Exploring Factors Impacting Innovation Rates: an analysis on OECD Countries (2001-2015)

By Elia Landini
Tra My Le
Veronica Manusardi

Panthéon-Sorbonne Master in Economics
Course: Econometrics

TABLE OF CONTENT

1. INTRODUCTION 1
2. LITERATURE REVIEW 3
3. DATA ANALYSIS 7
4. EMPIRICAL APPROACH 12
5. RESULTS 16
6. CONCLUDING REMARKS 21
7. REFERENCES 22

---

1. INTRODUCTION

America is the land of possibilities and personal growth, under many aspects. The American dream, the idea that through hard work, courage and determination it is possible to achieve an improved quality of life and economic prosperity, is what pushed people to migrate in the US since the late 1880s. The US has soon become a symbol of entrepreneurial opportunities for people coming from different parts of the globe and it started to be seen as the ideal place to cultivate ideas and transform them into revolutionary inventions - from Edison's light bulb in 1878 till the invention of the internet and the *www*.

However, behind this captivating and fascinating vision, access to opportunities to become an inventor is not uniformly distributed, as there are challenges and social characteristics that may hinder the path for some.

There are indeed various factors, but those who have the strongest influence on the possibility of becoming and inventor are especially the socio-economic ones, as highlighted by Raj Chetty in his research paper "*Who becomes an inventor in America? The importance of exposure to innovation",* published in 2018 on The Magazine for Economic Performance.

The paper explores the main socio-economic barriers that some individual face when pursuing a innovation career, concluding that what is making individual miss the opportunity to become an inventor is their characteristic at birth, so gender, ethnicity, the social condition of their families and the environment they were born in.

The author coined the expression of "*Lost Einsteins",* referring to all those people who could have made highly influential innovations if they had been provided with the opportunities they rightly earned, but that could not due their characteristics at birth.
In particular, the definition refers to students who are part of *social minorities,* such as women, ethnic minorities and low-income families, who due to misallocation and the limited resources of their families and the lack of a strong mentorship in the early stages of education, missed the possibility to delve into STEM educational fields and develop skills that may push them in creating something and becoming an inventor.
However, Chetty's investigation was focused exclusively on the United States, mainly because he argued that relatively little was known about the individuals who become inventors in the modern era in the United States and he decided to delve more into this geographical area.
However, his research *opened us a door* to a crucial question: does the American dream of limitless opportunities reflect reality or do these socio-economic barriers cast shadows on this ideal? Behind the façade of the United States as the land of opportunities, we must consider whether certain social boundaries exert such a significant influence that alternative setting may prove more conducive to inventive pursuits.
This question has inspired our research, aimed at exploring whether the United States is indeed the place where anything is possible for aspiring inventors, even considering the different social obstacles that could influence the path to innovative success.
Through this inquiry, we seek to understand if there are other countries or social environments where the dream of becoming an inventor is more easily achievable than in the United States, narrowing our research exclusively on a selected group, countries within the Organisation for Economic Co-operation and Development (OECD)[1], in order to ensure fair comparisons and draw meaningful insights specific to advanced economies.
Additionally, our research considered data from 2000 to 2015 and thought this is not an extended period of time, it was the base chosen for our research to avoid exogenous effects quite hard to be captured by our model. An example of it, imagining analysing data from the '70s, regardless their actual availability, we could have to capture also various generational changes that plays a determinant role in encouraging individual to apply for a patent, and perhaps do it in a specific country. Among them, more binding and widespread intellectual property laws, the strong diffusion of English as international language, or even the epiphany of social media. Hence, we consider these macro-trends as ceteris paribus over here, notwithstanding each country has its own factors that either

[1] To this date, OECD countries are 38, for the list of members visit: https://www.oecd.org/about/members-and-partners/

help or hinder innovation, for example economic structures, education systems, regulations and cultural attitudes. Therefore, we considered some of these factors there appeared to be common in the countries we analysed, such as GDP, trade openness, employment rate, gender gap and government expenditure (in R&D, education or healthcare).
To unravel the complexities of innovation dynamics and in order to understand those factors that set countries apart in terms of innovation, we have built a regression model with the country's rate of innovation as the dependent variable (y) and considering various independent variables representing representing economic, social, and institutional factors.
Our research aims to provide a contribution to the debate on innovation opportunities in an increasingly interconnected and competitive world, not only to highlight where is actually easier to become an inventor, but also to provide proof of those factors that have a high influence on this event.

2. LITERATURE REVIEW

Firstly, to understand this research, we consider that it is essential to recognize who the authors and literature consider as an inventor. An inventor is an individual that holds a patent, that gives exclusive rights to sell or profit from an invention and protects their intellectual property. The patent is the only element granting the right to exclude anyone else from the re-production or use of a specific new invention for a stated number of years (an average of 20 years) and to be such, it has to present three main characteristics: novelty (not previously published elsewhere in the world), non-obviousness (sufficient inventive step) and industrial applicability (commercial utility).
Even though the main role of a patent is to protect the intellectual property of the inventor for him to draw economic advantage from it[2], a patent promotes innovation and technical progress firstly by providing a temporary monopoly for the inventor and secondly, by generating new ideas and encouraging new inventions.
However, the advantages of a patent and their characteristics are not related solely on the inventor, on a micro level, but they must be considered on a macro level, involving countries, firms and a globe wide economy in order to fully comprehend their impact.
Literature has debated on the importance of patents for a country[3], examining their role in shaping market dynamics and their impact on market competitiveness. Patents not only show us a nation's

---

[2] Griliches, Zvi. "Patent statistics as economic indicators: a survey." R&D and productivity: the econometric evidence. University of Chicago Press, 1998. 287-343.
[3] Gambardella, Alfonso, Dietmar Harhoff, and Bart Verspagen. "The value of patents." Universita Bocconi, Ludwig-Maximiliens Universitaet, and Eindhoven University, Working Paper: http://www. creiweb. org/activities/sc_conferences/23/papers/gambardella. pdf (2005)

commitment to encourage creativity and protect intellectual property, but they also generate an economic value and represents for a country the ideal mechanism for incentivizing research, competitiveness, and technological progress in a country.

R&D	- patent provides legal protection for innovations - incentive for companies to engage in research and development activities
Tech	- attracts FDI - stimulate domestic investment - contributes to technology evolution
Competitiveness	- Patented technologies can serve as a barrier to entry for competitors - Influence on market dynamics
Market Dynamics	- patenting strategies influence firms and new entrants

Research and Development (R&D)

Investment in R&D encourages innovation, which in turn, spurs economic growth. Indeed, an economy with a higher degree of innovation - meant as original patented products - will tend to be more independent on the market and gain a competitive advantage. As a result, it will strengthen and enhances its connectivity, becoming more appealing to foreign investments. +
Literature has studied the relation existing between the number of patent applications and R&D expenditure and some[4] were able to build a regression model which showed a strong positive correlation between R&D expenditure and patent application, meaning that an increase of R&D expenditure will enlarge the frontier R&D activities, improve the probability of R&D success and eventually increase the number of patent applications. More literature[5] showed that both business and non-business R&D have a positive influence on patenting activity.

Technological progress

As emphasized throughout this study, patents provide inventors with exclusive rights to their invention for a certain timeframe. This exclusivity serves as an appealing incentive for companies to invest in innovation, explore new ideas and their subsequent development, and this incentive extends

[4] Prodan, Igor. "Influence of research and development expenditures on number of patent aplications: selected case studies in OECD countries and central Europe, 1981-2001." Applied Econometrics and International Development 5.4 (2005).
[5] Westmore, B. (2013), "R&D, Patenting and Growth: The Role of Public Policy", OECD Economics Department Working Papers, No. 1047, OECD Publishing, Paris,

especially to foreign investors, making the country more attractive for attracting Foreign Direct Investment (FDI). Nations boasting well-established patent systems tend to be magnets for FDI, as foreign companies are more willing to invest in countries where their intellectual property is protected, as this minimizes the risk of competitors copying or infringing upon their innovations.
The elevate presence of patent plays a role in fostering technology transfer among different companies, industries and countries. This means that a strong patent framework for a country does not only stimulate foreign and domestic investment, but also contributes to the collaborative evolution of technology across various sectors.

Market competitiveness and dynamics

Protecting individuals' intellectual property, patents promote fair market generate a incentive for innovation, research and development, and since patents may lead to significant profits, firms and countries are interested and willing to research and innovate more and more.
Additionally, firms and countries with strong patent protection will be better positioned on the market the operate in and are able to establish a competitive advantage by offering innovative products and services, generating entry barriers for potential competitors.
As stated above, the presence of a well-established patent framework may be appealing not only to domestic and foreign investors, but also to other companies that may be more likely to engage in synergies and joint ventures that can drive innovation and market competitiveness.
Certainly, patents wield a significant impact on the innovation rate of a country. They are undeniable instrumental in shaping a nation's innovation landscape. The correlation between patents and a country's innovation level is unmistakable. Without any doubt, the loss of potential inventors – those *Lost Einsteins* discussed above- due to a weak approach to patent protection can have extremely negative effect on a country's innovation environment. Especially if these individuals are left out because of extreme inequality.
Therefore, having a strong patent system is not just a legal necessity but a strategic move for any country looking to boost and maintain its innovation capabilities.
Raj Chetty's research on "Lost Einsteins has delve into and highlight the main differences among students that hinder them to becoming an inventor, focusing on socio-economic factors such as gender, race (differences between children from black or white families) or growing up in a low-income family or neighborhood. His research has shown what the impact of these factors is and how significant it is, concluding that losing this part of society may have a deep negative impact on US economic growth.

In fact, the critical point faced in his research is that the demographic groups found represent a significant portion of America population, meaning that US society may be missing out on a substantial number of potential inventors. As stated by D. Leonhardt[6], US society appears to be missing out on most potential inventors from minorities group, both ethnic and gender (in the first place, African American and women), and these groups together make up most of the American population.

However, he did not explain if these factors are especially correlated to the US environment or not and so it is not completely clear if they are the only ones that may hinder potential inventors. Our conclusion was that there may be a possibility that the US is not the most suitable place and country to become an inventor.

The rest of the literature that focuses on this topic often revolves around two main points. Firstly, there is a focus on the significance and influence of parental income, particularly father's income, on a child's possibility and likelihood of becoming an inventor and those other influent factors, such as ethnicity and gender. Secondly, more recent literature explores the relation between misallocation and economic growth, examining how the loss of potential inventors withing these marginalized groups could impact the overall welfare of the country.

The impact of innovation on economic growth is undeniable and missing out on a part of it will generate negative impacts on the economy of a country.

The rate of innovation, measured by the introduction of novel ideas, products, or processes, serves as a key indicator of a nation's competitive edge and it is fundamental to revolutionize industries, improve quality of life, and enhance global competitiveness.

However, it still seems difficult to measure the innovative output of a country. The main part of literature has chosen to use R&D spending in a country as a measure of technical change, due to availability and reliability of data, usually using a Cobb-Douglas approach. Studying how overall productivity grows, authors and research either use the amount of R&D capital in a total factor productivity regression or the intensity of R&D in a regression for changes in total factor productivity. These two methods are shown in the equations below.

1. $logTFP_t = logA + \gamma logRDK_t + \beta_t$

   Where RDK is the stock of R&D capital in time t, the parameter γ is knowledge. This equation yields a measure of the elasticity of output with respect to knowledge.

---

6 Lost Einsteins: The Innovations We're Missing, by D. Leonhardt, New York Times, 2017 https://www.nytimes.com/2017/12/03/opinion/lost-einsteins-innovation-inequality.html

2. $dlogTFP_t = \rho \frac{RD_t}{Q_t} + \beta$

Where RD is the flow of R&D in time t, the parameter ρ is return to knowledge. This equation yields a measure of the social gross (excess) rate of return to knowledge.

However, both these two approaches present problems. Firstly, it is not explained that knowledge is separable in the production function and there may be measurement problems.

Therefore, other methods to measure innovation rate exist, even though only a few authors have not considered the key role of R&D in measuring innovation.

For example, authors like Geroski (1989) and Budd and Hobbis (1989) conducted studies involving firms in the UK and their results indicated that patenting by UK firms had a significant and positive impact on productivity.

In any case, innovation and technical change has always consider in literature as the biggest responsible for economic growth of a country, starting from the work of Solow, and overlook the central role of innovation would represent a substantial loss for a country.

However, as Chetty states[7], it is not possible to deduce that American economy would be better if these individuals had the possibility to become inventors, but it is possible to notice the failure in society to consent these individuals to become inventors.

3. DATA ANALYSIS

In this section, we describe the data sources and define the key variables used in our analysis. In total, 11 databases of the 36 OECD countries from 2001 to 2015 are used. They are merged to form a panel dataset that contains 540 observations of per country innovation rate over the mentioned period across the corresponding GDP, GDP growth, Trade openness, FDI, GDP expenditure on R&D and education among other key determinants of innovation.

Data sources

The 2 data sources for the datasets are OECD.Stat, which includes data/ metadata for OECD countries among others and the World Bank Open data.

1. OECD.Stat

10/11 of our variables come from this platform, which consolidates data across OECD's many databases. The dependent variable, the "Regional Innovation rate" (specifically, the PCT patent

[7] Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, John Van Reenen "*Who Becomes an Inventor in America? The Importance of Exposure to Innovation*", The Quarterly Journal of Economics, Volume 134, Issue 2, May 2019, Pages 647–713

applications per million inhabitants) is taken from the "Regional Innovation" dataset, whose metadata collection is undertaken by the Centre for Entrepreneurship, SMEs, Regions and Cities via an annual questionnaire sent to the delegates of the Working Party on Territorial Indicators and via access to websites of National Statistical Offices and Eurostat.

2. The World Bank Open Data

Only one variable, the "GDP yearly growth rate" in OECD countries, is extracted from The World Bank Database, which uses World Bank national accounts data and OECD National Accounts data files. The GDP data represents the sum of value added by all its producers (before accounting for consumption of fixed capital in production) and Growth rates of GDP are calculated using the least squares method and constant price data in the local currency.

Data description

The definitions are highlighted below for the variables (per OECD country) of our research question. Specifically, we discuss those of 'Regional Innovation rate' as the dependent variable (Y_0= i) and 10 remaining independent variables detailed below ($\{X_i\}_1^{10}$).

- *Regional Innovation rate* (PCT patent applications per million inhabitants, Y0=i):

A patent is an exclusive right granted by a national or regional patent office for a limited period. Patent Co-operation Treaty (PCT), an international treaty, administered by the World Intellectual Property Organization (WIPO), enables applicants to seek patent protection simultaneously in each of many countries by filing one international patent application instead of several national or regional applications. The dependent variable, Innovation rate, is defined as PCT patent application data averaged per million inhabitants in a country.

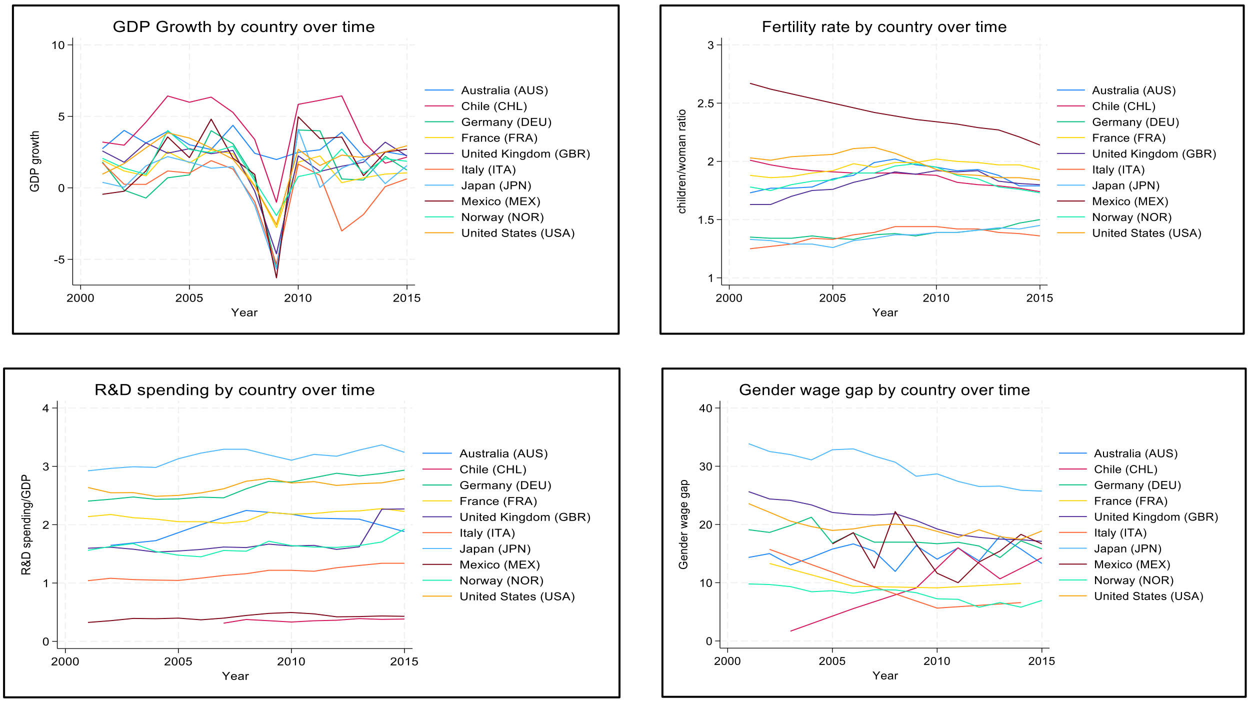
- *Gross Domestic Product* (GDP in million USDs, X1=y):

GDP is the standard measure of the value added created via production of goods and services in a country during a period and the income earned from that production or the total spending on final goods and services (less imports). It is based on nominal GDP (GDP at current prices).

- *GDP Growth* (annual %, X2=g):

GDP growth is the annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2015 prices, expressed in USD.

Graph 1: raw data, variables' trend by country over time



Source: our data elaboration on Stata (not all countries included)

- *Foreign direct Investment* (FDI) (million USDs, X3=s):

FDI flows record cross-border transaction value of direct investment in a period, often a quarter or a year. Financial flows include equity transactions, earnings reinvestment and intercompany debt transactions. Outward flows represent transactions that increase the investment that investors in the reporting economy have in enterprises in a foreign economy, less transactions that reduce such investment.

- *Trade Openness - Trade in Goods and Services* (million USDs, X4=t):

Trade in Goods and Services is the transactions in goods and services between residents and non-residents. It is measured in million USDs at 2015 constant prices and purchasing power for net trade (exports minus imports)

- *Employment Rate* (thousands of persons or % of working age population, X5=e):.

Employment rate is a measure of the extent to which available labor resources are being used, calculated as the ratio of the employed to the working age population. The employed are those aged 15 or over who report having worked in gainful employment for at least one hour in the previous week or who had a job but were absent from work during the reference week. The working age population refers to people aged 15-64. It is seasonally adjusted and measured in 'thousand persons' aged 15 and over.

- *Gender Wage Gap indicator* (X6=w):

Gender wage gap is the difference between median earnings of men and women relative to median earnings of men.

- *Fertility rate* (Total children per woman, X7=f):

Fertility rate in a year is the total number of children that would be born to each woman if she were to live to the end of her child-bearing years and give birth to children in alignment with the prevailing age-specific fertility rates. It is calculated by totaling the age-specific fertility rates as defined over 5-year intervals.

- *Public Spending on R&D* (%GDP) (X8=r):

GDP spending on R&D is the total expenditure on R&D carried out by all resident companies, research institutes, university and government laboratories, etc., in a country, including R&D funded from abroad, excluding domestic funds for R&D outside the domestic economy.

- *Public Spending on Education* (% GDP, X9=u):

Public spending on education includes direct expenditure on educational institutions and educational-related public subsidies given to households and administered by educational institutions. It is shown as a percentage of GDP, divided by primary, primary to post-secondary non-tertiary and tertiary levels. OECD Database on Education Statistics

- *Public spending on healthcare (% GDP, X10=h):*

Health spending measures the final consumption of health care goods and services including personal health care and collective services but excluding spending on investments. It is measured as a share of GDP, as a share of total health spending

Data description

Variable names (per country)	Units	Variables in regression (specified if log was taken)	Frequency	Years of available data	Source
Regional Innovation rate ***(Y0=i)***	PCT patent applications per million inhabitants	Regional innovation rate (Natural logarithm of i)	annually	1990-2015	OECD
Gross domestic product (GDP) ***(X1=y)***	Million USDs	National GDP (Natural logarithm of y)	annually	1970-2022	OECD
GDP Growth ***(X2=g)***	Annual %	GDP yearly growth rate	annually	1961-2022	World Bank
Foreign direct Investment (FDI) ***(X3=s)***	Million USDs	Foreign direct Investment (FDI) (Natural logarithm of s)	quarterly	2005-2022	OECD
Trade Balance (Trade in Goods and Services) ***(X4=t)***	Million USDs	Trade balance (Natural logarithm of t)	quarterly	1995-2022	OECD

Employment Rate ***(X5=e)***	thousands of persons	Employment rate (LFS) (Natural logarithm of e)	annually	1955-2021	OECD
Gender Wage Gap indicator ***(X6=w)***	employees	Gender wage gap indicator **Excluding the self-employed*	annually	1995-2022	OECD
Fertility rate ***(X7=f)***	Total children per woman ratio	Fertility rate (children/woman)	annually	1970-2022	OECD
Public Spending on R&D ***(X8=r)***	% GDP	R&D spending	annually	2000-2022	OECD
Public Spending on Education ***(X9=u)***	% GDP	Educational spending	annually	2000-2020	OECD
Public spending on healthcare ***(X10=h)***	% GDP	Healthcare system spending	annually	1970-2022	OECD

Data Manipulation

Firstly, disaggregated quarterly data was transformed into aggregated annually data for some variables.

Secondly, with regards to missing values, where existed for some variables, we employ a panel random effect estimator for each to generate a prediction on both possible past and future observations. As per Hausman test, this provides better estimation for our unadjusted data. Despite large number of variables, we could only employ those with a complete range of observations over time (otherwise the regression would continue predicting missing results). Notice that for some countries, data is complete for variables not listed in our model but to avoid over-noise-capturing bias which might affect only those countries, we opted for a uniformed prediction model, including those complete variables common to all countries.

Thirdly, in the case of log-transformed variables (specified in the table below), we proceed with retaking the log function for the new predicted values. We hold that doing a 2-stage log transformation makes the regressive forecast more robust to possible heteroskedasticity or high-skewness possible biases.

Next, in the case of log-transformed variables, we proceed with retaking the log function for the new predicted values. We hold that doing a 2-stage log transformation makes the regressive forecast more robust to possible heteroskedasticity or high-skewness possible biases.

At this point, stationarity tests are crucial to validate that statistical properties of mean and variance remain constant over time in our data frame. Ensuring stationarity is vital for accurate modelling, forecasting and reliable statistical inference as non-stationarity data may mislead models and compromise forecasting accuracy. Hence, stationarity verification is essential for maintaining the stability and interpretability of the time series models. To investigate the heterogenous range of non-stationarity biases, we performed 4 tests for each independent variable: Im-Pesaran-Shin (IPS) test, Breitung Panel Unit Root Test, Hadri M test and Levin, Lin and Chu (LLC) test. Where there existed non-consensus result, the test with the most reliable significant statistics would be taken. It is also worth noticing that in our case, traditional approaches such as Augmented Dickey Fuller test are not available given the multitude of panel belonging to the data frame. The results suggested that the variables 'y', 'g', 's', 't' and 'r' are likely non-stationary across the panel.

Given the results in terms of stationarity extrapolated from the employed stationarity test, we moved on to implement corrections to the non-stationary variables, by recurring to first differences, following the rationale of eliminating trends and making the series more amenable to statistical analyses. As a result, we found that the differenced variables *'y', 'log_y', 'g', 's', 't', 'log_t' and 'r'*, appear stationary across panels, supporting the effectiveness of the differencing transformation achieving stationarity.

4. EMPIRICAL APPROACH

Introduction

It is common practice in current literature to present past models in order to give strength to possible assumptions or avoiding specifications already widely described by other sources. However, the current literature on this side was quite meagre, and largely focused on social factors, giving macroeconomic factors a backstage role or sometimes even excluding them.

Hence, in addressing our research question, i.e., how and to what extent macroeconomic, social and government-biased factors affect the innovation rate in a country, we were not really backed from any previous research which could propose similar or feasible models to be apply in our case. Moreover, when it comes to deal with regional innovation there exist factors that are largely affecting individuals on their own micro-sphere, related to both intrinsic characteristics at birth and characteristics developed over time, but that are out of the scope of our research demand[8]. This idea of a macro-level analysis is probably the key added value of our research to the current literature, which can be translated into a radical switch from a literature that investigates individuals' factors leading to holding a patent, to research, ours, that investigates country's factors that influence the chance of having new patent holders on its domestic soil. It is worth to notice that we are mentioning patent holders and not number of patents released, since our dataset refers to a density value in terms of millions of inhabitants of people granted of this right and this leads to the potential risk of having people holding more than one patent.

[8] Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, John Van Reenen "*Who Becomes an Inventor in America? The Importance of Exposure to Innovation*", The Quarterly Journal of Economics, Volume 134, Issue 2, May 2019, Pages 647–713

It is paramount to underline that there exists always potential bias in the estimation of regional innovation rate, since we are aware of omitted variables that are not available, hard to measure and can be related to the economic development and not really to the innovation itself.

Model Specifications

In order to find the most adherent model for our research problem without incurring in both endogeneity and high-multicollinearity issues, we have developed multiple linear panel regression models to estimate the marginal impact of each factor on the regional innovation rate (i). Aside from our dependent variable, independent variables have been clustered in three groups according to their nature and then progressively added to the model to verify their relevance and models' fit (R-squared). But before giving insight into these different model structure, it is mandatory to explicate the choice of adopting a random effect estimator rather than more common methods of estimations for panel data, among them fixed effects or pooled OLS.

This econometric approach, grounded in the random effects model, is chosen to efficiently capture unobserved heterogeneity across countries, recognizing the diverse influences shaping innovation trajectories within the dynamic context of OECD member states. Pooled OLS is appropriate when you treat the panel data as if it were a single cross-sectional dataset, ignoring any individual-specific or time-specific effects. This assumes that there is no correlation between the individual effects and the independent variables, which is not our case. On the other hand, to verify whether random effects assumption was more suitable than the fixed effects one, we conducted the Hausman test. The results of the test indicate that the p-value is considerably high (0.9992), suggesting that you fail to reject the null hypothesis. The null hypothesis in the Hausman test is that the difference in coefficients between the fixed-effects (FE) and random-effects (RE) models is not systematic. Since the p-value is critical, there is no strong evidence to suggest that the differences in coefficients between the two models are systematic. Therefore, we chose the random-effects model as it is consistent with the assumptions of the test.

Table 4: Haussmann test

Variable	FE	RE	Difference	Std. Err.
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
log_y	142.44640	146.30180	-3.85538	-
g	-093209	-0.91844	-0.01365	-
log_s	67132.79	67984.52	-851.72360	-
log_t	-66392.85	-67234.6	84.74470	-
log_e	39.33977	40.10877	-0.76901	-
w	-0.87264	-0.71456	-0.15809	-
f	27.24597	29.78793	-2.54196	2.49745
r	1.99552	1.92523	0.07029	-
u	-626554	-6.06510	-0.20043	-
h	13.66471	14.19723	-0.53252	0.07353

Source: our data elaboration on Stata

The model specifications follow:

- *Macro-trends specification*

$$iit = \beta 0it + \beta 1i \ln(yit) + \beta 2igit + \beta 3i \ln(sit) + \beta 4i \ln(tit) + \beta 5i \ln(eit) + \varepsilon it$$

  where *i* represents regional innovation rate by country over time, *ln(y)* the logarithmic function of country's GDP, g the yearly GDP's growth rate, *ln(s)* the logarithmic function of foreign direct investment (FDI), *ln(t)* the logarithmic function of trade balance (trade openness), *ln(e)* the logarithmic function of employment rate, *εit* is the idiosyncratic error term with the assumption *E(εit) =0.*

- *Female empowerment specification*

$$i_{it} = \beta_{0i} + \beta_{1i} \ln(y_{it}) + \beta_{2i} g_{it} + \beta_{3i} \ln(s_{it}) + \beta_{4i} \ln(t_{it}) + \beta_{5i} \ln(e_{it}) + \beta_{6i} w_{it} + \beta_{7i} f_{it} + \varepsilon_{it}$$

  where *i* represents regional innovation rate by country over time, *ln(y)* the logarithmic function of country's GDP, g the yearly GDP's growth rate, *ln(s)* the logarithmic function of foreign direct investment (FDI), *ln(t)* the logarithmic function of trade balance (trade openness), *ln(e)* the logarithmic function of employment rate, *w* the gender wage gap, *f* the fertility rate, *εit* is the idiosyncratic error term with the assumption *E(εit) =0.*

- *Government expenditure targets specification*

$$i_{it} = \beta_{0i} + \beta_{1i} ln(y_{it}) + \beta_{2i} g_{it} + \beta_{3i} ln(s_{it}) + \beta_{4i} ln(t_{it}) + \beta_{5i} ln(e_{it}) + \beta_{6i} w_{it} + \beta_{7i} f_{it} + \beta_{8i} r_{it} + \beta_{9i} u_{it} + \beta_{10i} h_{it} + \varepsilon_{it}$$

  where *i* represents regional innovation rate by country over time, *ln(y)* the logarithmic function of country's GDP, g the yearly GDP's growth rate, *ln(s)* the logarithmic function of foreign direct investment (FDI), *ln(t)* the logarithmic function of trade balance (trade openness), *ln(e)* the logarithmic function of employment rate, *w* the gender wage gap, *f* the fertility rate, *r* the public expenditure on R&D, *u* the public expenditure for education, *h* the public expenditure for healthcare, *εit* is the idiosyncratic error term assuming *E(εit) =0.*

Once independent variables have been included in the model, we must depurate our dependent variable from its own recursive bias, given by its past values: the autocorrelation test. The analysis undertaken on the variable *i* and its first lag, denoted as *L.i*, discloses a robust positive autocorrelation with a correlation coefficient of approximately 0.9938. This latter underscores a pronounced and consistent linear relationship between the innovation rate (*i*) and its past values averagely for each panel. The strength of this positive autocorrelation signifies that the current level of innovation is strongly influenced by its historical performance, and the nearly perfect positive correlation suggests a persistent pattern in the innovation rate over time, indicating a substantial degree of inertia or momentum in the innovation process. This empirical finding leads us to have evidence of potential influence of past innovation dynamics on present outcomes, emphasizing the relevance of incorporating temporal dependencies in the modelling and analysis of innovation trends. However, including *lagged i* in the model to tackle autocorrelation and temporal dependencies would have bring to over capture the data noise and mislead eventual conclusions. Therefore, we have kept working with our original specifications, without including any lag variables, aware of the autocorrelation of innovation rate.

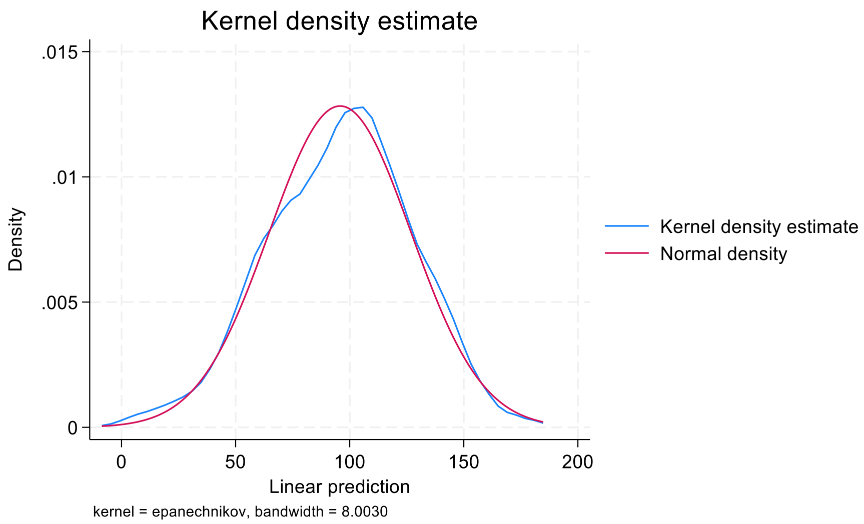
As regards the specification choice, the model fit statistics for the three distinct specifications, namely Specification 1 (Macro-trends), Specification 2 (Female Empowerment), and Specification 3 (Government Expenditure Targets), offer a comprehensive view of their respective explanatory prowess in elucidating variations in the innovation rate (*i*) across the observed countries. While Specification 1 captures an overall R-squared of 0.1311, denoting a modest but notable explanatory power based on macroeconomic indicators, Specification 2, integrating variables that deal with women empowerment, demonstrates a slightly higher R-squared of 0.0718. The inclusion of the wage gap (*w*) and fertility rate (*f*) as significant contributors suggests their relevance in shaping innovation dynamics, however risk of correlation but not causality must be considered. In contrast, Specification 3 (Government Expenditure Targets) stands out with the highest overall R-squared of 0.2877, underscoring the substantial impact of government-related variables such as R&D spending (*r*), education spending (*u*), and healthcare spending (*h*) on innovation rates. Reason why we then moved forward adopting this latter as the better explicative specification.

Moreover, and additional proof of the suitability of our choice was ultimately given by our checks on error terms' normality. For Specification 1, the Jacques-Bera test, Skewness and Kurtosis tests, and the Shapiro-Wilk test consistently reject the null hypothesis, suggesting a departure from normality for the error terms associated with the innovation rate (*i*). Similarly, Specification 2 follows suit, with all normality tests indicating a significant departure from normal distribution. However, in Specification 3, the p-values from the Jacques-Bera test, Skewness and Kurtosis tests, and the

Shapiro-Wilk test are relatively higher. While the Jacques-Bera test's p-value is not notably elevated, the other two tests provide some support for the hypothesis of normality. Therefore, based on these normality tests, the error terms in Specification 3 appear to be relatively close to a normal distribution compared to the other specifications. Together with the better model fit, the relevance of normality in ensuring the validity of statistical inferences, Specification 3 is again chosen as the preferred model, where the error terms demonstrate a comparatively better fit to a normal distribution.

Eventually, remaining on Specification 3, the functional form of the model had to be evaluated to address whether our dependent variable was following a different trend structure with respect to the one we have predicted ex ante. Running the test, the p-value resulted quite small (p = 0.0000), indicating that we reject the null hypothesis, suggesting that there might be a specification error in the model related to quadratic and cubic functional forms, which leads us to rule them out of the model specification. The test confirmed us that chosen functional form fits our data.

Graph 2: Kernel density estimate of se for Specification 3



Source: our data elaboration on Stata

5. RESULTS

Beginning with each specification's marginal effects estimates, notably, the natural logarithm of GDP (*log_y*) exhibits distinctive impacts across specifications, with Specification 2, centred around female empowerment, displaying the highest coefficient (5.152.036), suggesting that when accounting for gender-related dynamics, the positive influence of GDP on innovation is more pronounced. In contrast, Specification 3, emphasizing government expenditure targets, reports the smallest coefficient for *log_y* (1.463.018), which could lead to hold that the varying significance of *log_y* underscores the importance of considering gender-related factors in innovation processes, potentially indicating that fostering female empowerment contributes significantly to innovation, even though we have causality issue must be addressed.

Similarly, the GDP yearly growth rate exhibits nuanced relationships across specifications. Specification 3 stands out with the most substantial negative impact (-0.91844), indicating that higher GDP growth, within the context of government expenditure targets, is associated with a lower innovation rate. This may be attributed to policy choices and resource allocation strategies influenced by government expenditure targets, revealing a complex relationship between economic growth and innovation in this specific context, which actually goes against our expectations on this side since we were expecting that a favourable economic environment would have been largely more beneficial to the vitality of the entire research and development field, fostering patent grants. It is even worth to mention that the result could be biased by the fact that individuals are encouraged to apply for a patent during those periods characterized by solid economic growth and optimistic expectations, but that does not guarantee that the patent would be released in a year that preserve the same characteristics as at the application time (application and verification procedures can take years according to countries' bureaucracy).

Furthermore, the natural logarithm of FDI (*log_s*) shows intriguing patterns. In Specification 1, emphasizing macro-trends, *log_s* has the highest coefficient (82044.77), indicating a robust positive relationship between Foreign Direct Investment and innovation. This aligns with the understanding that a favourable economic environment, as captured by macroeconomic trends, attracts higher levels of FDI, thereby fostering innovation.

Public expenditure variables in Specification 3 offer nuanced insights into their impact on innovation rates within OECD countries. The positive coefficient for government spending on research and development (R&D) (1.925.225) aligns with expectations, indicating that higher R&D investments relative to GDP positively contribute to innovation. Unexpectedly, higher public spending on education as a percentage of GDP exhibits a negative association with the innovation rate (-6.065101). This counterintuitive result warrants further investigation, and suggest the risk of possible unobserved variables, since from one side the global trend of this "patent rush" has been pushing regional innovation to its highest historical data, on the other, given external and procyclin macro-dynamics, such as crisis, trade sanctions and ultimately war tensions, the government public spending in R&D has been progressively reduced to often leave the space to economic recovery. Conversely, increased public spending on healthcare relative to GDP (14.19723) is positively linked to higher innovation rates, suggesting a potential role of facilitating the access to the healthcare system and a healthy population in fostering innovation. These findings underscore the intricate relationship between public expenditure components and innovation dynamics within the OECD countries frame. The table below does not mention statistical significance data (p-value, z) since all the parameters result to be significant within our data frame and model structure.

Table 2: Specification Analysis

	SPECIFICATION 1		**SPECIFICATION 2**		**SPECIFICATION 3**	
	b	**se**	**b**	**se**	**b**	**se**
Naural logarithm of y (GDP)	2514.705	2677.466	5152.036	2583.135	1463.018	2542.915
GDP yearly growth rate	-0.3883	0.309913	-0.58297	0.301106	-0,91844	0.275109
Naural logarithm of s (FDI)	82044.77	37259.08	78259.32	35546.32	67984.52	32154.55
Naural logarithm of t (TB)	-81121.9	36772.13	-77377.9	35081.75	-67234.6	31734.32
Natural logarithm of e	4098.525	5268.398	4148.856	5022.998	4010.877	4546.576
Gender wage gap indicator	-	-	-149.357	0.316669	-0.71456	0.299045
Fertility rate (children/woman)	-	-	4601.328	1045.343	2978.793	9642.898
R&D spending	-	-	-	-	1925.225	2600.789
Education spending		-	-	-	-60.651	2837.728
Healthcare system spending	-	-	-	-	1419.723	1362.253
Constant	-394.487	6432.373	-453.711	6425.924	-522.374	5968.244
Observations	521	-	521	-	521	-

Source: our data elaboration on Stata

Chosen the model specification, we now display results for what were the potential estimators to employ to analyses our panels, i.e., Fixed Effects (FE), Fixed Effects with Clustered Standard Errors (FECL), Random Effects (RE), Random Effects with Robust Standard Errors (RERB), and lastly Pooled Ordinary Least Squares (pooled OLS). Notably, *log_y* (GDP) exhibits minimal variance across RE and RERB, suggesting robustness in this variable's estimation method. GDP yearly growth rate shows consistency across all models, indicating its stable impact on innovation rates. However, for FDI (*log_s*) and Trade Balance (*log_t*), the choice between RE and RERB is crucial, as OLS provides notably different estimates, emphasizing the significance of accounting for unobserved heterogeneity. Employment Rate demonstrates a consistent impact across all models, with marginal differences.

Table 3: Estimated Values

Variable	FE	FECL	RE	RERB	OLS
log_y	142.44644	142.44644	146.30182	146.30182	198.48275
g	-0.93209	-0.93209	-0.91844	-0.91844	-0.31454
log_s	67132.795	67132.795	67984.518	67984.518	65787.939
log_t	-66392.851	-66392.851	-67234.596	-67234.596	-65060.566
log_e	39.33977	39.33977	40.10877	40.10877	81.41465
w	-0.87264	-0.87264	-0.71456	-0.71456	3.83101
f	27.24597	27.24597	29.78793	29.78793	79.19797
r	1.99552	1.99552	1.92522	1.92522	4.25235
u	-6.26554	-6.26554	-6.06510	-6.06510	-9.32652
h	13.66471	13.66471	14.19724	14.19724	29.19655
_cons	-500.92249	-500.92249	-522.37432	-522.37432	-1286.28070

Source: our data elaboration on Stata

In our analysis, the risk of heteroskedasticity introduces a concern regarding the assumption of constant variance in the residuals. Heteroskedasticity, or uneven variability of the error terms, can compromise the reliability of standard errors and, consequently, the validity of statistical inferences. In testing these intrinsic risks in our analysis, the Breusch-Pagan test outcomes, with missing statistics and p-values, left the assessment inconclusive, possibly due to peculiarities in the residuals or data.
However, even though not such valuable, the test statistic (F-statistic) is 6.14 with a p-value of 0.0000. Since the p-value is less than the conventional significance level of 0.05, we would have rejected the null hypothesis. This provides potential preliminary evidence against the assumption of constant variance, suggesting the presence of heteroskedasticity in the model.
In addition, the subsequent White test offered deeper insights, revealing substantial evidence of heteroskedasticity in the residuals. The F-statistic of 6.14, coupled with a p-value of 0.0000, unequivocally indicated the rejection of the assumption of constant variance. Although specific test statistics (W0, W50, W10) and associated p-values were not available, the mean and standard deviation of squared residuals for each country ID underscored significant variability, bolstering the conclusion of heteroskedasticity. Acknowledging the implications of this issue on parameter estimates, we opted for robust standard errors to enhance the robustness of our panel data analysis, accounting for potential variations in error term variance across nations and reinforcing the validity of our statistical inferences.

Table 6: IV regressions

Variable	Legend	IV1	IV2	IV3
w	b	0.16271	30.37978	46.68347
	se	7.31529	12.95725	32.83681
	p	0.9823	0.019	0.1551
log_y	b	38.73575	684.79440	987.49227
	se	123.46527	323.72974	652.96308
	p	0.7537	0.0344	0.1305
log_s	b	11.98115	-260256.28	-511968.85
	se	40.20147	367416.75	724887.6
	p	0.7657	0.4787	0.48
log_t	b	-2.28751	256646.49	504819.36
	se	8.84477	362503.95	715371.43
	p	0.7959	0.479	0.4804
f	b	39.49464	176.35456	242.39084
	se	25.94485	58.37765	134.56462
	p	0.1279	0.0025	0.0717
r	b	7.56697	-20.65011	-36.12017
	se	13.32579	29,.5427	63.73141
	p	0.5701	0.4772	0.5709
u	b	5.71598	-2.46841	-2.05175
	se	10.15501	19.45304	32.62235
	p	0.5735	0.899	0.9499
h	b	20.59631	66.35703	86.96926
	se	4.67605	17,2541	41.66987
	p	0.0	0.0001	0.0369.
_cons	b	-162.46829	-126.29357	-1825.6752
	se	223.65626	452.59054	1145.7025
	p	0.4676	0.0053	0.111

Source: our data elaboration on Stata

Our original random effect regressive model, “xtreg, re” in Stata notation, assumed exogeneity of the explanatory variables. However, concerns about endogeneity prompted the need for rigorous testing. A detailed exploration revealed intricate relationships within the dataset, raising doubts about the independence of variables, such as public expenditure on education (*u*) and public expenditure on healthcare (*h*). The potential correlation of these variables with the error term necessitated endogeneity tests to validate the assumptions of our model.

Upon scrutiny, both the Durbin-Wu Hausman and Hansen Sargan tests yielded low p-values, signaling strong evidence against exogeneity. Consequently, we opted for an instrumental variable (IV) approach to enhance the robustness of our regression analyses. The selection of public expenditure on education (u) and public expenditure on healthcare (h) as instrumental variables stems from their theoretical relevance and orthogonality with the error term.

Our rationale for choosing u and h as instruments lies in their economic significance: public expenditure on education (u) reflects investments in human capital, and public expenditure on healthcare (h) is a key determinant of societal well-being.
The effectiveness of *u* and *h* as instruments is affirmed by IV1, IV2, and IV3 methodologies, where their coefficients exhibit consistency across specifications. Furthermore, correlation analyses confirm the weak correlation between the instrumental variables (*w, g, log_e*) and the chosen instruments (*u, h*), indicating their orthogonality. This comprehensive validation process reinforces our confidence in *u* and *h* as high-quality instruments, supporting the robustness of our regression models in the presence of endogeneity challenges.

6. CONCLUDING REMARKS

In this paper we developed a random effect panel regression model with robust stand errors for a panel of 36 OECD countries over the period from 2001 to 2015 to investigate what are the pivotal factors influencing a country's capability of attracting innovation. We started by considering those factors, as suggested by literature and our knowledge, that could directly influence the innovation rate. However, it was not always trivial to measure and to observe each variable and the external macro-dynamics that could influence them (crises, trade sanctions, war tensions), so we are still aware of the evidence of bias in the estimation, even though also confident with the reliability of our results and their respective robustness checks. Additionally, the presence of bias could be linked to the likelihood of individuals to apply for patents during periods of solid economic growth, which may not guarantee the same characteristics at the time of patent release.
By considering three different model specifications, Macro-trends, Female empowerment and Government expenditure targets, we observed a drastic increase in fitting when social variables, as female empowerment were added, but the ultimate contribution has been given by what were government expenditure targets.
Firstly, we noticed that a higher GDP growth is associated with a lower innovation rate, which was for us an unexpected result. We conclude that this effect may be attributed to policy choices and resource allocation strategies influenced by government expenditure targets, revealing a complex relationship between economic growth and innovation.s Also, the result describing the relation with public spending on education and innovation rate was unexpected, as we discovered a negative correlation that would require further investigation. The counterintuitivity of our result suggests the risk of possible unobserved variables, since from one side the global trend of this "patent rush" has been pushing regional innovation to its highest historical data, on the other, given external and

procyclin macro-dynamics, such as crisis, trade sanctions and ultimately war tensions, the government public spending in R&D has been progressively reduced to often leave the space to economic recovery.

However, we were delighted to notice that there exists a robust positive relationship between foreign Direct Investment (FDI) and innovation. In fact, both government spending on research and development (R&D) and public spending on healthcare positively contributes to a country innovation rate, aligning with expectations. In summary, the findings underscore the complex and multifaceted nature of the relationship between economic variables, government policies, and innovation within OECD countries, paving the way to future discussion, especially on those variables that could lift the veil on endogeneity issues. Ultimately, the analysis suggests the importance of considering gender-related factors and the need for further investigation into unexpected results, potential biases, and the intricate dynamics between different components of public expenditure and innovation.

7. REFERENCES

Bibliography

Griliches, Zvi. "*Patent statistics as economic indicators: a survey.*" R&D and productivity: econometric evidence. University of Chicago Press, 1998. 287-343.

Gambardella, Alfonso, Dietmar Harhoff, and Bart Verspagen. "*The value of patents*." Universita Bocconi, Ludwig-Maximiliens Universitaet, and Eindhoven University, (2005).

Prodan, Igor. "*Influence of research and development expenditures on number of patents applications: selected case studies in OECD countries and central Europe*, *1981-2001*." Applied Econometrics and International Development 5.4 (2005).

Westmore, B. (2013), "*R&D, Patenting and Growth: The Role of Public Policy*", OECD Economics Department Working Papers, No. 1047, OECD Publishing, Paris.

Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, John Van Reenen "*Who Becomes an Inventor in America? The Importance of Exposure to Innovation*", The Quarterly Journal of Economics, Volume 134, Issue 2, May 2019, Pages 647–713.

Cameron, Gavin. "*Innovation and economic growth"*. No. 277. Centre for Economic Performance, London School of Economics and Political Science, 1996.

D. Leonhardt, "*Lost Einsteins: The Innovations We're Missing*", New York Times, 2017
https://www.nytimes.com/2017/12/03/opinion/lost-einsteins-innovation-inequality.html

Hana, Urbancová. "*Competitive advantage achievement through innovation and knowledge*." Journal of competitiveness 5.1 (2013): 82-96.

Raj Chetty, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, *Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States*, The Quarterly Journal of Economics, Volume 129, Issue 4, November 2014, Pages 1553–1623.

Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, John Van Reenen "*Who Becomes an Inventor in America? The Importance of Exposure to Innovation*", The Quarterly Journal of Economics, Volume 134, Issue 2, May 2019, Pages 647–713.

Organization for Economic Co-operation and Development (OECD) (2005). Oslo manual: Guidelines for collecting and interpreting innovation data, Third Edition, Paris.

Akduğan, U.; Doğan, N. Factors Affecting Innovation in OECD Countries. EKOIST Journal of Econometrics and Statistics, [Publisher Location], v. 0, n. 36, p. 111-136, 2022.

Datasets

OECD. "National Accounts at a Glance", OECD National Accounts Statistics (database), 2024. https://doi.org/10.1787/data-00369-en

OECD. "Main Science and Technology Indicators", OECD Science, Technology and R&D Statistics (database), 2024.
https://doi.org/10.1787/data-00182-en

OECD. "Education at a glance: Educational finance indicators", OECD Education Statistics (database), 2024.
https://doi.org/10.1787/c4e1b551-en

OECD. "Health expenditure and financing: Health expenditure indicators", OECD Health Statistics (database), 2024
https://doi.org/10.1787/data-00349-en

OECD (REGPAT). Regional Innovation rate Data, 2010-2019.
https://stats.oecd.org/Index.aspx?DataSetCode=REGION_INNOVATION#

OECD. National GDP Data, 1970-2022.
https://data.oecd.org/gdp/gross-domestic-product-gdp.htm

OECD. GDP yearly growth rate Data, 1961-2022.
https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=OE

OECD. FDI Data, 2005-2022
https://data.oecd.org/fdi/fdi-flows.htm

OECD. Trade Balance Data, 1995-2022.
https://data.oecd.org/trade/trade-in-goods-and-services.htm

OECD. Employment rate Data,1955-2021.
https://stats.oecd.org/Index.aspx?DataSetCode=LFS_D

OECD. Gender Wage Gap indicator Data, 1955-2022.
https://data.oecd.org/earnwage/gender-wage-gap.htm

OECD (Family Indicators database). Fertility rate Data, 1970-2022.

https://data.oecd.org/pop/fertility-rates.htm

OECD (Database on Main Science and Technology Indicators). R&D Spending Data, 2000-2022. https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm

OECD. Education Spending Data, 2000-2020. https://data.oecd.org/eduresource/public-spending-on-education.htm

OECD. Healthcare System Spending Data, 1970-2022. https://data.oecd.org/healthres/health-spending.htm

```
*Who becomes an inventor today?-Experimental analysis-Do-file**

/////////////////////////////////////////////////////////////////////////////
//////////////////////////////Introductory setup/////////////////////////////
/////////////////////////////////////////////////////////////////////////////

clear mata
capture log close
cd "D:\Sorbonne\Econometrics\Project PSME1\Datasets"
clear

/////////////////////////////////////////////////////////////////////////////
/////////////////////////Data import and manipulation///////////////////////
/////////////////////////////////////////////////////////////////////////////

*0) Y0=i--->Regional innovation rate (PCT patent applications per million inhabitants)
*------------------------------------------------------------------------------------
*Data Source: OECD (https://stats.oecd.org/Index.aspx?DataSetCode=REGION_INNOVATION)
import delimited RIRi
describe
*save in Stata format
save dfi, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
keep value territoryleve~y reg_id region year
*equally, another specification must be introduced, i.e., we are only interested in country's level of regional innovation (patent application density per million inhabitants), given our data availability on macro variables
keep if territorylevelandtypology == "Country"
keep value reg_id year
*keep only target countries (OECD members)
keep if reg_id == "AUS" | reg_id == "AUT" | reg_id == "BEL" | reg_id == "CAN" | reg_id == "CHL" | reg_id == "CZE" | reg_id == "DNK" | reg_id == "EST" | reg_id == "FIN" | reg_id == "FRA" | reg_id == "DEU" | reg_id == "GRC" | reg_id == "HUN" | reg_id == "ISL" | reg_id == "IRL" | reg_id == "ISR" | reg_id == "ITA" | reg_id == "JPN" | reg_id == "KOR" | reg_id == "LVA" | reg_id == "LTU" | reg_id == "LUX" | reg_id == "MEX" | reg_id == "NLD" | reg_id == "NZL" | reg_id == "NOR" | reg_id == "POL" | reg_id == "PRT" | reg_id == "SVK" | reg_id == "SVN" | reg_id == "ESP" | reg_id == "SWE" | reg_id == "CHE" | reg_id == "TUR" | reg_id == "GBR" | reg_id == "USA"
*rename (and relabel) original variables' names to better fit the model
rename reg_id country_code
label var country_code "OECD member country code"
rename value i
label var i "Regional innovation rate"
*sort by country_code and year
sort country_code year -i
*we know check for duplicates or disaggregate observations not usefull for our analysis
duplicates drop country_code year, force
*normal works, take the log
generate logi = log(i)
label var logi "Natural logarithm of i"
*verify the better fit of the log transformation through Kernel density plots (example=USA)
kdensity i if country_code == "USA"
kdensity logi if country_code == "USA"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*ultimately manipulate the variables order and sort them
order country_id country_code year i
sort country_code year
save datai, replace
browse
*display a bar chart to a better visualization of aggregate average value of i per each year within the OECD frame
egen mean_i = mean(i), by(year)
summ mean_i, detail
twoway (bar mean_i year, bargap(20)), ///
       title("OECD average i per year") ///
```

```
        xtitle("Year") ytitle("Mean i") ///
        yscale(range(`r(min)' `r(max)'))

*1) X1=y--->Gross domestic product (GDP)
*---------------------------------------
*Data Source: OECD (https://data.oecd.org/gdp/gross-domestic-product-gdp.htm)
clear
import delimited GDPy
describe
*save in Stata format
save dfy, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*some national statistics are displayed in USD_CAP, and some others in MLN_USD depending on data availability. Since we hold that MLN_USD is a more wide-spread and immediate way of measurement and since we are committed to mantain an uniform unit of scale for each variable, we decided to pick this latter and henceforth we will give this assumption as granted when it comes to deal with GDP.
keep if measure == "MLN_USD"
*drop non-involved variables
drop indicator measure subject frequency flagcodes
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value y
label var y "National GDP"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*normal works, take the log
generate logy = log(y)
label var logy "Natural logarithm of y"
*verify the better fit of the log transformation through Kernel density plots (example=USA)
kdensity y if country_code == "USA"
kdensity logy if country_code == "USA"
*sort by coutry_code and year
order country_id country_code year y
sort country_code year
save datay, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code == "FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline y, overlay i(country_code) t(year) ///
    title("GDP by country over time") ///
    xtitle("Year") ytitle("GDP") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*2) X2=g--->GDP growth
*---------------------
```

```
*Data Source: World Bank (https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=OE)
clear
import delimited GDPg
describe
*save in Stata format
save dfg, replace
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*drop non-involved variables
drop indicator measure subject frequency flagcodes
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
rename value g
label var g "GDP yearly growth rate"
*unlike our previous datasets, this dataframe containes not only yearly data, but also disaggregated quarterly data. With respect to our research purpose, we hold that quarterly data would not show statistically significant variations if compared to yearly ones (Chetty, Raj, and Nathaniel Hendren, 2018), hence, we will keep on working with the latter.
*avoiding complicated drops function conditional to having a string variable rather than a numeric one, as in the case of quarterly data expressed as YYYY-Qn, we can just easily drop those observations containing a Q letter in their respective year's value.
drop if strpos(year, "Q") > 0
*destring year to make it a numeric variable
destring year, generate(year_n)
drop year
rename year_n year
label var year "Year"
*first delete those observations going beyond our target time frame
keep if year >= 2001 & year <= 2015
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*sort by coutry_code and year
order country_id country_code year g
sort country_code year
save datag, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code == "FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline g, overlay i(country_code) t(year) ///
    title("GDP Growth by country over time") ///
    xtitle("Year") ytitle("GDP growth") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*3) X3=s--->Foreign direct investement (FDI)
*--------------------------------------------
*Data Source: OECD (https://data.oecd.org/fdi/fdi-flows.htm)
clear
import delimited FDIs
describe
*save in Stata format
save dfs, replace
*As for g (GDP growth), this dataframe containes not only yearly data, but also disaggregated
```

```
quarterly data, for the same reasons mentioned above, we will refuse quarterly data, and keep on working on yearly ones.
*avoiding complicated drops function conditional to having a string variable rather than a numeric one, as in the case of quarterly data expressed as YYYY-Qn, we can just easily drop those observations containing a Q letter in their respective year's value.
drop if strpos(time, "Q") > 0
*destring year to make it a numeric variable
destring time, generate(year_n)
drop time
rename year_n year
label var year "Year"
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if year >= 2001 & year <= 2015
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*drop non-involved variables
drop indicator measure subject frequency flagcodes
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename value s
*the dataset still appears not perfectly structured since, even though the time variable (year) should present yearly data for s (FDI), the same year is repeated 4 time, leading us to deduce that this repetition may refer to data with a quarterly frequency. To validate this hypothesis we have cross-checked this data with the ones presented by the World Bank (BX.KLT.DINV.CD.WD), and we actually confirmed our deductions, making a correction to aggregate results a necessary step to be undertaken. This action must be retained mandatory also considering the nature of our dataframe (panel). Stata would not be able to declare this latter as a panel df if it is not able to uniquelyvocably associate each observation to both a country_code and a year.
collapse (sum) s, by(country_code year)
label var s "Foreign direct investement (FDI)"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*normal works, take the log
*unlike previous variables, s (IDF) can potentially assume both positive and negative values, making critical to deal with log since the logarithmic function is undefined for non-positive numbers. To approach this issue we added a constant to "s" before taking the logarithm, small enough to ensure us to be able to take the log without bringing any significant bias to the analysis. When, we say "small enough", we mean a portion of the minimum value taken by s in our data range.
egen min_s = min(s)
generate log_s = s + abs(min_s) + 1
generate logs = log(log_s)
label var logs "Natural logarithm of s"
drop min_s
drop log_s
*verify the better fit of the log transformation through Kernel density plots (example=USA)
kdensity s if country_code == "USA"
kdensity logs if country_code == "USA"
*plot results through a line graph for each country
tsline s, by(country_code) ///
    title("") ///
    xtitle("Year") ytitle("FDI")
*sort by coutry_code and year
order country_id country_code year s
sort country_code year
save datas, replace
```

```
browse

*4) X4=t--->Trade Openess (Trade in Good and Services)
*-----------------------------------------------------
*Data Source: OECD (https://data.oecd.org/trade/trade-in-goods-and-services.htm)
clear
import delimited TGSt
describe
*save in Stata format
save dft, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*since we hold that MLN_USD is a more wide-spread and immediate way of measurement and since we are committed to mantain an uniform unit of scale for each variable, we decided to pick this latter and henceforth we will give this assumption as granted when it comes to deal with GDP. In this sense, we have refused other measurement units such as %GDP.
keep if measure == "MLN_USD"
*data are also classified by their subject, i.e., if the value represents either exports or imports, but it also already displays the net difference between these latter according to the 2007 notation (exp-imp). As measure of trade openess, we hold that the trade balance is a representative and robust index to adress this measure.
*keep only trade balances' results
keep if subject == "NTRADE"
*drop non-involved variables
drop indicator measure subject frequency flagcodes
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value t
label var t "Trade Balance (TB)"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*as for s (IDF), t (TB) can potentially assume both positive and negative values, making critical to deal with log since the logarithmic function is undefined for non-positive numbers. To approach this issue we added a constant to "t" before taking the logarithm, small enough to ensure us to be able to take the log without bringing any significant bias to the analysis. When, we say "small enough", we mean a portion of the minimum value taken by t in our data range.
egen min_t = min(t)
generate log_t = t + abs(min_t) + 1
generate logt = log(log_t)
label var logt "Natural logarithm of t"
drop min_t
drop log_t
*verify the better fit of the log transformation through Kernel density plots (example=USA)
kdensity t if country_code == "USA"
kdensity logt if country_code == "USA"
*sort by coutry_code and year
order country_id country_code year t
sort country_code year
save datat, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code ==
```

```
"FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline t, overlay i(country_code) t(year) ///
    title("Trade openess by country over time") ///
    xtitle("Year") ytitle("Trade Balance") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*5) X5=e--->Employment rate (LFS)
*---------------------------------
*Data Source: OECD (https://stats.oecd.org/Index.aspx?DataSetCode=LFS_D)
clear
import delimited LFSe
describe
*save in Stata format
save dfe, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
*keep only OECD members countries
rename country location
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*data are also disaggregated by gender, but at this tage we are only interested in aggregated results (MW).
*keep only aggregated results
keep if sex == "MW"
*equally we are interested in the whole labourforce (LF), whithout furthering micro-level consideration in different age-range clusters
keep if v6 == "Total"
*same for our target series (E=LFS employment)
keep if series == "E"
*drop non-involved variables
drop v2 sex v4 age v6 series v8 frequency v10 v12 unit unitcode powercodecode powercode referenceperiodcode referenceperiod flagcodes flags
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value e
label var e "Employment rate (LFS)"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*normal works, take the log
generate loge = log(e)
label var loge "Natural logarithm of e"
*verify the better fit of the log transformation through Kernel density plots (example=USA)
kdensity e if country_code == "USA"
kdensity loge if country_code == "USA"
*sort by coutry_code and year
order country_id country_code year e
sort country_code year
save datae, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code ==
```

```
"FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline e, overlay i(country_code) t(year) ///
    title("Employment rate by country over time") ///
    xtitle("Year") ytitle("Employment rate (LFS)") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*6) X6=w--->Gender wage gap indicator (difference between median earnings of men and women relative to median earnings of men)
*-----------------------------------------------------------------------------------------------------------------------------
*Data Source: OECD (https://data.oecd.org/earnwage/gender-wage-gap.htm)
clear
import delimited GWGw
describe
*save in Stata format
save dfw, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*moreover, we hold that self-employed women represent outliers in the overall frame of female labour force, bringing on the table a considearble risk of bias if included in the model with suitable specifications. Given the limitedness of our data availability, for most pf the countries listed within our dataframe, data regarding the share of self-employed women over the whole female labour force were not available, makimg impossible to run a weighted average to also include these values in our estimation. Hence, we retain more representative and robust to consider the employee component, eventhough we signal that these ruled out observations presents the highest wage gap between male and female individuals.
drop if subject == "SELFEMPLOYED"
*drop non-involved variables
drop indicator subject measure frequency flagcode
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value w
label var w "Gender wage gap indicator (difference between median earnings of men and women relative to median earnings of men)"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*sort by coutry_code and year
order country_id country_code year w
sort country_code year
save dataw, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code == "FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline w, overlay i(country_code) t(year) ///
    title("Gender wage gap by country over time") ///
    xtitle("Year") ytitle("Gender wage gap") ///
```

```
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*7) X7=f--->Fertility rate (children/woman ratio)
*-------------------------------------------
*Data Source: OECD (https://data.oecd.org/pop/fertility-rates.htm)
clear
import delimited FERf
describe
*save in Stata format
save dff, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*drop non-involved variables
drop indicator subject measure frequency flagcode
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value f
label var f "Fertility rate (children/woman)"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*sort by coutry_code and year
order country_id country_code year f
sort country_code year
save dataf, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code == "FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline f, overlay i(country_code) t(year) ///
    title("Fertility rate by country over time") ///
    xtitle("Year") ytitle("children/woman ratio") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*8) X8=r--->R&D spending (public spending on research and development as percentage of GDP)
*-----------------------------------------------------------------------------------------
*Data Source: OECD (https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm)
clear
import delimited RDSr
describe
*save in Stata format
save dfr, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
```

```
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*henceforth to evaluate public spending in each sector we will consider percentage of GDP as unit of scale, since our demand scope is to evaluate whether a greater either goverment or private (aggreagate) focus on a specific public sector may impact an individual's chances of being granted of a patent, and not whether this impact is due to higher absolute values where obviously more populated and richer countries would be advantaged.
keep if measure == "PC_GDP"
*drop non-involved variables
drop indicator subject measure frequency flagcode
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value r
label var r "R&D spending (public spending on research and development as percentage of GDP)"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*sort by coutry_code and year
order country_id country_code year r
sort country_code year
save datar, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code == "FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline r, overlay i(country_code) t(year) ///
    title("R&D spending by country over time") ///
    xtitle("Year") ytitle("R&D spending/GDP") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*9) X9=u--->Education spending (public spending on education as percentage of GDP)
*--------------------------------------------------------------------------------
*Data Source: OECD (https://data.oecd.org/eduresource/public-spending-on-education.htm)
clear
import delimited EDUu
describe
*save in Stata format
save dfu, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*we take in account the PRY_NTRY index, since according to the current literature, it is the more reliable and widespread pre-processed index to measure public spending in education.
```

```
keep if subject == "PRY_NTRY"
*drop non-involved variables
drop indicator subject measure frequency flagcode
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value u
label var u "Education spending (public spending on education as percentage of GDP)"
*declare the data as a panel dataset
egen country_id = group(country_code)
label var country_id "Time-serie country ID number"
xtset country_id year
*sort by coutry_code and year
order country_id country_code year u
sort country_code year
save datau, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code == "FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline u, overlay i(country_code) t(year) ///
    title("Education spending by country over time") ///
    xtitle("Year") ytitle("Education spending/GDP") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

*10) X10=h--->Healthcare system spending (public spending on healthcare as percentage of GDP)
*-----------------------------------------------------------------------------------------
*Data Source: OECD (https://data.oecd.org/healthres/health-spending.htm)
clear
import delimited HEAh
describe
*save in Stata format
save dfh, replace
*now, we do not need all the included variables, but only the only usefull ones for hour regression or functional to our merging process
*first delete those observations going beyond our target time frame
keep if time >= 2001 & time <= 2015
*keep only OECD members countries
keep if location == "AUS" | location == "AUT" | location == "BEL" | location == "CAN" | location == "CHL" | location == "CZE" | location == "DNK" | location == "EST" | location == "FIN" | location == "FRA" | location == "DEU" | location == "GRC" | location == "HUN" | location == "ISL" | location == "IRL" | location == "ISR" | location == "ITA" | location == "JPN" | location == "KOR" | location == "LVA" | location == "LTU" | location == "LUX" | location == "MEX" | location == "NLD" | location == "NZL" | location == "NOR" | location == "POL" | location == "PRT" | location == "SVK" | location == "SVN" | location == "ESP" | location == "SWE" | location == "CHE" | location == "TUR" | location == "GBR" | location == "USA"
*we consider the total aggregate value over national GDP
keep if subject == "TOT"
*as for r and u (R&D spending and education spending over GDP) we take %GDP as unit of scale for the same reasons above mentioned
keep if measure == "PC_GDP"
*drop non-involved variables
drop indicator subject measure frequency flagcode
*rename and relabel targetted variables
rename location country_code
label var country_code "OECD member country code"
rename time year
label var year "Year"
rename value h
label var h "Healthcare system spending (public spending on healthcare as percentage of GDP)"
*declare the data as a panel dataset
egen country_id = group(country_code)
```

```
label var country_id "Time-serie country ID number"
xtset country_id year
*sort by coutry_code and year
order country_id country_code year h
sort country_code year
save datah, replace
browse
*plot results through a line graph for each country
keep if country_code == "AUS" | country_code == "CHL" | country_code == "DEU" | country_code == "FRA" | country_code == "GBR" | country_code == "ITA" | country_code == "JPN" | country_code == "MEX" | country_code == "NOR" | country_code == "USA"
xtline h, overlay i(country_code) t(year) ///
    title("Healthcare spending by country over time") ///
    xtitle("Year") ytitle("Healthcare spending/GDP") ///
    legend(label(1 "Australia (AUS)") label(2 "Chile (CHL)") label(3 "Germany (DEU)") ///
           label(4 "France (FRA)") label(5 "United Kingdom (GBR)") label(6 "Italy (ITA)") ///
           label(7 "Japan (JPN)") label(8 "Mexico (MEX)") label(9 "Norway (NOR)") ///
           label(10 "United States (USA)"))

/////////////////////////////////////////////////////////////////////////////
///////////////////////////////Final merging/////////////////////////////////
/////////////////////////////////////////////////////////////////////////////

*in this section we progressively merge each dataset to work with a unique dataframe enclosing all our variables and observations in a panel format
use datai, clear
merge 1:1 country_id year using "datay"
drop _merge
save data1, replace
*we repeat the same procedure for all the disaggregate dataframes referred to each variable
use data1, clear
merge 1:1 country_id year using "datag"
drop _merge
save data2, replace

use data2, clear
merge 1:1 country_id year using "datas"
drop _merge
save data3, replace

use data3, clear
merge 1:1 country_id year using "datat"
drop _merge
save data4, replace

use data4, clear
merge 1:1 country_id year using "datae"
drop _merge
save data5, replace

use data5, clear
merge 1:1 country_id year using "dataw"
drop _merge
save data6, replace

use data6, clear
merge 1:1 country_id year using "dataf"
drop _merge
save data7, replace

use data7, clear
merge 1:1 country_id year using "datar"
drop _merge
save data8, replace

use data8, clear
merge 1:1 country_id year using "datau"
drop _merge
```

```
save data9, replace

use data9, clear
merge 1:1 country_id year using "datah"
drop _merge
save merged_data, replace
describe
browse
*declare the data as a panel dataset
xtset country_id year

////////////////////////////////////////////////////////////////////////////
//////////////////Missing values regressive forecasts /////////////////////
////////////////////////////////////////////////////////////////////////////

*since we have missing values for some variables, we employ a panel random effect estimator for
each of these latter to generate a prediction on both possible past and future observations,
which, according to Hausman test, provides the better estimation for our unadjusted data.
Notwithstanding our large number of involved variables, we are only able to employ those
presenting a complete range of observations over time, otheriwise the regression would keep
predicting missing results. It is worth to notice that for some countries data are complete also
for variables not listed in our estimation model, but in order to avoid over-noise-campturing
bias, that would then possibly affect only these exceptional countries, we have opted for a
uniformed prediction model, exclusively including those complete variables common to all countries.

use merged_data, clear
xtset country_id year

*1) predicted s
*--------------
xtreg s f logt g logy, re
predict pred_s
*Replace missing values in 's' with predicted values
replace s = pred_s if missing(s)
drop pred_s
describe
*in case of log transformed variables, we also need to retake the log function for the new
predicted values. We hold that doing a 2-stage log transformation makes the regressive forecast
more robust to possible heteroschedacity or high-skewness possible biases.
*Redo log for new predicted values
egen max_s = max(s)
generate log_s1 = t + abs(max_s) + 1
generate log_s = log(log_s1)
label var log_s "Natural logarithm of e"
drop max_s
drop log_s1
drop logs

*2) predicted e
*--------------
xtreg e f logt g logy, re
predict pred_e
*Replace missing values in 'e' with predicted values
replace e = pred_e if missing(e)
drop pred_e
describe
*Redo log for new predicted values
egen max_e = max(e)
generate log_e1 = t + abs(max_e) + 1
generate log_e = log(log_e1)
label var log_e "Natural logarithm of e"
drop max_e
drop log_e1
drop loge

*3) predicted w
*--------------
xtreg w f logt g logy, re
```

```
predict pred_w
*Replace missing values in 'w' with predicted values
replace w = pred_w if missing(w)
drop pred_w

*4) predicted r
*--------------
xtreg r f logt g logy, re
predict pred_r
*Replace missing values in 'w' with predicted values
replace r = pred_r if missing(r)
drop pred_r

*5) predicted u
*--------------
xtreg u f logt g logy, re
predict pred_u
*Replace missing values in 'w' with predicted values
replace u = pred_u if missing(u)
drop pred_u

*uniformely rename variables
rename logi log_i
rename logy log_y
rename logt log_t
*adjust the order of varibales to better reflect our format
*sort by coutry_code and year
order country_id country_code year i log_i y log_y g s log_s t log_t e log_e w f r u h
sort country_code year
save pred_data, replace
*declare the data as a panel dataset
xtset country_id year
browse

/////////////////////////////////////////////////////////////////////////////
///////////////////////////Stationarity Tests//////////////////////////////
/////////////////////////////////////////////////////////////////////////////

*At this point, stationarity tests are crucial to validate that statistical properties like mean and variance remain constant over time throught our dataframe. Ensuring stationarity is vital for accurate modeling, forecasting, and reliable statistical inference. Non-stationary data may mislead models and compromise forecasting accuracy. Overall, stationarity verification is essential for maintaining the stability and interpretability of time series models. To investigate the heterogenous range of non-stationarity biases we have involved four different tests: Im-Pesaran-Shin (IPS) test, Breitung Panel Unit Root Test, Hadri LM Test and Levin, Lin, and Chu (LLC) Test. It is worth to notice that in our case, traditional approaches, such as Augmented Dickey-Fuller test, are not available given the multitude of panel belonging to the dataframe.

*1) stationarity analysis for y
*-----------------------------
*The Im-Pesaran-Shin (IPS) test is assessing whether the target variable is stationary across individual time series. It provides insights into the common stochastic trend shared by the panel.
xtunitroot ips y
*The Im-Pesaran-Shin (IPS) unit-root test results for the variable "y" indicate a positive test statistic of 8.8557 and a p-value of 1.0000. These values fail to reject the null hypothesis of unit roots, suggesting that "y" is likely non-stationary across the panel.

*The Breitung Panel Unit Root Test examines stationarity in the target variable, considering cross-section dependence. It's robust in the presence of correlated data among entities.
xtunitroot breitung y
*The test yielded a test statistic of 13.5466 and a p-value of 1.0000. These results indicate a failure to reject the null hypothesis, suggesting that y is likely non-stationary across the panel.

*The Hadri LM Test assesses stationarity in the target variable for panel data, particularly considering the impact of cross-sectional dependence on the results.
xtunitroot hadri y

```

```
*The Levin, Lin, and Chu (LLC) Test evaluates stationarity in the target variable in panel data while accounting for both individual effects and cross-section dependence.
xtunitroot llc y
*The Hadri LM test yielded a highly significant test statistic of 47.9938 and a p-value of 0.0000. This strong evidence led to the rejection of the null hypothesis that all panels are stationary, indicating that variable "y" is likely non-stationary across the panel

*2) stationarity analysis for g
*------------------------------
*IPS test
xtunitroot ips g
*The test resulted in a test statistic of -6.6116 and a p-value of 0.0000. These findings lead to the rejection of the null hypothesis, providing evidence that the variable "g" is likely non-stationary across the panel.

*Breitung Panel Unit Root Test
xtunitroot breitung g
*The test yielded a test statistic of -9.6198 and a p-value of 0.0000. These results indicate a rejection of the null hypothesis, providing evidence that the variable "g" is likely non-stationary across the panel.

*Hadri LM Test
xtunitroot hadri g
*The test yielded a test statistic of 4.6573 and a p-value of 0.0000. These results indicate a rejection of the null hypothesis, providing evidence that the variable "g" is likely non-stationary across the panel.

*LLC test
xtunitroot llc g
*These results indicate a rejection of the null hypothesis, providing strong evidence that the variable "g" is likely non-stationary across the panel

*3) stationarity analysis for s
*------------------------------
*IPS test
xtunitroot ips s
*The test resulted in a test statistic of -3.5428 and a p-value of 0.0002. These findings lead to the rejection of the null hypothesis, providing evidence that the variable "s" is likely non-stationary across the panel.

*Breitung Panel Unit Root Test
xtunitroot breitung s
*The test yielded a test statistic of -2.5754 and a p-value of 0.0050. These results indicate a rejection of the null hypothesis, providing evidence that the variable "s" is likely non-stationary across the panel.

*Hadri LM Test
xtunitroot hadri s
*The test yielded a test statistic of 19.2455 and a p-value of 0.0000. These results strongly support the rejection of the null hypothesis, indicating that the variable "s" is likely non-stationary across the panel.

*LLC test
xtunitroot llc s
*These results strongly indicate a rejection of the null hypothesis, providing evidence that the variable "s" is likely non-stationary across the panel.

*4) stationarity analysis for t
*------------------------------
*IPS test
xtunitroot ips t
*The test resulted in a test statistic of -1.1581 and a p-value of 0.9861. The critical values for rejection are -1.830 (1%), -1.740 (5%), and -1.690 (10%). Given that the test statistic is less extreme than the critical values, we fail to reject the null hypothesis. This indicates that variable "t" is likely non-stationary across the panel.

*Breitung Panel Unit Root Test
xtunitroot breitung t
```

```
*The test statistic yielded a value of 1.3425 with a corresponding p-value of 0.9103. Given that the test statistic is not extreme and the p-value is high, there is insufficient evidence to reject the null hypothesis. This indicates that variable "t" is likely non-stationary across the panel.

*Hadri LM test
xtunitroot hadri t
*The test statistic yielded a value of 24.9452 with a corresponding p-value of 0.0000. The high test statistic and very low p-value provide strong evidence to reject the null hypothesis. This suggests that variable "t" is likely non-stationary across the panel.

*LLC test
xtunitroot llc t
*The test statistic resulted in a value of -9.1361 with a corresponding p-value of 0.0000. The negative test statistic and very low p-value provide strong evidence to reject the null hypothesis, indicating that variable "t" is likely non-stationary across the panel.

*5) stationarity analysis for e
*------------------------------
*IPS test
xtunitroot ips e
*The test statistics did not provide sufficient evidence to reject the null hypothesis, as the p-value associated with Z-t-tilde-bar was 0.6817, exceeding the common significance levels. Consequently, there is insufficient evidence to conclude that variable "e" is stationary based on the IPS test, implying the potential presence of unit roots.

*Breitung Panel Unit Root Test
xtunitroot breitung e
*The test statistic lambda was -0.0312, and the p-value was 0.4876. The p-value exceeds common significance levels, indicating that there is insufficient evidence to reject the null hypothesis. Consequently, based on the Breitung test, variable "e" may contain unit roots, implying non-stationarity across panels.

*Hadri LM test
xtunitroot hadri e
*The test statistic, lambda, yielded a value of -0.0312, with an associated p-value of 0.4876. Given that the p-value exceeds common significance levels (e.g., 0.05), there is insufficient evidence to reject the null hypothesis. Consequently, based on the Breitung test, variable "e" may exhibit unit roots, indicating non-stationarity across panels.

*LLC test
xtunitroot llc e
*The test results yielded a significant test statistic, with an unadjusted t-value of -9.2830 and an adjusted t-value of -4.2159, both implying strong evidence against the null hypothesis. Therefore, based on the LLC test, variable "e" appears to be stationary across panels, suggesting a lack of unit roots.

*In evaluating the reliability of the unit-root tests for variable "e," the Levin–Lin–Chu (LLC) test stands out as more trustworthy. The LLC test incorporates a lagged regression and addresses heteroskedasticity through the Bartlett kernel, providing a comprehensive assessment of stationarity. The notable and negative adjusted t-value in the LLC test signals a rejection of the null hypothesis, supporting the argument for stationarity.

*6) stationarity analysis for w
*------------------------------
*IPS test
xtunitroot ips w
*The Im–Pesaran–Shin (IPS) unit-root test results for variable "w" indicate a significant test statistic, with a Z-t-tilde-bar value of -6.9155 and a p-value of 0.0000. This suggests evidence against the null hypothesis, implying non-stationarity for the variable.

*Breitung Panel Unit Root Test
xtunitroot breitung w
*The test statistic, lambda, was calculated as -0.4973 with a p-value of 0.3095. This result does not provide sufficient evidence to reject the null hypothesis, indicating that variable "w" may have unit roots, suggesting non-stationarity.

*Hadri LM test
```

```
xtunitroot hadri w
*The test yielded a significant statistic, with a z-value of 5.5229 and a p-value of 0.0000. This strong evidence against the null hypothesis suggests that variable "w" is likely stationary across panels, indicating a lack of unit roots.

*LLC test
xtunitroot llc w
*The test yielded a significant unadjusted t-value of -7.3412 and an adjusted t-value of -3.2765, providing strong evidence against the null hypothesis. This implies that variable "w" appears to be stationary across panels, indicating a lack of unit roots.

*Given the results of the Im-Pesaran-Shin (IPS), Breitung, Hadri LM, and Levin-Lin-Chu (LLC) unit-root tests for variable "w," the significant p-value in the Hadri LM test (p-value = 0.0000) and the adjusted t-value in the LLC test (t = -3.2765) suggest strong evidence against the null hypothesis of unit roots, indicating that the variable "w" is likely stationary across panels. Therefore, based on these findings, it is reasonable to conclude that variable "w" is stationary in the panel dataset.

*7) stationarity analysis for f
*------------------------------
*IPS test
xtunitroot ips f
*Based on the results of the Im-Pesaran-Shin (IPS) unit-root test for variable "f," the test statistic (Z-t-tilde-bar) is 2.5333 with a p-value of 0.9944. These values suggest weak evidence against the null hypothesis of unit roots, indicating that variable "f" is likely non-stationary across panels.

*Breitung Panel Unit Root Test
xtunitroot breitung f
*The Breitung unit-root test for variable "f" yielded a test statistic (lambda) of 2.6238 with a p-value of 0.9957. These results provide insufficient evidence to reject the null hypothesis, suggesting that variable "f" likely contains unit roots and is non-stationary across panels.

*Hadri LM test
xtunitroot hadri f
*The Hadri LM test for variable "f" produced a test statistic of 29.8304 with a p-value of 0.0000. The highly significant p-value suggests strong evidence against the null hypothesis, indicating that variable "f" is likely stationary across panels.

*LLC test
xtunitroot llc f
*The Levin-Lin-Chu unit-root test (LLC) for variable "f" resulted in a test statistic of -4.9840 and an adjusted t-value of -0.2989, with a p-value of 0.3825. The unadjusted and adjusted t-values both fail to reject the null hypothesis of unit roots, suggesting a lack of stationarity.

*The Hadri LM test accounts for potential violations of homoskedasticity assumptions, making it suitable for this dataset. Additionally, the highly significant test statistic and low p-value reinforce its credibility in rejecting the null hypothesis of unit roots, leading to the conclusion of stationarity for variable "f".

*8) stationarity analysis for r
*------------------------------
*IPS test
xtunitroot ips r
*The test results reveal a test statistic of -1.3947 and a corresponding p-value of 0.9999. The high p-value fails to provide evidence against the null hypothesis, suggesting a lack of stationarity for variable "r."

*Breitung Panel Unit Root Test
xtunitroot breitung r
*The test yielded a statistic of 2.9776 and a corresponding p-value of 0.9985. The high p-value fails to provide evidence against the null hypothesis, suggesting a lack of stationarity for variable "r".

*Hadri LM test
xtunitroot hadri r
*The test resulted in a significant statistic of 29.1043 with a p-value of 0.0000, providing strong evidence against the null hypothesis. This indicates that variable "r" is likely to be
```

```
non-stationary across panels, suggesting the presence of unit roots.

*LLC test
xtunitroot llc r
*The Levin-Lin-Chu (LLC) unit-root test for variable "r" in the panel dataset resulted in a non-significant test statistic, with an adjusted t-value of -0.9005 and a p-value of 0.1839. This suggests insufficient evidence to reject the null hypothesis that panels contain unit roots, implying a lack of clear stationarity for variable "r" across panels.

*9) stationarity analysis for u
*------------------------------
*IPS test
xtunitroot ips u
*The test results yielded a significant test statistic, with a Z-t-tilde-bar value of -3.4928 and a p-value of 0.0002, providing strong evidence against the null hypothesis. Therefore, "u" appears to be stationary across panels, suggesting a lack of unit roots.

*Breitung Panel Unit Root Test
xtunitroot breitung u
*The test results produced a highly significant test statistic, with a lambda value of -5.7768 and a p-value of 0.0000, providing strong evidence against the null hypothesis. Therefore, "u" appears to be stationary across panels, indicating a lack of unit roots.

*Hadri LM test
xtunitroot hadri u
*The test results revealed a significant test statistic, with a z-value of 17.4116 and a p-value of 0.0000, providing robust evidence against the null hypothesis. Therefore, "u" is deemed to be stationary across panels, implying the absence of unit roots.

*LLC test
xtunitroot llc u
*The test results yielded a significant test statistic, with an unadjusted t-value of -11.6567 and an adjusted t-value of -4.8094, both implying strong evidence against the null hypothesis. Therefore, "u" appears to be stationary across panels, suggesting a lack of unit roots.


*10) stationarity analysis for h
*-------------------------------
*IPS test
xtunitroot ips h
*The test did not yield a statistically significant result, as the test statistic did not exceed the critical values at conventional significance levels (1%, 5%, 10%). Therefore, we do not reject the null hypothesis, suggesting that variable "h" may exhibit unit roots across the panels.

*Breitung Panel Unit Root Test
xtunitroot breitung h
*The test statistic, represented by lambda, was found to be 3.0187 with a p-value of 0.9987. As the p-value exceeds conventional significance levels, we fail to reject the null hypothesis. Therefore, "h" is likely to have unit roots across the panels, indicating non-stationarity.

*Hadri LM test
xtunitroot hadri h
*The test statistic, represented by "z," was found to be 37.4329 with a p-value of 0.0000, indicating strong evidence against the null hypothesis. Therefore, "h" is likely to be stationary across all panels, suggesting the absence of unit roots.

*LLC test
xtunitroot llc h
*The test results yielded a significant test statistic, with an unadjusted t-value of -7.2781 and an adjusted t-value of -3.5677, both implying strong evidence against the null hypothesis. Therefore, based on the LLC test, variable "h" appears to be stationary across panels, suggesting a lack of unit roots.

*Im-Pesaran-Shin (IPS) and Breitung tests do not provide conclusive evidence in this case to clearly assert whether the variable can be retained stationary or not. Therefore, based on the more reliable Hadri LM and LLC tests, we can infer that variable "h" is likely stationary across panels, as the significancy of their results shows.

```

```
////////////////////////////////////////////////////////////////////////////
//////////////////////////Stationarity Correction//////////////////////////
////////////////////////////////////////////////////////////////////////////

*Given the results in terms of stationarity extrapolated from the employed stationarity tests, we want now to implement corrections to those variables that were identified as non-stationary. To do so, we will recur to first differences, following the rationale of eliminating trends and making the series more amenable to statistical analyses.

*recall the unadjusted dataset
use pred_data, clear
xtset country_id year

*1) y adjusted
*-------------
*generate first differences
gen y_diff = y - L.y
*drop the original variable if you want to keep only the differenced variable
drop y
*rename the differenced variable to the original variable name
rename y_diff y
*LLC test
xtunitroot llc y
*The differenced variable 'y' appears to be stationary across panels, supporting the effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg y g log_s log_t log_e w f r, re
predict pred_y
*Replace missing values in with predicted values
replace y = pred_y if missing(y)
drop pred_y
label var y "GDP"

*2) log_y adjusted
*----------------
*generate first differences
gen log_y_diff = log_y - L.log_y
*drop the original variable if you want to keep only the differenced variable
drop log_y
*rename the differenced variable to the original variable name
rename log_y_diff log_y
*LLC test
xtunitroot llc log_y
*The differenced variable 'log_y' appears to be stationary across panels, supporting the effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg log_y g log_s log_t log_e w f r u h, re
predict pred_log_y
*Replace missing values in 's' with predicted values
replace log_y = pred_log_y if missing(log_y)
drop pred_log_y
label var log_y "Naural logarithm of y (GDP)"

*3) g adjusted
*-------------
*generate first differences
gen g_diff = g - L.g
*drop the original variable if you want to keep only the differenced variable
drop g
*rename the differenced variable to the original variable name
rename g_diff g
*LLC test
xtunitroot llc g
*The differenced variable 'g' appears to be stationary across panels, supporting the effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg g log_y log_s log_t log_e w f r u h, re
predict pred_g
```

```
*Replace missing values in with predicted values
replace g = pred_g if missing(g)
drop pred_g
label var g "GDP yearly growth rate"

*4) s adjusted
*-------------
*generate first differences
gen s_diff = s - L.s
*drop the original variable if you want to keep only the differenced variable
drop s
*rename the differenced variable to the original variable name
rename s_diff s
*LLC test
xtunitroot llc s
*The differenced variable 's' appears to be stationary across panels, supporting the effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg s log_y g log_t log_e w f r u h, re
predict pred_s
*Replace missing values in with predicted values
replace s = pred_s if missing(s)
drop pred_s
label var s "Foreign direct investement (FDI)"

*5) log_s adjusted
*----------------
*generate first differences
gen log_s_diff = log_s - L.log_s
*drop the original variable if you want to keep only the differenced variable
drop log_s
*rename the differenced variable to the original variable name
rename log_s_diff log_s
*LLC test is not applicable in this case, since it requires strongly balanced data. We will then try to deploy IPS test
*xtunitroot ips log_s
*None of the previous, and further, tests actually produces statistically significant results for this variable analysis, leading to uncertain conclusions on the effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg log_s g log_y log_t log_e w f r u h, re
predict pred_log_s
*Replace missing values in 's' with predicted values
replace log_s = pred_log_s if missing(log_s)
drop pred_log_s
label var log_s "Naural logarithm of s (FDI)"

*6) t adjusted
*-------------
*generate first differences
gen t_diff = t - L.t
*drop the original variable if you want to keep only the differenced variable
drop t
*rename the differenced variable to the original variable name
rename t_diff t
*LLC test
xtunitroot llc t
*The differenced variable 't' appears to be stationary across panels, supporting the effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg t log_y g log_s log_e w f r u h, re
predict pred_t
*Replace missing values in with predicted values
replace t = pred_t if missing(t)
drop pred_t
label var t "Trade Balance (TB)"

*7) log_t adjusted
```

```
*----------------
*generate first differences
gen log_t_diff = log_t - L.log_t
*drop the original variable if you want to keep only the differenced variable
drop log_t
*rename the differenced variable to the original variable name
rename log_t_diff log_t
*LLC test
xtunitroot llc log_t
*The differenced variable 'log_t' appears to be stationary across panels, supporting the
effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg log_t g log_y log_s log_e w f r u h, re
predict pred_log_t
*Replace missing values in 's' with predicted values
replace log_t = pred_log_t if missing(log_t)
drop pred_log_t
label var log_t "Naural logarithm of t (TB)"

*8) r adjusted
*-------------
*generate first differences
gen r_diff = r - L.r
*drop the original variable if you want to keep only the differenced variable
drop r
*rename the differenced variable to the original variable name
rename r_diff r
*LLC test
xtunitroot llc r
*The differenced variable 'r' appears to be stationary across panels, supporting the
effectiveness of the differencing transformation in achieving stationarity.
*repredict missing values to have complete data from 2001 to 2015
xtreg r log_y g log_s log_t log_e w f u h, re
predict pred_r
*Replace missing values in with predicted values
replace r = pred_r if missing(r)
drop pred_r
label var r "R&D spending (public spending on research and development as percentage of GDP)"

*reorder the final dataframe according to the new changes
*sort by coutry_code and year
order country_id country_code year i log_i y log_y g s log_s t log_t e log_e w f r u h
sort country_code year
save adj_data, replace
*declare the data as a panel dataset
xtset country_id year
browse

/////////////////////////////////////////////////////////////////////////////
////////////////////////////Model Specification/////////////////////////////
/////////////////////////////////////////////////////////////////////////////

use adj_data, clear
xtset country_id year
ssc install outreg2
ssc install estout

*---------------------------------------------------------------------------
*Specification_1----->Macro-trends
xtreg i log_y g log_s log_t log_e, re
predict pred_i1
outreg2 using "Specification_1", replace ctitle(Baseline)
est store spec1

*Specification_2----->Female empowerment
xtreg i log_y g log_s log_t log_e w f, re
predict pred_i2
outreg2 using "Specification_2", replace ctitle(Baseline)
```

```
est store spec2

*Specification_3----->Goverment expenditure targets
xtreg i log_y g log_s log_t log_e w f r u h, re
predict pred_i3
outreg2 using "Specification_3", replace ctitle(Baseline)
est store spec3
*----------------------------------------------------------------------------
*export results
esttab spec1 spec2 spec3 using "raw_res.csv", se label replace wide plain

////////////////////////////////////////////////////////////////////////////
/////////////////////////Panel data specifications//////////////////////////
////////////////////////////////////////////////////////////////////////////

use adj_data, clear
xtset country_id year

*1) Introductory setup
*---------------------
*we deploy this test to investigate which panel estimator works better with our data
*FIXED EFFECTS----->without clustered SEs
xtreg i log_y g log_s log_t log_e w f r u h , fe
estimates store FE
*FIXED EFFECTS----->with clustered SEs
xtreg i log_y g log_s log_t log_e w f r u h , fe cluster(country_id)
estimates store FECL
*RANDOM EFFECTS----->without robust SEs
xtreg i log_y g log_s log_t log_e w f r u h , re
estimates store RE
*RANDOM EFFECTS----->with robust SEs
xtreg i log_y g log_s log_t log_e w f r u h , re robust
estimates store RERB
*POOLED OLS
reg i log_y g log_s log_t log_e w f r u h
estimates store OLS
*Pooled OLS is appropriate when you treat the panel data as if it were a single cross-sectional dataset, ignoring any individual-specific or time-specific effects. This assumes that there is no correlation between the individual effects and the independent variables, which is not our case.

*comparision of different estimated results
estimates table FE FECL RE RERB OLS

*2) Hausman test
*---------------
hausman FE RE
*The results of the Hausman test indicate that the p-value is very high (0.9992), suggesting that you fail to reject the null hypothesis. The null hypothesis in the Hausman test is that the difference in coefficients between the fixed-effects (FE) and random-effects (RE) models is not systematic. Since the p-value is high, there is no strong evidence to suggest that the differences in coefficients between the two models are systematic. Therefore, we choose the random-effects model as it is consistent with the assumptions of the test.

*hence, generalizing these findings to our previous estimations, we can assert that RE is the best estimator we can deploy within our dataframe for this regression model.

*3) First-order autocorrelation check
*------------------------------------
sort country_id year
gen lagi = L.i
br country_id year i lagi
correlate i L.i
*The correlation analysis reveals a strong positive autocorrelation between the variable i and its first lag L.i, with a correlation coefficient of approximately 0.9938. This indicates a substantial linear relationship between the variable and its past values.

////////////////////////////////////////////////////////////////////////////
///////////////////////////Functional form test/////////////////////////////
```

```
////////////////////////////////////////////////////////////////////////////

use adj_data, clear
xtset country_id year

*1) non-normality of error terms checks
*---------------------------------------
/////////SPEC1/////////
xtreg i log_y g log_s log_t log_e, re
predict pred_i1
*Kernel density distribution vs normal
kdensity pred_i1, normal
*Jacques_Bera test
sktest pred_i1
*Shapiro-Wilk Test
swilk pred_i1
*The Jacques-Bera test for skewness and kurtosis, the Skewness and Kurtosis tests, as well as the Shapiro-Wilk test, all indicate a departure from normality. The p-values associated with these tests are all very low (close to or equal to zero), leading to the rejection of the null hypothesis that the data follows a normal distribution.

/////////SPEC2/////////
xtreg i log_y g log_s log_t log_e w f, re
predict pred_i2
*Kernel density distribution vs normal
kdensity pred_i2, normal
*Jacques_Bera test
sktest pred_i2
*Shapiro-Wilk Test
swilk pred_i2
*The Jacques-Bera test for skewness and kurtosis, the Skewness and Kurtosis tests, as well as the Shapiro-Wilk test, all provide evidence against the hypothesis that the data follows a normal distribution. The p-values associated with these tests are very low (close to or equal to zero), leading to the rejection of the null hypothesis. Therefore, based on these normality tests, data do not appear to be normally distributed.

/////////SPEC3/////////
xtreg i log_y g log_s log_t log_e w f r u h, re
predict pred_i3
*Kernel density distribution vs normal
kdensity pred_i3, normal
*Jacques_Bera test
sktest pred_i3
*Shapiro-Wilk Test
swilk pred_i3
*The p-values from the Jacques-Bera test for skewness and kurtosis, the Skewness and Kurtosis tests, and the Shapiro-Wilk test are relatively high. While the p-value from the Jacques-Bera test for skewness and kurtosis is not very elevate, the other two tests provide some support for the hypothesis of normality. Therefore, based on these normality tests, data appears to be relatively close to a normal distribution.

*Lower values of AIC and BIC indicate a better fit. Therefore, Specification 3 (Goverment expenditure targets) appears to have the best fit among the three models also according to R-squared. Moreover, R-squared values increase from Spec1 to Spec3, indicating a better ability to explain the variance in the dependent variable.

*2) Ramsey reset test
*--------------------
xtreg i log_y g log_s log_t log_e w f r u h, re
predict pred_i
gen form2 = pred_i^2
gen form3 = pred_i^3
*fit test including form2 and form3
xtreg i log_y g log_s log_t log_e w f r u h form2 form3, re
*RESET test over the new experimental variables
test form2 form3
*The p-value for the RESET test is very small (p = 0.0000), indicating that we reject the null hypothesis. This suggests that there might be a specification error in the model related to form2
```

```
and form3, leading us to rule them out of the model specification- The test confirm us that we are working with the right functional form.

////////////////////////////////////////////////////////////////////////////
/////////////////////Heteroscedasticity robustness test/////////////////////
////////////////////////////////////////////////////////////////////////////

use adj_data, clear
xtset country_id year
ssc install whitetst

*1) Breusch-Pagan test
*---------------------
xtreg i log_y g log_s log_t log_e w f r u h, re
predict uhat, u
*Manually run Breusch-Pagan LM test
robvar uhat, by(country_id)

*The result of the Breusch-Pagan test for heteroscedasticity indicates that the test statistics for all three weighting matrices (W0, W50, and W10) are missing (denoted by .), and the corresponding p-values are also missing. This suggests that the test results are inconclusive or invalid due to some issue with the residuals or the data.

*The test statistic (F-statistic) is 6.14 with a p-value of 0.0000. Since the p-value is less than the conventional significance level of 0.05, you would reject the null hypothesis. This provides evidence against the assumption of constant variance, suggesting the presence of heteroskedasticity in the model.

*2) White test
*-------------
*Manually run White test
gen uhat_sq = uhat^2
robvar uhat_sq, by(country_id)
*The results indicate the presence of heteroscedasticity in the residuals. The test statistics (W0, W50, W10) are not reported, and the associated p-values are also not available (denoted as "."), suggesting potential issues with the estimation or the grouping variable. However, the mean and standard deviation of squared residuals for each country ID reveal substantial variability, supporting the conclusion of heteroscedasticity

*given these results, to avoid highly probable heteroskedasticity biases, we now want Stata to estimate the standard errors using a method that is less sensitive to the presence of heteroskedasticity. To do so, we employ robust standard errors
xtreg i log_y g log_s log_t log_e w f r u h, re
est store BASE
xtreg i log_y g log_s log_t log_e w f r u h, robust
est store ROBUST
*export results
esttab BASE ROBUST using "Spec_3_Base&Robust.csv", se label replace wide plain

////////////////////////////////////////////////////////////////////////////
///////////////////////Endogeneity robustness checks///////////////////////
////////////////////////////////////////////////////////////////////////////

*To ensure the robustness of our analysis, we employ exogeneity tests, which scrutinize whether our chosen explanatory variables are free from correlation with the error term. This test is instrumental in validating the assumptions underlying our model, providing insights into the potential bias introduced by endogeneity.

use adj_data, clear
xtset country_id year
ssc install estout

*1) IV regression
*----------------
*pretend to use model variables as instrumental variables (IV validity)
ivregress 2sls i log_y log_s log_t f r u h ( w=g), first
est store IV1
ivregress 2sls i log_y log_s log_t f r u  h ( w= g log_e), first
```

```
est store IV2
ivregress 2sls i log_y log_s log_t f r u  h ( w=log_e), first vce(robust)
est store IV3
*display results
est table IV1 IV2 IV3, se p
esttab IV1 IV2 IV3, se

*2) Durbin Wu Hausman test
*-------------------------
estat endogenous
*Both the robust score chi-squared test and the robust regression F-test have low p-values (0.0000), indicating strong evidence against the null hypothesis. Therefore, we reject the null hypothesis, suggesting that the model variables are endogenous. Hence, w and log_e must be included in our model.

*3) Hansen Sargan test
*---------------------
estat overid
*The robust score chi-squared statistic is 57.7889 with a p-value of 0.0000, and the robust regression F-statistic is 37.0072 with a p-value of 0.0000. These findings provide robust support for the rejection of the hypothesis that the included variables are exogenous, suggesting that they are likely endogenous in your model. Hence, g and log_e must be included.

*4) verify the quality of our instrumental variables
*--------------------------------------------------
corr w g log_e
reg w g log_e log_y log_s log_t f r u h, robust
test g log_e
*the test indicates that there is no significant evidence to reject the null hypothesis that both g and log_e coefficients are zero. The F-statistic is 1.59, and the p-value is 0.2044. This suggests that these variables do not contribute significantly to explaining the variation in the dependent variable.
```