Social Indicators Research
https://doi.org/10.1007/s11205-023-03152-3

ORIGINAL RESEARCH

A Model for Evaluating Inequalities in Sustainability

Ida Camminatiello[1] · Rosaria Lombardo[1] · Mario Musella[2] · Gianmarco Borrata[2]

Accepted: 24 May 2023

Abstract
On 25 September 2015, the *United Nations General Assembly* adopted the 2030 Agenda for sustainable development, which includes seventeen Sustainable Development Goals, among them the 10th Goal aims to reduce inequalities. Convinced of the importance of this goal, in this paper we propose to study the socio-economic determinants which affect the inequalities among the 20 Italian regions by applying a suitable regression model. The socio-economic literature suggests that the most important determinants of inequalities are government spending, income, employment and educational attainment, so we focus our attention on the indicators of the Sustainable Development Goals related to these determinant factors. Given that the number of indicators is extremely high, while the number of observations is low, we consider the partial least squares regression as the most suitable statistical methodology to deal with this dependence modeling.

Keywords Inequalities · Sustainable development goals · Multicollinearity · Partial least squares regression

1 Introduction

The 2030 Agenda for Sustainable Development is a global plan adopted by the United Nations in 2015. It includes 17 Sustainable Development Goals (SDGs) that cover a wide range of issues whose main objective is the achievement of a *sustainable and inclusive development*. The 17 SDGs cover three macro areas which are

✉ Ida Camminatiello
ida.camminatiello@unicampania.it

Rosaria Lombardo
rosaria.lombardo@unicampania.it

Mario Musella
mario.musella2@unina.it

Gianmarco Borrata
gianmarco.borrata@unina.it

[1] Economics Department, University of Campania, Corso Gran,Priorato di Malta, 81043 Capua, CE, Italy

[2] Department of Social Sciences, University of Naples Federico II, Vico Monte della Pietà, 80138 Napoli, Italy

Published online: 12 June 2023

- Social area: combating poverty, reducing economic and social inequalities and access to basic education and healthcare.
- Economic area: promoting innovation and economic growth, creating decent jobs and increasing the resilience of economies.
- Environmental Area: conserving biodiversity, combating climate change and protecting ecosystems.

The 2030 Agenda is designed to be an universal and global action plan to address the challenges of the 21st century and aims to achieve a series of objectives by 2030. It has also been adopted by all member countries of the United Nations and represents a global commitment to promoting sustainable and fair development for all.

The definition of *sustainable and inclusive development* is highly debated in literature. The classical definition of *sustainable development* was first given by the Brundtland Commission in 1987 which says that the sustainable development is "the development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (Brundtland et al., 1987). *Inclusive development*, on the other hand, refers to the economic growth that ensures fair distribution across society and creates opportunities for all (OECD, 2018). For a detailed description of the topic refer to Kamran et al. (2023).

Therefore, the *sustainable and inclusive development* aims to create a growth model that is fair when guaranteeing the access to essential resources and opportunities for everyone, while improving the quality of life and preserving natural resources for future generations (Griessler & Littig, 2005).

In literature (Smith, 1937), inequalities have been defined as "differences in the distribution of income and wealth, as well as economic and social opportunities, among different groups within a society".

Reducing *economic and social inequalities* is one of the most pressing challenges that countries around the world should address for the promotion of the sustainable and inclusive development.

Inequalities can be based on a variety of factors, including race, ethnicity, gender, sexuality, class, or disability. Inequalities can have a negative impact on people's well-being, social cohesion and stability of societies. Specifically, *economic inequalities* refers to differences in wealth and income between individuals or groups. For example, they can be measured through the gap between the highest and lowest incomes within a society (Lanza, 2015). *Social inequalities*, instead, refers to differences in access to rights, opportunities, and services between individuals or groups. They may be caused by factors such as gender, ethnicity, sexual orientation, or physical ability; for a comprehensive review of social inequalities, see Neckerman (2004).

In brief, economic and social inequalities are a complex and multi-faceted phenomenon which can affect people's lives, their health, education, work, and overall well-being. They can also have negative consequences on social cohesion and stability of societies and can limit a society's potential for development, preventing many individuals from reaching their full potential.

According to a quote by Nelson Mandela

There can be no justice when opportunities are not equally distributed.

However, despite efforts to promote equality, such inequalities persist in all societies and represent an impending challenge. Therefore, the issue of reducing economic and social inequalities is an important matter that should be addressed at a global, national and/or local level.

At a local level, various organizations and groups, such as local authorities, non-governmental organizations, local communities, and citizens can promote initiatives and programs to reduce economic and social inequalities.

At a national level, the governments of the world have adopted various initiatives and programs to reduce economic and social inequalities within their regions.

At a global level, the 2030 Agenda for Sustainable Development represents one of the main initiatives to reduce economic and social inequalities.

Here, our aim is to study at a global level the main determinants of the socio-economic inequalities in Italy, highlighting similarities and differences between north, centre and south regions, in order to provide a decision-making tool for policy-makers. Therefore, we briefly describe some of the Sustainable Development Goals on which the 2030 Agenda is based. In this paper, among the various goals of the 2030 Agenda, we focus our attention on Goal 10, namely "reducing inequalities within and among regions", and on Goals 3, 4 and 8 which refer to the socio-economic determinants of inequality.

Every country in the world is required to contribute to addressing these major challenges towards a sustainable path, by developing its own *National Strategy for Sustainable Development*.

In Italy, the National Sustainable Development Strategy (NSDS) was presented to the Council of Ministers and officially approved by the Inter-ministerial Committee for Economic Programming in 2017. It was prepared by the Italian Ministry of Environment Land and Sea in consultation with the Ministry of Foreign Affairs and International Cooperation and all line Ministries, including other national authorities (National Statistical Institute—ISTAT, National Institute for Environmental Research—ISPRA, etc.). The NSDS aims to be the strategic framework for guiding the implementation of the 2030 Agenda in Italy and represents the national reference framework for planning and programming new policies. It represents a step forward in equipping Italy with a governance structure for the 2030 Agenda, a tool that will enable the government to promote equitable and sustainable well-being through the definition of new approaches and policies (ATC, 2015).

To study how the different determinants impact economic inequality between the different Italian regions, we considered the data produced by ISTAT regarding the 2030 Agenda. The data consists of simple indicators related to the seventeen goals.

To know what are the most important socio-economic variables/indicators which influence inequality, we propose to consider a suitable dependence model, i.e. Partial Least Squares (PLS) regression. PLS regression is a statistical technique that is used to predict one or more response variable based on one or more predictor variables. It is similar to multiple linear regression, but it is particularly useful in cases where the number of predictor variables is large with respect to the number of observations.

The paper is structured as follows. Section 2 concerns a brief overview of the literature on the measurement of economic inequalities and how they are measured in Goal 10 of the 2030 Agenda. Section 3 involves an analysis of the socio-economic determinants of inequalities, measured in Goals 3, 4 and 8. Section 4 describes the statistical methodology. Section 5 presents the model for evaluating the inequalities and some conclusion remarks are made in Sect. 6.

2 Inequalities in Sustainability

Reducing socio-economic inequalities is a crucial factor for the sustainable and inclusive development. In this section, we give a brief literature review on the main methodologies for measuring and predicting socio-economic inequalities.

To measure inequality, the Gini index (Gini, 1912) is still one of the most widely used indicators in the literature. Despite its limitations in obtaining estimates that vary depending on individual aversion to inequality (Atkinson, 1970), several authors such as Kakwani (1980), Yitzhaki (1983), and Donaldson and Weymark (1983) have developed extended versions of the Gini index. For a complete discussion of the Gini index, see Farris (2010). Apart from Gini's index, one may consider the Theil index (Theil, 1967), the Gamma index (Lorenz, 1905), the Gini-Simpson index (Simpson, 1949), the Generalized Entropy Index (Shorrocks, 1984, 1994, 1999), and the Oaxaca-Blinder decomposition (Oaxaca, 1973; Blinder, 1973).

To predict inequality many researchers have used a quantile regression model. For example, Piketty and Saez (2003) used quantile regression to study the distribution of income in the United States. Lynch et al. (2004) used quantile regression to study the relationship between income and self-rated health in the United States.

However, quantile regression like any other ordinary least squares model can suffer from multicollinearity which occurs when two or more predictor variables in a model are highly correlated with each other. Therefore, it is important to check for multicollinearity before fitting a regression model, and to use appropriate techniques if it is present.

Here, we consider partial least squares regression which can handle correlated predictor variables appropriately. We study the dependence relationship between the simple indicators of the Goal 10 and the indicators of the socio-economic determinants (which refer to Goals 3, 4 and 8, and are described in Sect. 3), collected on the 20 Italian regions. The response variable, i.e. the Goal 10 "Reduce inequality within and among regions" is composed of 10 targets, each measured by one or more indicators. However, note that we take into consideration only those indicators whose data are available in the most recent year, i.e. 2019. Table 1 shows the considered target and the associated indicator, while the first two rows of Table 2 concern their label and a brief description. For further information on the construction and composition of indicators, see "SDGs 2021 Report: Statistical information for the 2030 Agenda in Italy" published by ISTAT (https://www.istat.it/it/benessere-e-sostenibilit%C3%A0/obiettivi-di-sviluppo-sostenibile/gli-indicatori-istat).

Table 1 Description of Goal 10 and its indicators, data available for 2019

Goal 10	Indicator description
10.1 By 2030 progressively achieve and sustain income growth of the bottom 40% of the population at a rate higher than the national average	10.1.1 Growth rates of household expenditure or income per capita among the bottom 40% of the population and the total population
10.2 By 2030 empower and promote the social, economic and political inclusion of all, irrespective of age, sex, disability, race, ethnicity, origin, religion or economic or other status	10.2.1 Proportion of people living below 50% of median income, by sex, age and persons with disabilities

Table 2 The SDGs indicators concerning the Goals 10, 3, 5 and 8 of the 2030 Agenda

Var	Goal	Label	Indicator
Y1	10	Income	Gross disposable income per capita
Y2	10	Poverty	Risk of poverty
X1	3	HealthWellB1	Probability of death under age 5
X2	3	HealthWellB2	Excess weight (standardized rates); healthy life expectancy at birth
X3	3	HealthWellB3	Diabetes (standardized rates); arterial hypertension (standardized rates); percentage of deliveries with more than 4 check-ups during pregnancy; day-hospital beds in public and private healthcare institutions; Beds in ordinary hospitalization in public and private healthcare institutions
X4	3	HealthWellB4	Dentists; pharmacists; nurses and midwives; doctors
X5	4	QualityEdu1	Inadequate literacy skills (class II secondary school students; first grade secondary school class III students; class V second grade secondary school students; class II secondary school students); Inadequate numerical skills (first grade secondary school class III students; class V second grade secondary school students); Inadequate listening and reading comprehension of the English language (first grade secondary school class III students; class V second grade secondary school students)
X6	4	QualityEdu2	Early exit from the education and training system
X7	4	QualityEdu3	Places authorized in socio-educational services (nursery schools and supplementary services for early childhood) for 100 children aged 0–2
X8	4	QualityEdu4	Rate of participation in educational activities (preschool and first year of primary school) for 5-year-old
X9	4	QualityEdu5	Pupils with disabilities (kindergarten, primary school, lower secondary school, upper secondary school); participation in continuing education
X10	4	QualityEdu6	At least basic digital skills; High digital skills
X11	4	QualityEdu7	Graduates and other tertiary qualifications (ages 30–34); graduates in technical-scientific disciplines (STEM)
X12	4	QualityEdu8	Physically accessible schools; schools with pupils with disabilities due to the presence of adapted computer workstations (in primary school, lower secondary school, upper secondary school); physically inaccessible schools
X13	8	WorkGrowth1	Annual growth rate of real GDP per employee; annual growth rate of value added in volume per employee; Annual growth rate of value added in volume per hour worked
X14	8	WorkGrowth2	Employed in fixed-term jobs for at least 5 years; involuntary part-time; Unemployment rate; non-participation rate at work; employment rate (20–64 years)

Table 2 (continued)

Var	Goal	Label	Indicator
X15	8	WorkGrowth3	Young people who do not work and do not study (NEET); young people who are not working or studying (NEET) (15–24 years)

3 Inequalities Determinants

Socio-economic inequalities can be determined by a large number of variables/indicators, such as income, education, employment, family composition, gender, ethnicity, and access to quality healthcare services (Mackenbach et al., 2002). For example, people with lower incomes are more likely to suffer from chronic diseases such as diabetes and cardiovascular disease due to poor diet, sedentary lifestyle, and lack of access to quality healthcare services (Timmis et al., 2022).

To address socio-economic inequalities, it is crucial to implement strategies that target the reduction of income and education disparities and enhance the availability of high-quality healthcare services for all individuals. This can be accomplished through various measures, such as: (a) elevating minimum wages; (b) investing in education and training; (c) improving access to quality healthcare services; (d) promoting healthy lifestyles. Additionally, it is important to raise public awareness about the importance of reducing health inequalities and to promote community participation in the formulation and implementation of healthcare policies.

Indeed, education is a crucial factor in reducing inequality globally. Over the past few decades there have been advances in increasing access to education worldwide, but significant inequalities persist in access to quality education among social, economic and geographic groups. Educational inequality can have negative long-term consequences on individuals' and communities' personal, social and economic development. To reduce these inequalities globally, systemic and targeted interventions are needed at the social, economic and political levels (Schmidt et al., 2015).

The issue of decreasing joblessness among young people is a relatively new topic on the global stage, but it is a crucial problem for many nations around the world. The number of people in the age group 15–24 who are not working, studying or receiving training has become alarmingly high and this has led to the creation of a new term, "NEET" (Not Education Employment Training). This term refers not just to a social category but also to the individuals themselves who fit into that category.

According to the analysis conducted by Caroleo et al. (2020), countries with lower economic development tend to have higher rates of NEET. Furthermore, in countries where the school-to-work transition institutions are more developed, the proportion of NEET individuals is lower.

The International Monetary Fund (IMF) argues that economic inequality can have negative effects on long-term economic growth and financial stability, as well as on social cohesion and on democracy quality.

Three instrumental reasons for pursuing economic policies that engender less income inequality are (Birdsall, 2001)

- Inequality can inhibit growth and slow poverty reduction.
- Inequality often undermines the political process: this may lead to an inadequate social contract and may trigger bad economic policies-with ill effects on growth, human development, and poverty reduction.
- Inequality can undermine civil, social, and political life and inhibit certain forms of collective decision-making.

In this paper, to study the determinants of socio-economic inequalities we emphasize the indicators of three SDGs of the Italian NSDS, in particular

- Goal 3: "Ensure healthy lives and promote well-being for all at all ages". It is composed of 10 targets and 14 indicators which aim to measure progress towards the Goal of ensuring healthy lives and promoting well-being for all. In particular, it aims to ensure that everywhere people can enjoy good health and physical and mental well-being. It is expected to be achieved through a range of actions, expanding healthcare systems and promoting healthy lifestyles, such as regular physical activity and healthy diet. It is also expected to face emerging challenges such as the obesity epidemic and non-communicable diseases, and to protect people from epidemics and pandemics;
- Goal 4: "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". It is composed of 8 targets and 12 indicators which aim to measure progress towards the goal of ensuring quality education for all. It aims to ensure that everyone, regardless of their background or circumstances, can access to high-quality education and lifelong learning opportunities. So, this means promoting literacy, at primary and secondary schooling, as well as at tertiary education and vocational training. It also aims to promote gender equality and reduce the education gap between countries and within countries. Achieving this goal will contribute to creating more inclusive and sustainable societies, as well as promoting economic growth and sustainable development;
- Goal 8: "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all, and an enhanced productive capacity for least developed regions". It is composed of 8 targets and 8 indicators which aim to measure progress towards the goal of promoting inclusive and sustainable economic growth. This goal aims to promote inclusive and sustainable economic growth that creates decent and dignified employment opportunities for all. It is expected to be achieved by supporting entrepreneurship and innovation, strengthening entrepreneurial capacities, improving access to credit and infrastructure, and by promoting fair and sustainable international trade. Achieving this goal will contribute to create a stronger and more stable global economy, reducing poverty and inequality, and promoting sustainable development.

The selected indicators of Goals 3, 4 and 8 of the 2030 Agenda, available for the year 2019, are listed in Table 2 of Sect. 5.

4 Methodology

Ordinary least squares (OLS) model allows explaining one or more quantitative response variables in terms of predictors. The model is widely used in many fields, however, it can be affected by the multicollinearity problem, which occurs when two or more of the predictors are highly correlated. This can lead to unreliable and unstable estimates of the regression coefficients, as well as large standard errors on these estimates. It can also make hard to determine the unique effect of each predictor on the response variable (Camminatiello et al., 2017).

In presence of multicollinearity, the stepwise selection of the predictors could be performed. *However, the stepwise procedure will tend to choose the predictors that best match the data sample, and thus to favor variables that, by the luck of the draw of the sample, happen to have high-magnitude coefficients for that sample but not necessarily for the underlying population. Therefore,* Ridge regression estimator (Hoerl & Kennard, 1970) and its modifications,

continuum regression (Stone & Brooks, 1990), least absolute shrinkage regression (Tibshirani, 1996), partial least squares (PLS) regression (Wold, 1966, 1975, 1985, 1978), principal components regression and latent root regression (Webster et al., 1974) have been proposed as alternative tools for facing multicollinearity.

Among these methods we consider PLS for its mathematical and statistical properties Tenenhaus (1998); Frank et al. (1993).

4.1 PLS Regression

PLS regression has been originally developed by Wold (1966) and is also known as Nonlinear Iterative Partial Least Squares (NIPALS). There are several different variants of the initial PLS algorithm (Wold, 1966, 1975), including one called SIMPLS (De Jong, 1993).

PLS regression has several advantages over traditional OLS regression (Durand, 2001), indeed it can handle

- a large number, p, of predictor variables and a small number, n, of observations;
- correlated predictor variables;
- missing data;
- both continuous and categorical response variables;
- both continuous and categorical predictor variables.

In PLS regression, the columns of the matrices $\mathbf{X}$ and $\mathbf{Y}$ representing the p independent and q dependent variables, respectively, are centered and scaled to have zero mean and unit length. The first step in the PLS procedure involves carrying out uncorrelated latent variables. These latent variables, namely PLS components, are linear combinations of the original independent variables, which maximize the covariance between the independent and dependent variables (Helland, 1988).

After creating the latent variables, a least squares regression is performed using a subset of the extracted variables. The PLS regression model can be written as

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\Gamma} + \mathbf{F} \tag{1}$$

where $\mathbf{T}$ is the $n \times a$ matrix of the PLS components, $\boldsymbol{\Gamma}$ is the $a \times q$ matrix of the PLS regression coefficients, and $\mathbf{F}$ is the $n \times q$ matrix of the disturbances.

PLS regression leads to biased but lower variance estimates of the regression coefficients compared to OLS regression. Goutis (1996) showed that the estimates of PLS coefficients are always shrunken compared to OLS ones, that is

$$\|\hat{\boldsymbol{\Gamma}}_1\| \leq \|\hat{\boldsymbol{\Gamma}}_2\| \leq \cdots \leq \|\hat{\boldsymbol{\Gamma}}_a\| \leq \ldots \|\hat{\mathbf{B}}\| \tag{2}$$

where $\hat{\boldsymbol{\Gamma}}_a$ is the PLS estimator obtained after the extraction of the a components and $\hat{\mathbf{B}}$ is the OLS matrix of coefficient estimators.

For assessing the goodness of fitting of the PLS model, the determination coefficient R^2 is commonly used, while for measuring the predictive ability of the model, among different criteria, we consider the *PRESS* (PRediction Error Sum of Squares) statistic.

4.2 Model Calibration and Validation

The number a of latent variables to be retained in the model can be selected according to different criteria. The most common ones have been based upon test set and Cross Validation (CV) or its generalization (Lombardo et al., 2009). By using a test set, the criterion is given by the Residual Sum of Squares (RSS) related to each PLS component. Although this approach is fast, as we run the PLS algorithm once for each component, it is not often considered, because a test set, independent from the training set, is only exceptionally available. Differently, when using cross-validation (Wold, 1978) the criterion being used is the *PRESS* which evaluates the accuracy and predictive power of a model. The cross-validation involves dividing a dataset into multiple subsets, say l, where each subset (or fold) is used as both a training set and a validation set. To determine the optimal number of latent variables using *PRESS*, start with a minimum number of latent variables, such as 1. Use cross-validation (e.g., l-fold cross-validation) to fit the model with the chosen number of latent variables. Calculate the *PRESS* for each omitted subset, and then restore the omitted subset. The procedure is repeated for the number h (for $h = 1, 2, ...a$) of latent variables until each subset has been left out once. The l individual *PRESS*'s are summed up to give the $PRESS_k$ of each response variable. Finally, the $PRESS_{total}$ is calculated by summing up the $PRESS_k$, for $k = 1, .., q$. The number a of latent variables that corresponds to the lowest $PRESS_{total}$ represents the optimal choice in terms of minimizing prediction errors.

Here we perform a leave-one-out-cross validation, therefore $l = n$, and for each h the $PRESS_k$ related to the response k can be computed as

$$PRESS_k = \sum_{i=1}^{n} (y_{ik} - \hat{y}_{-ik(h)})^2 \tag{3}$$

where the predicted values $\hat{y}_{-ik(h)}$ are based on the PLS estimates (with h components) obtained by removing the ith subset.

4.3 Bootstrap Confidence Interval

Bootstrapping is a useful method for estimating the sampling distribution of the coefficients in partial least squares regression model when the assumption of normality is not met or when the sample size is small.

In partial least squares regression, a sample of predictor and response variables is used to fit the model. The model is then used to predict the response for each predictor in the sample, and the prediction errors are used to estimate the coefficients of the model. The bootstrap procedure involves sampling with replacement from the original sample and fitting the PLS model to the resampled data (Tenenhaus & De Jong, 2003; De Jong & Wieringa, 2002). This is repeated many times to create a distribution of estimates for the coefficients, which can be used to compute standard errors and confidence intervals (Magnanensi et al., 2017; Knecht & Hautle, 2013; Naes & Naes, 2005).

In the programming environment R, there are numerous packages and functions (for example, see www.jf-durand-pls.com) that perform PLS. We consider the bootpls and pls packages to perform bootstrapped partial least squares regression.

For constructing confidence intervals from bootstrapped samples, the percentile method has been used. It involves computing the percentiles of the bootstrapped

samples that correspond to the desired confidence level. For example, to compute a 95% confidence interval, the 2.5th and 97.5th percentiles of the bootstrapped samples are taken as the lower and upper limits of the interval.

5 Evaluating Inequalities in Sustainability through PLS Regression

The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, provides a shared blueprint for peace and prosperity for people and the planet, now and for the future. At its heart are the 17 Sustainable Development Goals (SDGs), which represent an urgent call for action by all countries—developed and developing—in a global partnership. They recognize that ending poverty and other deprivations must go hand-in-hand with strategies that improve health and education, reduce inequality, and spur economic growth—all while tackling climate change and working to preserve our oceans and forests (https://sdgs.un.org/Goals).

As highlighted in Sect. 3, here we aim to evaluate the most important determinants of inequalities across the Italian regions, looking at the ISTAT indicators of the SDGs (https://www.istat.it/it/benessere-e-sostenibilit%C3%A0/obiettivi-di-sviluppo-sostenibile/gli-indicatori-istat). For the sake of data explanation, we consider the variables listed in Table 2 related to Goals 10, 3, 5 and 8 of the 2030 Agenda. The variables/indicators of Table 2 have been collected on the 20 Italian regions, i.e. *Piemonte, Valle d'Aosta, Liguria, Lombardia, Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna*. Some of them are simple indicators and others are composite indicators (Lafortune et al., 2018; Alaimo et al., 2021a; Alaimo & Maggino, 2020; OECD, 2008). For each composite indicator, we checked the reliability using the Cronbach's Alpha and they all were consistent, showing an alpha index greater than 0.7 (Table 3). *By definition, Cronbach's alpha is used to evaluate internal consistency for indicators measured by more than one variable. Therefore, no Cronbach's alpha was computed for the indicators HealthWellB1, QualityEdu2, QualityEdu3, and QualityEdu4 as they are single-item indicators.* These simple indicators have been normalized.

Table 3 Cronbach's Alpha for composite Indicators

Composite indicator	Value
HealthWellB2	0.79
HealthWellB3	0.78
HealthWellB4	0.71
QualityEdu1	0.99
QualityEdu5	0.74
QualityEdu6	0.97
QualityEdu7	0.70
QualityEdu8	0.77
WorkGrowth1	0.90
WorkGrowth2	0.95
WorkGrowth5	0.99

Table 4 PLS regression performance

PLS model				
dimension	Income	Poverty	% Var	Cum Var
1	0.808	0.912	85.998	85.998
2	0.103	0.023	6.289	92.288
3	0.001	0.037	1.932	94.220

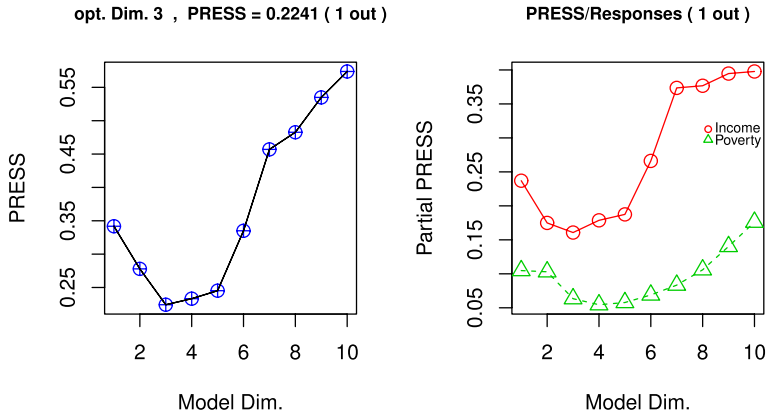

Fig. 1 The PRESS criterion for the choice of the number of PLS components

5.1 The PLS Model

A very desirable condition in a set of regression data is that there is no multicollinearity among the predictors included in the model. For this reason we calculate the condition index (*CI*) that allows us to check the level of collinearity. The condition index is given by

$$CI = \sqrt{\lambda_{max}/\lambda_{min}}$$

where λ_{max} and λ_{min} are the maximum and minimum eigenvalues, respectively, of the correlation matrix among predictors.

For our predictors listed in Table 2, the *CI* is equal to 79.36, which indicates a very strong level of collinearity (Lucadamo et al., 2021). Therefore the use of the PLS regression is surely more suitable for getting a reliable parameter estimation.

We consider the multivariate PLS model where the responses are the two simple indicators of Goal 10 listed in the first two rows of Table 2 (see also Table 1), related to *Income* and *Poverty* condition. The correlation between them is negative and very high equal to −0.89.

The choice of the suitable number of PLS components has been made by using the PRESS criterion (Sect. 4.2). It results that the predictive ability of the model is excellent for each dependent variable, we get that $PRESS_{total} = 0.22$, and in detail $PRESS_{Income} = 0.16$ and $PRESS_{Poverty} = 0.06$, see also Fig. 1. After determining the suitable number of PLS components, we perform the PLS model with three components which explain 65.23% of **X** variability and the 94.21% of **Y** variability. Table 4 shows the percentage of **Y** variability

with respect to each model dimension. Furthermore, it results that model fits each response very well, i.e. $R^2_{Income} = 0.91$ and $R^2_{Poverty} = 0.97$.

When performing a PLS regression, it is possible to visualize the relationship between the responses, the predictors and the latent variables by using a graph called the PLS loading plot which is a correlation circle and allows to study the correlation among original and latent variables. Also, to portray the relationship between the observations and the PLS components one can consider the PLS score plot. Being the percentage of explained variance by the first two components very high, equal to 92.28 (see Table 4), we consider only two components for visualizing the results in Fig. 2. The left side of Fig. 2 shows the PLS loading plot on the plane t_1, t_2. While the right side of Fig. 2 shows the PLS score plot. Looking at the PLS loading plot of Fig. 2, we can observe a high positive correlation among:

- the response *Poverty* (risk of poverty) and the predictors *QualityEdu1* (inadequate literacy, numerical and English skills), *WorkGrowth2* (unemployment/employment rate) and *WorkGrowth3* (young people who do not work and not study-NEET).
- the predictors *QualityEdu2* (early exit from the education and training system) *QualityEdu4* (rate of participation in educational activities for 5-year-olds), and *HealthWellB1* (under-5 mortality rate);
- the response *Income* (gross disposable income per capita) and the predictors *QualityEdu3* (places authorized in socio-educational services per 100 children aged 0–2) and *QualityEdu6* (digital skills);
- the predictors *QualityEdu7* (graduates and other tertiary qualifications) and *HealthWellB4* (dentists, pharmacists, nurses and midwives, doctors).

Moreover the correlation circle of Fig. 2 shows a high negative correlation between the two responses—*Income* (gross disposable income per capita) and *Poverty* (risk of poverty). Note that the response *Poverty* is negatively correlated with *QualityEdu3* (places authorized in socio-educational services per 100 children aged 0–2) and *QualityEdu6* (digital skills), while the response *Income* is inversely related to *QualityEdu1* (inadequate literacy, numerical and English skills) *WorkGrowth2* (unemployment/employment rate), and *WorkGrowth3* (young people who do not work and do not study-NEET).

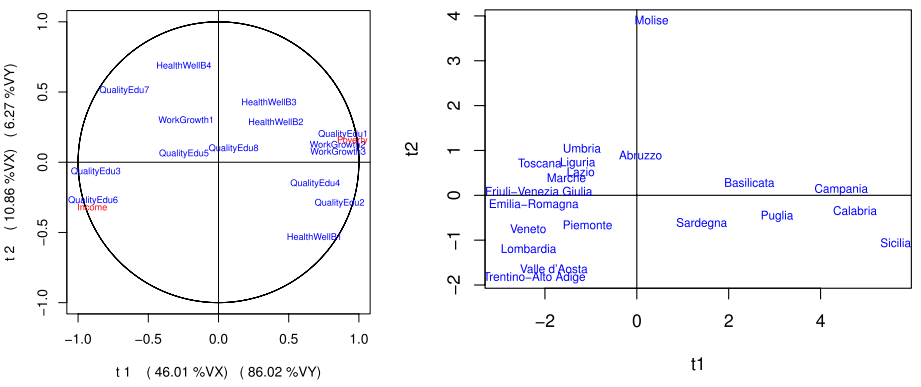

Fig. 2 The PLS plots on the components t_1, t_2. On the left side, the PLS loading plot. On the right side, the PLS score plot

Furthermore, *QualityEdu2* (early exit from the education and training system), *QualityEdu4* (rate of participation in educational activities for 5-year-olds) and *HealthWellB1* (under-5 mortality rate) are negatively correlated with *QualityEdu7* (graduates and other tertiary qualifications) and *HealthWellB4* (dentists, pharmacists, nurses and midwives, doctors) (Fig. 2).

Looking at the observation or score plot of the right side of the Fig. 2, we can observe that there are many differences across regions with respect to the response variables, i.e. *Income* and *Poverty*. In particular, we can see that

- *Campania* and *Basilicata* are characterised by *Poverty* (risk of poverty), mainly due to *QualityEdu1* (inadequate literacy, numerical and English skills), to *WorkGrowth2* (unemployment/employment rate), and to *WorkGrowth3* (young people who do not work and do not study- NEET);
- *Puglia, Calabria, Sicilia* and *Sardegna* are characterised by *QualityEdu2* (early exit from the education and training system), by *QualityEdu4* (the high rate of participation in educational activities for 5-year-olds), and by *HealthWellB1* (under-5 mortality rate);
- *Emilia-Romagna, Friuli-Venezia Giulia, Lombardia, Piemonte, Trentino-Alto Adige, Veneto* and *Valle d'Aosta* present high values of *Income*, *QualityEdu3* (number of places authorized in socio-educational services per 100 children aged 0–2), and of *QualityEdu6* (good digital skills);
- *Lazio, Liguria, Marche, Umbria, Molise, Abruzzo* and *Toscana* are principally characterised by *QualityEdu7* (graduates and other tertiary qualifications) and by *HealthWellB4* (dentists, pharmacists, nurses and midwives, doctors).

To assess the relevance of these predictors we look at Table 5, Figs. 3 and 4 that show the corresponding PLS coefficients. They are computed from three extracted components. As they relate to the centered and scaled data, they express the strength of the relationship between the two dependent variables and all the significant variables in the model. It results that the most important predictor of *Income* is *QualityEdu6*

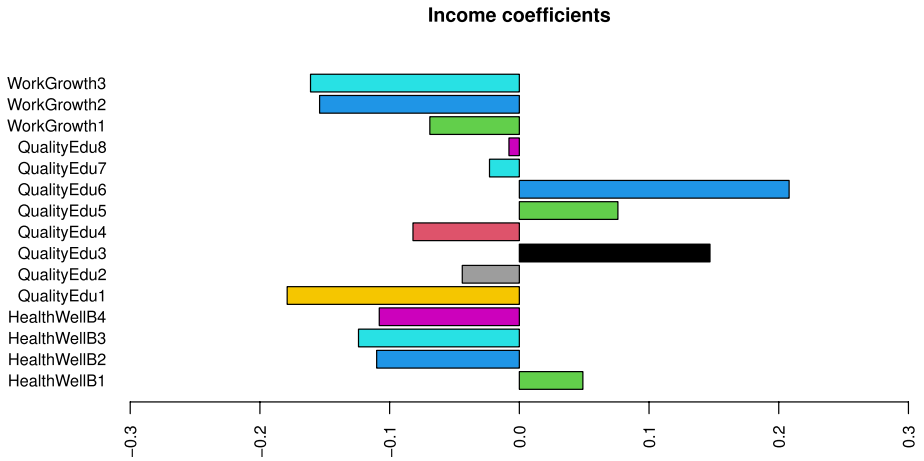

Fig. 3 The PLS coefficients for the Income response

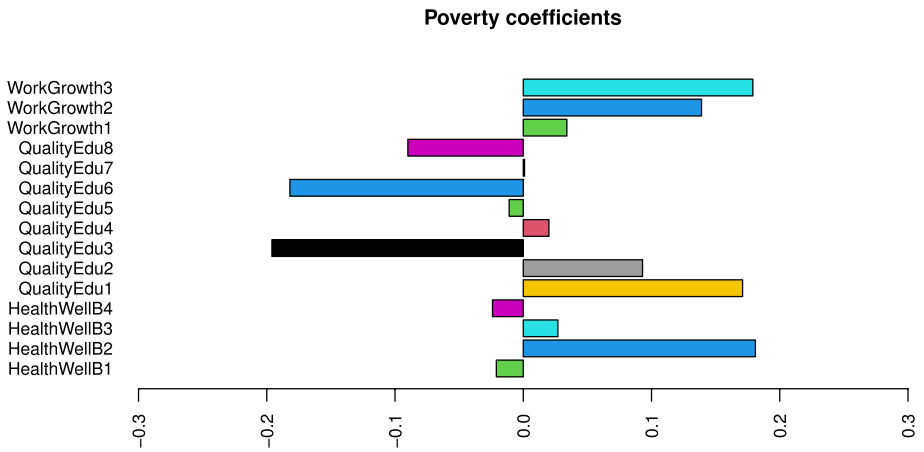

Fig. 4 The PLS coefficients for the Poverty response

Table 5 Estimates of the PLS regression coefficients for the two responses

PLS coefficients		
Label	Income	Poverty
HealthWellB1	0.049	− 0.021
HealthWellB2	− 0.110	**0.181**
HealthWellB3	− 0.124	0.027
HealthWellB4	− 0.109	− 0.026
QualityEdu1	**− 0.180**	**0.171**
QualityEdu2	− 0.044	**0.093**
QualityEdu3	**0.147**	**− 0.196**
QualityEdu4	**− 0.083**	0.020
QualityEdu5	0.076	− 0.010
QualityEdu6	**0.208**	**− 0.182**
QualityEdu7	− 0.023	0.001
QualityEdu8	− 0.008	− 0.090
WorkGrowth1	− 0.068	0.034
WorkGrowth2	**− 0.154**	0.139
WorkGrowth3	**− 0.161**	**0.179**

Highlighted in bold the statistical significant coefficients

(digital skills) which is related positively (i.e. the higher *QualityEdu6* is, the higher the *Income* is). The second, third and fourth predictor in order of importance are *QualityEdu1* (inadequate skills), *WorkGrowth3* (young people who do not work and do not study- NEET) and *WorkGrowth2* (unemployment/employment rate), respectively, but the estimates of their regression coefficients are negative indicating an inverse relationship with the response (i.e. the higher the indicators *QualityEdu1*, *WorkGrowth3* and *WorkGrowth2* are, the lower the *Income* is). It follows *QualityEdu3* (places authorized in socio-educational services) that has a positive effect on *Income*.

After bootstrapping the model, it results that all these predictors are statistically significant within a 95% confidence interval. In Table 6, the coefficient values of the significant predictors are highlighted in bold, while the coefficient values which do not result statistically significant are highlighted in italics.

Looking at the most important and significant predictors, broadly speaking we can conclude that education and employment play a very important role in increasing income across the Italian regions.

Differently, the two most important predictors of the response *Poverty* are *QualityEdu3* (places authorized in socio-educational services) and *QualityEdu6* (digital skills) which are related negatively (i.e. the higher the indicators *QualityEdu3* and *QualityEdu6* are, the lower the *Poverty* is). The next five most important predictors of *Poverty* are *HealthWellB2* (excess weight, healthy life expectancy at birth), *WorkGrowth3* (young people who do not work and do not study-NEET), *QualityEdu1* (inadequate literacy, numerical and English skills), *WorkGrowth2* (unemployment/employment rate) and *QualityEdu2* (early exit from the education and training system). All these predictors are statistically significant within a 95% confidence interval. Indeed, Table 6 shows the quantile confidence intervals of the PLS regression coefficients from bootstrapped samples of the responses and predictor variables It is evident that for the two responses of Goal 10 the statistically significant predictors are not the same, however in common between them there are some indicators belonging to Goal 4, i.e. *QualityEdu1* (inadequate skills), *QualityEdu3* (places authorized in socio-educational services), *QualityEdu6* (digital skills), and some belonging to Goal 8, i.e. *WorkGrowth2* (unemployment/employment rate) and *WorkGrowth3* (young people who do not work and do not study-NEET). These predictors principally characterise two Italian south regions for high values, i.e. *Basilicata* and *Campania* and for low values *Friuli-Venenzia-Giulia,*

Table 6 Confidence intervals of the PLS regression coefficients for the two responses

Label	Income		Poverty	
	Lower bound	Upper bound	Lower bound	Upper bound
HealthWellB1	*– 0.071*	*0.098*	*– 0.048*	*0.102*
HealthWellB2	*– 0.189*	*0.016*	**0.054**	**0.233**
HealthWellB3	*– 0.173*	*0.013*	*– 0.042*	*0.118*
HealthWellB4	*– 0.176*	*0.015*	*– 0.084*	*0.050*
QualityEdu1	**– 0.208**	**– 0.114**	**0.105**	**0.193**
QualityEdu2	*– 0.100*	*0.036*	**0.041**	**0.137**
QualityEdu3	**0.074**	**0.179**	**– 0.222**	**– 0.102**
QualityEdu4	**– 0.191**	**– 0.019**	*– 0.014*	*0.112*
QualityEdu5	*– 0.021*	*0.222*	*– 0.087*	*0.077*
QualityEdu6	**0.150**	**0.230**	**– 0.201**	**– 0.111**
QualityEdu7	*– 0.090*	*0.072*	*– 0.077*	*0.024*
QualityEdu8	*– 0.100*	*0.119*	*– 0.112*	*0.022*
WorkGrowth1	*– 0.182*	*0.036*	*– 0.073*	*0.074*
WorkGrowth2	**– 0.175**	**– 0.090**	**0.095**	**0.160**
WorkGrowth3	**– 0.181**	**– 0.100**	**0.117**	**0.189**

The coefficient values of the significant predictors are highlighted in bold, while the coefficient values which do not result statistically significant are highlighted in italics

Toscana, Marche, Lazio, Liguria and *Umbria*. Note that only for the response *Poverty* the predictor *HealthWellB2* (excess weight, healthy life expectancy at birth) related to Goal 3, results statistically significant. This predictor variable well describes *Basilicata* and *Campania* regions.

6 Conclusion

Socio-economic inequalities are present in various levels in all European countries and available data suggest that the income and health gap is increasing. Many studies have been conducted to explain socio-economic inequalities and much has been learned about the various factors that underlie them.

In this paper, we have investigated inequalities across Italian regions modeling the dependence relationship through the multivariate Partial Least Squares regression model which can be suitable to analyse a data set of simple and composite indicators (predictors highly correlated). As in literature, a great debate concerns the construction of such composite indicators (Fattore, 2017; Alaimo et al., 2021a, b), a comprehensive discussion of the variants of aggregation of simple indicators and their effects when modeling shall be left for future consideration.

Here the inequality has been described by the two simple indicators related to *Income* and *Poverty* (responses of the multivariate model) and different determinants have been considered to explain and predict such indicators. Doing so, it results very relevant to improve literacy, numerical and English skills (*QualityEdu1*), to increase the places authorized in socio-educational services for children aged 0–2 (*QualityEdu3*), to improve digital skills (*QualityEdu6*), to increase the employment rate of young people (*WorkGrowth2*) and decrease the number of young people that do not study and do not work (*WorkGrowth3*), specially for *Campania, Basilicata, Puglia, Calabria, Sicilia and Sardegna*.

In general, knowing how different (or not) the inequality is across Italian regions can help policy-makers in deciding what could be a suitable reform to introduce in each region of Italy. Such reforms should require focusing on the reorganisation of the education system or on social operators and their incentives.

Also this study highlights that Italy can be divided into three distinct macroregions, each exhibiting distinctive characteristics. Northern Italy and some regions of the Centre show higher values than the national average across many indicators of equality, including income, employment, educational opportunities, and health. Some central regions have indicators that fall around the national average. While most of the southern regions experience living conditions significantly below the national average, with limited employment and educational opportunities.

The regional divide is a phenomenon that joins several other European countries besides Italy (like Sweden, Finland, Romania, and Estonia).

The severity of regional disparities in Italy, particularly about the labor market, necessitates the implementation of territorial development policies which focus on providing opportunities to people residing in the southern regions and leveraging untapped potential (Fina et al., 2023). In this study, we propose a reorientation of regional policies to achieve the following objectives:

– Increase public investment in health and education to stimulate short-term economic activity and enhance the potential for long-term economic growth.

– Enhance employment support measures to alleviate the challenges faced by individuals in the southern regions. It is essential to emphasize that regional equality not only contributes to social and political cohesion but also serves as a means to face social inequalities and foster sustainable and inclusive economic development.

Funding Open access funding provided by Università degli Studi della Campania Luigi Vanvitelli within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Agency for Territorial Cohesion. (2015). Agenda 2030 per lo sviluppo sostenibile. https://www.agenziacoesione.gov.it/comunicazione/agenda-2030-per-lo-sviluppo-sostenibile.

Alaimo, L. S., Arcagni, A., Fattore, M., & Maggino, F. (2021). Synthesis of multi-indicator system over time: A poset-based approach. *Social Indicators Research, 157*, 77–99.

Alaimo, L. S., Ciacci, A., & Ivaldi, E. (2021). Measuring sustainable development by non-aggregative approach. *Social Indicators Research., 157*, 101–122.

Alaimo, L. S., & Maggino, F. (2020). Sustainable development goals indicators at territorial level: Conceptual and methodological issues-the Italian perspective. *Social Indicators Research, 147*(2), 383–419.

Atkinson, A. (1970). On the measurement of inequality. *Journal of Economic Theory, 2*(3), 244–263.

Birdsall, N. (2001). Why inequality matters: Some economic issues. *Ethics & International Affairs, 15*, 3–28.

Blinder, A. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources, 8*(4), 436–455.

Brundtland, G. H., Khalid, M., Agnelli, S., Al-Athel, S., & Chidzero, B. (1987). *Our common future*. Report of the World Council for Economic Development.

Camminatiello, I., Lombardo, R., & Durand, J. F. (2017). Robust partial least squares regression models for the evaluation of justice court delay. *Quality & Quantity, 51*, 813–827.

Caroleo, F. E., Rocca, A., Mazzocchi, P., et al. (2020). Being NEET in Europe before and after the economic crisis: An analysis of the micro and macro determinants. *Social Indicator Research, 149*, 991–1024.

De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems, 18*, 251–263.

De Jong, S., & Wieringa, J. A. (2002). Bootstrapping partial least squares regression. *Journal of Chemometrics, 16*(4), 259–268.

Donaldson, D., & Weymark, J. A. (1983). Ethically flexible Gini indices for income distributions in the continuum. *Journal of Economic Theory, 29*(2), 353–358.

Durand, J. F. (2001). Local polynomial additive regression through PLS and splines: PLSS. *Chemometrics and Intelligent Laboratory Systems, 58*, 235–246.

Farris, F. A. (2010). The Gini index and measures of inequality. *The American Mathematical Monthly, 117*(10), 851–864.

Fattore, M. (2017). Synthesis of indicators: The non-aggregative approach. In F. Maggino (Ed.), *Complexity in society: From indicators construction to their synthesis* (pp. 193–212). Springer.

Fina, S., Heider, B., & Prota, F. (2023). Unequal Italy. Regional socio-economic disparities in Italy. Friedrich-Ebert-Stiftung. https://feps-europe.eu/wp-content/uploads/2021/07/Unequal-Italy-Regional-socio-economic-disparities-in-Italy.pdf.

Frank, I. E., Friedman, J. H., Wold, S., Hastie, T., & Mallows, C. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*(2), 109–148.

Gini, C. (1912). *Variabilità e mutuabilità. Contributo allo studio delle distribuzioni e delle relazioni statistiche*. Bologna, C. Cuppini.

Goutis, C. (1996). Partial least squares algorithm yields shrinkage estimators. *The Annals of Statistics, 24*, 816–824.

Griessler, E., & Littig, B. (2005). Social sustainability: A catchword between political pragmatism and social theory. *International Journal for Sustainable Development, 8*, 65–79.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistics—Simulation and Computation, 17*, 581–607.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation of nonorthogonal problems. *Technometrics, 12*, 55–67.

Kakwani, N. C. (1980). On a class of poverty measures. *Econometrica, 48*, 437–446.

Kamran, M., Rafique, M. Z., Nadeem, A. M., & Anwar, S. (2023). Does inclusive growth contribute toward sustainable development? Evidence from selected developing countries. *Social Indicator Research, 165*, 409–429.

Knecht, A., & Hautle, M. (2013). Consistent bootstrapping for partial least squares regression. *Journal of Chemometrics, 27*(6), 274–280.

Lafortune, G., Fuller, G., Moreno, J., Schmidt-Traub, G. & Kroll, C. (2018). SDG index and dashboards detailed methodological paper. Sustainable development solutions network. https://www.sdgindex.org/reports/sdg-index-and-dashboards-2018/

Lanza, G. (2015). *La misurazione della disuguaglianza economica: Approcci, metodi e strumenti*. Franco Angeli.

Lombardo, R., Durand, J. F., & De Veaux, R. (2009). Model building in multivariate additive partial least squares splines via the GCV criterion. *Journal of Chemometrics, 23*, 605–617.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publication of the American Statistical Association, 9*, 209–219.

Lucadamo, A., Camminatiello, I., & D'Ambra, A. (2021). A statistical model for evaluating the patient satisfaction. *Socio-Economic Planning Sciences, 73*, 100797.

Lynch, J., Smith, G. D., Harper, S., Hillemeier, M., Ross, N., Kaplan, G. A., et al. (2004). Is income inequality a determinant of population health? Part 1. A systematic review. *Milbank Quarterly, 82*(1), 5–99.

Mackenbach, J. P., Bakker, M., & Benach, J. (2002). *Reducing inequalities in health: A European perspective*. London: Routledge.

Magnanensi, J., Bertrand, F., Maumy-Bertrand, M., et al. (2017). A new universal resample-stable bootstrap-based stopping criterion for PLS component construction. *Statistical Computing, 27*, 757–774.

Naes, T., & Naes, E. (2005). Bootstrap confidence intervals for PLS regression: A comparison of six methods. *Journal of Chemometrics, 19*, 441–450.

Neckerman, K. (2004). *Social inequality*. Russell Sage Foundation.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review, 14*(3), 693–709.

OECD. (2008). *Handbook on constructing composite indicators methodology and user guide*. Organisation for Economic Cooperation and Development. ISBN 978-92-64-04345-9

OECD. (2018). *Opportunities for all: A framework for policy action on inclusive growth*. Organisation for Economic Cooperation and Development.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics, 118*(1), 1–39.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The Role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher, 44*(7), 371–386.

Shorrocks, A. F. (1984). Ranking income distributions. *Econometrica, 52*, 1369–1380.

Shorrocks, A. F. (1994). Inequality decomposition by population subgroup. *Econometrica, 62*, 1369–1396.

Shorrocks, A. F. (1999). Inequality, mobility and growth. *Journal of Economic Growth, 4*, 63–89.

Simpson, E. H. (1949). Measurement of diversity. *Nature, 163*(4148), 688–688.

Smith, A. (1937). *The wealth of nations*. The Modern Library, Random House Inc.

Stone, M., & Brooks, R. (1990). Continuum regression: Cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society Series B, 52*(2), 237–269.

Tenenhaus, M. (1998). *La régression PLS, théorie et pratique*. Paris: Editions Technip.
Tenenhaus, M., & De Jong, S. (2003). Bootstrapping in partial least squares regression. *Journal of Chemometrics, 17*(3), 253–263.
Theil, H. (1967). *Economic forecasts and policy*. Amsterdam: North Holland Publishing Company.
Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *Journal of Royal Statistical Society Series B, 58*, 267–288.
Timmis, A., Vardas, P., Townsend, N., et al. (2022). European society of cardiology: Cardiovascular disease statistics 2021. *European Heart Journal, 43*(8), 716–799.
Webster, J. T., Gunst, R. F., & Mason, R. L. (1974). Latent root regression analysis. *Technometrics, 164*, 513–522.
Wold, H. (1985). *Partial least squares*. In S. Kotz, & N. L. Johnson (Eds.) *Encyclopedia of statistical sciences* (Vol. 6, pp.581–591). Wiley.
Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 391–420). Academic Press.
Wold, H. (1975). Soft modelling by latent variables: Non linear iterative partial least squares approach. In J. Gani (Ed.), *Perspectives in probability and statistics: Papers in honour of Bartelett* (pp. 117–142). Academic Press.
Wold, S. (1978). Cross-validation estimation of the number of components in factor and principal components analysis. *Technometrics, 24*, 397–405.
Yitzhaki, S. (1983). On an extension of the Gini inequality index. *International Economic Review,* 24(3), 617–628.