focus on the period where the data was complete and reliable, we filtered the dataset to include only the years 2000 to 2020. This subset of data was considered sufficient for our analysis, ensuring that we had a consistent and reliable set of observations for the variable RPDNN.

After filtering the data, we performed additional checks for outliers and anomalies in the relevant variables, especially in SOVGDE and RPDNN, which could potentially influence the results of the regression model. Any outliers were either investigated further or removed if deemed to be erroneous or extreme.

Lastly, we ensured that all variables were formatted correctly for analysis. For example, we made sure that the RPDNN variable was in percentage form and that time series data, such as SOVGDE, was aligned with the correct year range (2000 to 2020). Additionally, any variables with non-numeric entries or inconsistencies were cleaned and standardized.

After the data cleaning process, we ended up with a more focused dataset that included only the years 2000 to 2020. This allowed us to proceed with a more robust and reliable analysis, ensuring that the key variable of interest (RPDNN) had valid data for the entire period under study. The resulting dataset was now ready for statistical modeling and hypothesis testing regarding the relationship between early school leavers and labor productivity.

Finally, cleaning process helped us address the issue of incomplete data for RPDNN prior to 2000, and by focusing on the period from 2000 to 2020, we were able to conduct a more accurate and meaningful analysis of the impact of early education leavers on labor productivity.

# Time Series

In this study, we primarily explored the relationship between early leavers from education (RPDNN) and labor productivity (SOVGDE), but a crucial aspect of time series analysis that should be further examined is serial correlation of the errors and the potential shocks affecting the system.

Serial correlation refers to the correlation of a variable with itself over successive time intervals. In time series analysis, serial correlation of residuals (errors) occurs when the error terms from one period are correlated with those of another period. This can indicate that there are patterns in the data that have not been accounted for by the model—potentially due to omitted variables, lags, or shocks.

For example, if the residuals from a regression model are not random, but instead show a pattern (e.g., they are positively or negatively correlated with past residuals), it suggests that the model might be missing something important, such as a lagged variable or a shock event. These correlations in the errors could be interpreted as a form of shock or unexplained dynamics in the data that require further exploration.

In the context of time series analysis, shocks refer to sudden, unforeseen changes that disrupt the regular behavior of the system. These could be economic disruptions, policy changes, or any major event that alters the normal course of trends. Shocks typically introduce serial correlation in the error terms because their effects do not immediately dissipate and may persist over several periods.

In the model we're working with, if serial correlation is present in the residuals, it suggests that there are shocks (unmodeled events) that are affecting the relationship between early leavers from education (RPDNN) and labor productivity (SOVGDE). These shocks might require adjustments to the model to properly account for their impact, often using techniques such as ARIMA models (Auto-Regressive Integrated Moving Average) or other time series methods that handle serial correlation and incorporate shocks.

A graph of a graph

AI-generated content may be incorrect.

# Modeling

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.