Universidad de Granada

Departamento de Métodos Cuantitativos para la Economía y la Empresa

Tesis Doctoral en Ciencias Económicas y Empresariales

Técnicas Cuantitativas Avanzadas en el Ámbito Económico y Empresarial

---

Generalization of the residualization procedure

Properties and environmental applications

Claudia García García

Directores:

Dña. Catalina B. García García

D. Román Salmerón Gómez

Granada, 2020

Editor: Universidad de Granada. Tesis Doctorales
Autor: Claudia García García
ISBN: 978-84-1306-563-2
URI: http://hdl.handle.net/10481/63333

AGRADECIMIENTOS

En primer lugar, mi más sincero agradecimiento a mis Directores, la Dra. Catalina B. García García y el Dr. Román Salmerón Gómez. Con esta gran experiencia he podido saber cuál es mi vocación profesional, y he tenido la suerte de teneros a vosotros para guiarme. Gracias por confiar en mí y darme esta oportunidad, creo que no habría podido tener mejores mentores. Gracias por vuestro apoyo y ayuda constante, por la orientación, por la motivación y por todos vuestros consejos. Habéis sido y seréis un ejemplo a seguir, tanto personal como profesionalmente. Os agradezco el haber estado ahí para todo lo que necesitaba y haber sido tan cercanos conmigo.

En segundo lugar, gracias a la Universidad de Granada por brindarme este reto y permitirme llevarlo a cabo. Y gracias también a todo el Departamento de Métodos Cuantitativos para la Economía y la Empresa.

Mi más cordial agradecimiento a la Universidad del Pireo (Grecia), por darme la oportunidad de desarrollar una parte de mi investigación allí. Concretamente, gracias al *Department of Banking and Financial Management*, y al Dr. Nicholas Apergis, por guiarme durante los meses de mi estancia. En este sentido, mi agradecimiento también a la Universidad de Granada por el apoyo financiero durante los meses en el extranjero a través de la beca de Movilidad Internacional de Estudiantes de Programas de Doctorado durante el curso académico 2018/2019.

Me gustaría expresar mi especial agradecimiento al Dr. José García Pérez (Universidad de Almería), por sus aportaciones a los trabajos desarrollados, por su afabilidad y profesionalidad.

También me gustaría dar las gracias a mis padres, Raquel y Alfredo. Vosotros me dais la fuerza necesaria para cumplir todos mis retos, siempre empujándome a seguir hacia delante y a superar todos los obstáculos que se me ponen por delante, aunque

parezcan imposibles. Gracias por vuestro positivismo y por vuestro esfuerzo, por escucharme, por mantener la calma en los momentos difíciles e influir en mí siempre de la mejor manera. Vosotros me habéis enseñado a contemplar el mundo de una manera realista pero sin dejar de creer en mí misma y en mis metas. Me incitáis y motiváis siempre a mejorar como persona, y a intentar ser la mejor en todo lo que hago. Gracias por vuestro apoyo ilimitado, por ayudarme a crecer no sólo personalmente, sino también en el ámbito profesional. Y finalmente, gracias por enseñarme que de todas las experiencias se aprende, sean malas o buenas. Gracias también al resto de mi familia, a todos los que me apoyáis; en especial, a mi padrino Luis Ángel, que siempre ha creído en mí. Y gracias a mis abuelos, por lo orgullosos que estaban y estarían de mí.

Para terminar, gracias a todos mis amigos/as. Con vosotros he crecido durante muchas etapas diferentes de mi vida, y sois uno de los bloques imprescindibles de ella. En especial, gracias a Marta por ser mi compañera de batallas durante estos años de Doctorado, y a Sergio, que estás ahí día a día ayudándome a mantener la calma.

La verdadera ciencia enseña, por encima de todo, a dudar y a ser ignorante.

Miguel de Unamuno

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Econometrics could be defined as the combination of statistics, mathematics and economics in order to give empirical support to different theoretical models (Tintner (1968)). Samuelson et al. (1954) defined it as *the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.* Thus, an econometric model allows us to analyse how a dependent or explained variable ($\mathbf{Y}$) is influenced by other explanatory, independent or predictor variables ($\mathbf{X}$). Usually, the relationship between the dependent and independent variables is expressed with the following linear regression:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{1.1}$$

where $\mathbf{u}$ is the random disturbance, and $\mathbf{X}$ is a $n \times p$ matrix (n observations and p variables).

The expression above can be rewritten as follows:

$$\mathbf{Y} = \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + ... + \beta_p\mathbf{X}_p + \mathbf{u}. \tag{1.2}$$

Usually, $\mathbf{X}_1$ from model (1.2) is an all-ones matrix ($\mathbf{X}_1{}^t = \begin{pmatrix} 1 & 1 & ... & 1 \end{pmatrix}$), thus $\beta_1 \mathbf{X}_1 = \beta_1$, which represents the intercept of the model. Hence, the researcher is using $p-1$ observed explanatory variables. The key here is the definition of the matrix $\mathbf{X}$ of independent variables: is the researcher considering the intercept, the constant of the model, as another explanatory variable or not? There are authors who uphold both alternatives. Johnston and Dinardo (2001); Wooldridge (2008); Stock and Watson (2012), among others, do not contemplate the intercept as an explanatory variable, while Uriel (1997); Novales (1993); Gujarati (2003), among others, consider that the constant is another independent variable in the econometric model. With the clarification above, the reader can see that the authors take the second perspective.

The Ordinary Least Squares (OLS) estimator ($\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$) is commonly applied to estimate model (1.1). According to the Gauss-Markov Theorem, the OLS estimator is the Best Linear Unbiased Estimator (BLUE) if the random disturbance is considered spherical (homocedasticity and incorrelation). Apart from the hypothesis imposed to the random disturbance, it also needs to be verified that the range of the matrix $\mathbf{X}$ will be equal to the number of explanatory variables of the model and, consequently, the explanatory variables will not present any perfect linear relationship between them. If this requirement is not satisfied, the determinant of $\mathbf{X}^t\mathbf{X}$ will be zero and it will not be possible to obtain a unique solution for the estimates. This problem is known as perfect multicollinearity.

It will be also possible that the explanatory variables have a strong but not perfect relationship. This case is known as imperfect multicollinearity. With this type of multicollinearity, the estimation by OLS will be unique but the

determinant of matrix $\mathbf{X}^t\mathbf{X}$ will be very small. Additionally, large variances of OLS estimators may arise, as well as greater confidence intervals, insignificant t ratios and a high coefficient of determination, unstable results, wrong signs for the estimated coefficients and difficulty in determining the individual effects of the independent variables to the dependent variable and to the coefficient of determination.

It is worth noting that collinearity refers to the relationship between only two explanatory variables, while multicollinearity concerns more than two variables, so collinearity can be interpreted as a particular case of multicollinearity (Belsley (1991, 2004); Chennamaneni et al. (2011); Gujarati (2010); Holland (2014); Leighton (1985)). From now, to simplify the reading, the dissertation will henceforth use multicollinearity and collinearity as synonyms.

Some authors have stated that multicollinearity is a sample phenomenon (see for example Gujarati (2010), Johnston (1972), Stock and Watson (2012) or Wooldridge (2008)), but in many studies it is difficult or even imposible to obtain "ideal" or experimental data, and this fact (the use of real data) sometimes results in the presence of collinearity. Efforts to address multicollinearity are usually limited to deleting variables or, at best, the model is estimated with alternative traditional methods, such as ridge regression or partial least squares, which are recommended by Wei (2011) for prediction purposes but not for the analysis of causal effects. In any case, even if the goal of the study were to predict, where collinearity is not a major issue, it is highly recommended to mitigate the problem due to the continuity of the relationships between explanatory variables in the future. If this continuity is not verified, the forecast based on the initial model may be unreliable as well (Gujarati (2010); Wooldridge (2008)).

In addition, it should be highlighted that in some cases the consequences

of multicollinearity may not be troubling but its mitigation could still be recommended in order to analyse the causal effects between the variables. Baird and Bieber (2016) proposed an alternative methodology to OLS, based on ordered variable regression (Woolf (1951)), which resolves the issue of related predictors by creating and using predictors that are perfectly unrelated. Additionally, Shapley (2016) presents a different strategy for assessing the contribution of regressor variables to the dependent variable. These are the basis of regression with orthogonal variables (Novales et al. (2015); Salmerón et al. (2016)), also known as residualization methodology, which is applied in previous research articles published in major social science journals in many different fields, such as linguistics (Ambridge et al. (2012); Cohen-Goldberg (2012); Jaeger (2010); Kuperman et al. (2008, 2010); Lemhöfer et al. (2008)), environmental issues (Jorgenson (2006); Jorgenson and Burns (2007); Jorgenson and Clark (2009)) or economic development and policies (Bandelj and Mahutga (2010); Bradshaw (1987); Kentor and Kick (2008); Mahutga and Bandelj (2008); Walton and Ragin (1990)). Despite its application having been widespread, the theoretical background of the method has not been developed fully in these earlier works. The lack of specification of this methodology leads to different criticisms, such as the one in York (2012).

In order to offer a brief explanation about the application of the residualization procedure, let us consider a basic regression model with two observed explanatory variables plus the constant. Starting from model (1.1), the model will be $\mathbf{Y} = \beta_1 + \beta_2\ \mathbf{X}_2 + \beta_3\ \mathbf{X}_3 + \mathbf{u}$. Let us also suppose that variable $\mathbf{X}_2$ could be expressed as an approximate linear relationship of $\mathbf{X}_3$, which implies near collinearity. This method allows the researcher to isolate the effect of variable $\mathbf{X}_2$ from variable $\mathbf{X}_3$ by using the estimated residuals from the auxiliary regression $\mathbf{X}_2 = \alpha_1 + \alpha_2\ \mathbf{X}_3 + \mathbf{v}$ in the original model instead

of the original variable $\mathbf{X}_2$. But why is variable $\mathbf{X}_2$ isolated from variable $\mathbf{X}_3$? The answer is easy to understand: due to the OLS estimation properties, the estimated residuals of any regression by OLS are orthogonal to all the explanatory variables used in the analysis; therefore, the estimated residuals from the auxiliary regression represent the part of variable $\mathbf{X}_2$ that has no relationship with variable $\mathbf{X}_3$. It can be said that the principle *ceteris paribus* is strictly fulfilled. As the reader may note, another type of interpretation of the modified variable $\mathbf{X}_2$ is made. However, one important issue has to be taken into account: not all variables are susceptible to having their effects isolated from the others, so it is very important to choose the appropriate variable or variables from the specific model. In addition, apart from isolating the effect of variable $\mathbf{X}_2$, the method is simultaneously mitigating potential collinearity problems due to the foregoing. Furthermore, it is interesting to note that the method could be used more than once for a specific model.

With the above in mind and taking into account the residualization procedure and its properties, which are going to be analysed throughout this Thesis, it is clear that this method allows the researcher to deal with multicollinearity problems and, furthermore, it also introduces another interpretation of the modified variable(s). The main goal of this dissertation is to undertake an in-depth exploration of residualization, not only theoretically but also empirically. Regarding the empirical part of this Thesis, the focus is centred on the environment from an economic point of view. Usually, the purpose of environmental works is to study the impact of certain variables on the environment, thus, the particular research covering this field is to estimate the effects of the environmental impact factors. As has been shown earlier, in empirical research real data are generally used and this results in the presence of collinearity in many studies. That is the case with environmental studies: it

is likely that factors affecting environment will be strongly correlated. Variables like population, GDP per capita, industry, technological development, policies, etc., clearly influence environmental damage, but they influence each other as well. This fact implies multicollinearity and, as has been shown earlier, the presence of multicollinearity lead to distorted results. Despite the evidence of the presence of collinearity in environmental studies, efforts to address collinearity are usually disregarded or, at best, limited to eliminating variables, to using first differences, or to applying partial least squares or ridge regression (see Chapter 5). However, it may be considered necessary to analyse whether residualization, which is focused not only on mitigating collinearity but also on obtaining new interpretation of the variables, could be an alternative. This idea can be extended in particular to social sciences, and sciences in general.

For that purpose, the remainder of the dissertation is organised as follows:

- Chapter 2 addresses in depth the problem of multicollinearity: it reviews the concept and its causes and consequences. Furthermore, apart from the concept, the reader is introduced to the principal methodologies in the diagnosis and treatment of collinearity: the main measures for detecting the problem in a specific model and some prior methodologies for mitigating it.

- Chapter 3 introduces the reader to residualization. First, some critical views of the method are reviewed, and then the methodological preliminaries of residualization are presented. Then, the method for a linear model with two and three standardized explanatory variables is explained based on the work of Salmerón et al. (2016), developed by the author of this dissertation, her supervisors, and Dr. José García Pérez

(University of Almería, Spain), and published in *Boletín de Estadística e Investigación Operativa*.

- Chapter 4 further explains the theoretical development of the methodology for the case of p independent variables. This chapter represents the main contribution of this Thesis, and it corresponds to the work "Residualization: justification, properties and application", developed by the author of this dissertation, her supervisors, and Dr. José García Pérez (University of Almería, Spain), and published in *Journal of Applied Statistics (JAS)* (García et al. (2019c)).

- Chapter 5 introduces the empirical part. The well known STIRPAT model, which primarily studies environmental degradation, is used for three different examples. The first one is focused on the residualization procedure and its use in mitigating strong collinearity problems: it compares residualization with three other methodologies explained in Chapter 2, concluding that residualization is a good alternative for dealing with strong collinearity problems. The second one uses residualization mainly to mitigate collinearity problems, but the residualization procedure is applied in three different ways to show the applicability of the method. Finally, the third example applies residualization to show the reader the use of the method for empirical purposes. The first example is based on the work of García et al. (2020) (which presents updated data), the second uses data from the research of Apergis and García (2019), and the third is one of the examples presented in García et al. (2019c), all research developed by the doctoral candidate with her supervisors and other academics during her PhD.

- Finally, Chapter 6 offers an overall conclusion and provides some implications and future lines of research.

Chapter 2

The problem of multicollinearity

As stated by Gujarati (2010) or Novales (1988), the key question in an empirical analysis is not to discuss the existence of multicollinearity because it always exists (whatever the two economic variables, they are always correlated). According to some authors like Gujarati (2010) or Novales (1988) above or others like Johnston (1972) or Stock and Watson (2012), for empirical research real data are generally used and, as has been stated in Chapter 1, the use of real data results in the presence of collinearity in many studies. So, the debate is the choice of whether or not to ignore the problem because it is or not significant. Thus the dilemma is in fact the degree of multicollinearity that exists in an empirical study, i.e. whether the existing multicollinearity is of concern or not.

It was said in Chapter 1 that some authors have stated that multicollinearity is a sample phenomenon (Fox (1984); Gujarati (2010); Johnston (1972); Novales (1988); Schroeder (1990); Spanos and McGuirk (2002); Stock and

Watson (2012)), but let us look in-depth at the existing types of collinearity in order to clarify the concept, causes and consequences of collinearity.

As Belsley and Klema (1974) reveal, there are three principal questions about the multicollinearity problem which can be rewritten as follows:

1. What is multicollinearity, when does it appear, what are its consequences and what are the causes of the problem?
2. How can we detect the presence of multicollinearity in a specified model?
3. Is it possible to mitigate the problem? How?

These three questions will be treated throughout this Chapter. Section 2.1 answers the first question, i.e. the concept and causes and consequences of the problem. Section 2.2 takes an in-depth look at the second question and reviews the main methods for detecting the problem. Finally, Section 2.3 presents the most commonly-used methodologies that allow the researcher to deal with the problem.

2.1 Concept, causes and consequences of multicollinearity

With regard to the first question presented above, a general definition of multicollinearity is that it is a problem that consists of a lack of independence or presence of interdependence between explanatory variables, Farrar and Glauber (1967). Novales (1988), Silvey (1969) or Paul (2006) among others, distinguish between two main types of multicollinearity, paying special attention to the nature of the relationship:

- Perfect or exact multicollinearity: this occurs when one of the explanatory variables is a perfect linear combination of the rest of the variables (or only some of them), i.e. one of the explanatory variables can be expressed as an exact linear relationship of other independent variables from the initial model.

- Imperfect or near multicollinearity: this appears when one of the explanatory variables is approximately equal to a linear combination of at least one of the independent variables from the model.

Perfect multicollinearity usually indicates a logical error in the specification of the model, but imperfect multicollinearity is essentially a characteristic of the data. Therefore, if the variables included in the model are the only ones the researcher can include, then high near multicollinearity implies difficulties for obtaining accurate results (Stock and Watson (2012)), and it is desirable to mitigate it.

As Gujarati (2010) reveals, *in practice, we rarely encounter perfect multicollinearity, but cases of near or very high multicollinearity where explanatory variables are approximately linearly related frequently arise in many applications*. The following subsections will look in-depth these two types of collinearity.

2.1.1 Perfect multicollinearity

In the case of perfect multicollinearity, the $\mathbf{X}$ matrix does not have complete range and, consequently, the determinant of $\mathbf{X}^t\mathbf{X}$ will be zero, which means it cannot be inverted, so it is a singular matrix (Lazaridis (2015); Novales (1988)). With this, the Ordinary Least Squares (OLS) estimator does not have a unique solution, and the results are not decisive (Gujarati (2010);

Stewart (1987)) because the estimation of the coefficients will have an infinite number of solutions. In other words, the set of explanatory variables includes duplicated information and involves mathematical problems.

To illustrate this issue, let us assume a model with two explanatory variables and the constant[1] (see Gujarati (2010)):

$$\mathbf{Y} = \beta_1 + \beta_2\ \mathbf{X}_2 + \beta_3\ \mathbf{X}_3 + \mathbf{u}. \tag{2.1}$$

Assuming variable $\mathbf{X}_2$ is an exact linear combination of variable $\mathbf{X}_3$. If the researcher regresses $\mathbf{X}_2$ on $\mathbf{X}_3$, it will obtain that $\mathbf{X}_2 = \widehat{\alpha}_1 + \widehat{\alpha}_2\ \mathbf{X}_3$ due to $\mathbf{e} = \mathbf{0}$ (the estimated residuals from this last regression are zeros), where $\widehat{\alpha}_1$ and $\widehat{\alpha}_2 \in \mathbb{R}$ (these are the estimated values from the regression of $\mathbf{X}_2$ on $\mathbf{X}_3$, so they are known values). For this particular case, it is assumed that $\mathbf{X}_2$ is an exact linear combination of variable $\mathbf{X}_3$, thus it is logical that R^2 from model $\mathbf{X}_2 = \alpha_1 + \alpha_2\ \mathbf{X}_3 + \mathbf{v}$ has a value equal to one.

By substituting the previous results into the main equation (2.1), the following is obtained:

$$\begin{aligned}\mathbf{Y} &= \beta_1 + \beta_2\ (\widehat{\alpha}_1 + \widehat{\alpha}_2\ \mathbf{X}_3) + \beta_3\ \mathbf{X}_3 + \mathbf{u},\\ &= \beta_1 + \widehat{\alpha}_1\ \beta_2 + \widehat{\alpha}_2\ \beta_2\ \mathbf{X}_3 + \beta_3\ \mathbf{X}_3 + \mathbf{u},\\ &= \delta_1 + \delta_2\ \mathbf{X}_3 + \mathbf{u}.\end{aligned} \tag{2.2}$$

where:

$$\delta_1 = \beta_1 + \widehat{\alpha}_1\ \beta_2.$$

$$\delta_2 = \widehat{\alpha}_2\ \beta_2 + \beta_3.$$

If the researcher estimates model (2.2), values $\widehat{\delta}_1$ and $\widehat{\delta}_2$ will be obtained. With these, and bearing in mind the previous changes in variable, it is clear

[1] Based on the idea expressed in model (1.2) regarding the intercept.

that there are three unknown parameters and only two equations, so there is no way to calculate the unknowns:

$$\begin{cases} \widehat{\delta}_1 = \beta_1 + \widehat{\alpha}_1 \ \beta_2 \\ \widehat{\delta}_2 = \widehat{\alpha}_2 \ \beta_2 + \beta_3 \end{cases}$$

In conclusion, as has been shown earlier, the principal consequence of perfect collinearity is that the OLS estimator does not have a unique solution, thus estimation and hypothesis testing of individual regression coefficients in a multiple regression is not possible, Gujarati (2010). What the researcher could obtain are the estimates of the linear combination ($\widehat{\delta}_1$ and $\widehat{\delta}_2$), but not all the "unknowns" individually (β_1, β_2 and β_3).

This problem is not very usual because the first step to follow in every work or study is to make decisions about the variables used and, hence, about the model, with the aim of choosing the best option. It means making a further diagnosis of a set of variables and to check and select the correct ones, in order to avoid including either redundant or unnecessary information (duplicated data) in the model. In any case, the appearance of perfect multicollinearity is easy to solve by deleting the redundant variables from the model (Alauddin and Nghiem (2010); Grewal et al. (2004); Leamer (1973)). Also, software can be used to automatically detect the existence of this type of collinearity by noticing any errors in the calculation.

2.1.2 Imperfect multicollinearity

Imperfect multicollinearity does not imply an exact linear relationship between variables, but an approximate linear relationship between them. This type is more difficult to manage because it usually persists due to the own characteristics of the variables. In other words, approximate collinearity

is difficult to delete because the researcher is modeling a reality in which, generally, there is always a type of relationship between the empirical variables (Gujarati (2010); Novales (1988)).

Starting from model (2.1), with two explanatory variables and the constant, and assuming variable $\mathbf{X}_2$ is an approximate linear combination of variable $\mathbf{X}_3$, now if the researcher regresses $\mathbf{X}_2$ on $\mathbf{X}_3$, a level of R^2 near to 1 will be obtained, $\mathbf{X}_3$ explains close to 100% of the variation of $\mathbf{X}_2$, so it could be concluded that the explained and the explanatory variable ($\mathbf{X}_2$ and $\mathbf{X}_3$, respectively) are closely related but not exactly related, Wooldridge (2008). When variables are highly (not perfectly) correlated, *OLS estimators still remain BLUE [Best Linear Unbiased Estimator] even though one or more of the partial regression coefficients in a multiple regression can be individually statistically insignificant*, Gujarati (2010). In this case, matrix $\mathbf{X}$ has a full rank and is not singular, so the OLS estimator does have a unique solution but the estimation of the coefficients will be unstable. Therefore, even though the researcher may be able to estimate the model, high multicollinearity can lead to the following practical consequences (Alin (2010); Farrar and Glauber (1967); Gujarati (2010); Holland (2014); Leamer (1973); Leighton (1985); Meloun et al. (2002); Murray (2005); Rockwell (1975); Stewart (1987); Wooldridge (2008)):

- Inflated variances of the estimators.
- Greater confidence intervals.
- Tendency to consider the estimated parameters as non-significant. For the tests of individual significance, the null hypothesis is likely not to be rejected.

- High R^2, which means there is a tendency to consider the model globally significant and well determined.

- Difficulty in fixing the individual effects of the independent variables to the explained variable, and hence, to the explained sum of squares and to the coefficient of determination. It is not possible to separate the individual effects of the explanatory variables: the related variables *are so highly collinear that when one moves the other moves with it almost automatically*, Gujarati (2010).

- Non-robust results. The estimates have a high sensitivity to small changes in the initial data.

- A considerable possibility of the appearance of incorrect signs for the estimated coefficients. The estimated parameters and their importance in the model are distorted, so the outcome of the specific study will show unrealistic results. In other words, it is likely that results will be inconsistent with theory.

Bearing the above in mind, Paul (2006) affirms that *if the goal is simply to predict* $\mathbf{Y}$ *from a set of variables* $\mathbf{X}$*, then [high] multicollinearity is not a problem [because] the predictions will still be accurate, and the overall* R^2 *(or adjusted* R^2*) quantifies how well the model predicts the* $\mathbf{Y}$ *values.* But, this work also says *if the goal is to understand how the various* $\mathbf{X}$ *variables impact* $\mathbf{Y}$*, then multicollinearity is a big problem*, so the researcher has to take it into account the following implications: *one problem is that the individual p values can be misleading [...]. The second problem is that the confidence intervals on the regression coefficients will be very wide [...], [and] excluding a subject (or adding a new one) can change the coefficients dramatically and*

may even change their signs. The multicollinearity problem indeed affects least squares estimations, but not residuals or predictions, e.g. Belsley et al. (1980); Chatterjee and Hadi (1988); Giacalone et al. (2018); Gujarati (2010); Lauridsen and Mur (2006); Wooldridge (2008). Multicollinearity is significant when the aim of the study is to obtain reliable estimations but not to make predictions. In any case, even if the goal of the study were to make predictions, it is highly that the problem be mitigated since the researcher needs to be very sure of the continuity of the relationships between explanatory variables, because if the relationship changes in the future, the forecast based on the initial model may be unreliable as well (Gujarati (2010); Wooldridge (2008)).

On the other hand, Spanos and McGuirk (2002) *revisit the traditional account regarding near-multicollinearity in an attempt to reconsider its nature and consequences*. This work stated that the problem with near multicollinearity could be summarise into two different issues that are usually mixed:

- The structural problem, which increases systematic volatility. Systematic volatility is a parameter problem and it could be said to be predictable. It is concerned with changes of the coefficient estimates associated with high correlation among the regressors, therefore it is related to the presence of high correlation among regressors. This is known as systematic multicollinearity (Salmerón and Rodríguez (2017)).

- The numerical problem, which increases erratic volatility. This type of volatility is due to the characteristics of the data and is unpredictable. It is concerned with the sensitivity of the coefficient estimates to proportional changes in $\mathbf{X}^t\mathbf{X}$ and $\mathbf{X}^t\mathbf{Y}$, so it is related to the presence

of ill-conditioning in the regressor data matrix $\mathbf{X}^t\mathbf{X}$. This is known as erratic multicollinearity (Salmerón and Rodríguez (2017)).

Hence, in light of the above works, the reader could think that the multicollinearity problem is not always a sample phenomenon, although many authors (see, for example, Fox (1984) or Schroeder (1990) among others) state that collinearity is commonly interpreted as a data problem rather than a model-specification problem.

In closing, a very important classification of near collinearity is given by the works of Marquardt (1980); Marquardt and Snee (1975); Snee and Marquardt (1984). In light of these, the analyst can also distinguish between essential collinearity and non-essential collinearity. The first one concerns the relationship between explanatory variables, excluding the intercept, while the second one regards the specific relationship between the intercept and at least one of the observed independent variables of the model. So it could be interpreted that both, taken individually, are measuring the relationships among the real variables used in a specific model (the numerical problem) and both of them together are measuring the structural problem of the model.

To sum up, near or imperfect multicollinearity can be split into two groups:

- Those regarding structure: data structure (erratic collinearity) or model structure (systematic collinearity).

- Those regarding relationships: taking into account the relationship between the constant of the model and the rest of independent variables (non-essential collinearity) or considering only the relationship between explanatory variables without considering the constant (essential collinearity).

Throughout this Thesis, the group that concerns us most is the one that takes into account the relationship among explanatory variables.

Before reviewing the treatment of collinearity, it is important to further illustrate how multicollinearity can be detected (Section 2.2). Once it is discussed, in Section 2.3 the principal techniques and methodologies used in earlier literature to deal with multicollinearity problems will be individually explained.

2.2 Detection of multicollinearity in a model

Considering the general linear regression model for p explanatory variables and n observations, model (1.1), the objective is to estimate $\boldsymbol{\beta}$.

In the presence of multicollinearity there will be high instability in the estimation of $\boldsymbol{\beta}$: with small changes in the $\mathbf{X}$ matrix, there will be big changes in the estimation of $\boldsymbol{\beta}$ and the regressors may have a high sampling variance, as has been mentioned in Section 2.1. But, how the problem could be detected?

The distinction was made above between perfect and near multicollinearity. Perfect multicollinearity is directly detected by observing the matrix $\mathbf{X}$: if it is singular, there is perfect multicollinearity. This section therefore deals with near or imperfect multicollinearity. It has also been noticed that there is always collinearity in empirical modelling, hence the issue that concerns us here in that of detecting whether strong collinearity problems exist.

Imperfect collinearity can be detected in several ways. The researcher may perform some informal checks to have an initial approach of the problem. If the model has a high R^2 and is globally significant, while the estimated parameters are individually insignificant, there may be strong collinearity in

the model. If the Pearson coefficient of correlation is higher than 0.8 (Farrar and Glauber (1967); Grewal et al. (2004); Kumar (1975)), the problem is likely to exist. Note that this threshold only measures the relationship between pairs of variables and, in addition, this value does not ensure the existence of strong collinearity in the model: it would better to use a value of 0.9 (see García et al. (2017b) for more information). In any case, although these informal checks may help the researcher to form an initial idea about the existence of strong near multicollinearity in the model, there are formal checks. The following sections take an in-depth look at this fact.

2.2.1 Farrar and Glauber tests

As stated by Neeleman (1973), Farrar and Glauber (1967) designed a test on multicollinearity for detecting the problem, localising it, and finding the multicollinearity pattern. In line with these authors, let us assume that $\mathbf{x}$ is a standardized matrix of n observations and $p-1$ explanatory variables. From now, bold lowercase letters will represent standardized variables and bold capital letters will represent non-standardized variables.

It is known that the determinant of $\mathbf{x}^t\mathbf{x}$ takes the value 0 when there is complete dependency between variables and 1 when the variables are orthogonal, so $0 \leq \left|\mathbf{x}^t\mathbf{x}\right| \leq 1$. Based on this idea, the following test for checking if variables are mutually independent (if multicollinearity does not exist) has been developed:

$$\chi^2_{|\mathbf{x}^t\mathbf{x}|} = -[n-1-\frac{1}{6}\left(2(p-1)+5\right)]\ln\left|\mathbf{x}^t\mathbf{x}\right|,$$

which is distributed approximately as a χ^2 with $\left(\frac{1}{2}\ (p-1)\ (p-2)\right)$ degrees of freedom.

The second test developed by Farrar and Glauber (1967) is the one that localises the problem:

$$F_i = (r_{ii} - 1)\frac{n-p-1}{p-2}, \tag{2.3}$$

which has a Snedecor's F-distribution with $(n-p-1)$ and $(p-2)$ degrees of freedom. In this expression, $r_{ii} = 1 - (1/R_i^2)$, where R_i^2 represents the coefficient of determination of following regression (2.4):

$$\mathbf{x}_i = \mathbf{x}_{-i}\boldsymbol{\alpha} + \mathbf{v}, \tag{2.4}$$

where $\mathbf{v}$ is spherical, $\mathbf{x}_{-i}$ is the result obtained after eliminating column (variable) i from matrix $\mathbf{x}$ and $\mathbf{x}_i$ represent the variable i. That is, $\mathbf{x} = (\mathbf{x}_{-i}\ \mathbf{x}_i)$.

This previous test could be applied for each variable in order to detect the collinear variables.

Finally, to identify the multicollinearity pattern, the last test developed by Farrar and Glauber (1967) is the following:

$$t_{ij} = \frac{\rho_{ij}\sqrt{n-p-1}}{\sqrt{1-\rho_{ij}^2}},$$

which has a Student's t-distribution with $(n-p-1)$ degrees of freedom, and where ρ_{ij} is the coefficient of correlation between variables $\mathbf{x}_i$ and $\mathbf{x}_j$. This test can be used to study the pattern of the mutual relationships in the collinear subset detected with the F test above.

Although the set of these measures could be seen as a good tool for detecting and identifying the multicollinearity pattern, the procedure has received a lot of criticism. The principal disadvantage of the Farrar and Glauber technique is that, in words of O'Hagan and McCabe (1975), they made *a fundamental mistake in interpreting their diagnostics: a very simple conceptual error*. This work refers to the missinterpretation of the use of a t statistic, by providing a

fundamental measure of the severity of collinearity. Their statistic does not provide any more information than $|\mathbf{x}^t\mathbf{x}|$ because it *is simply some scalar times* $\log|\mathbf{x}^t\mathbf{x}|$ *and only has the relatively small advantage of being corrected from sample to sample for degrees of freedom.* Similarly, Haitovsky (1969) argues that *the only relevant requirement in the context of multicollinearity is the so-called full rank requirement*, and strong collinearity problems should be shown by the singularity of the matrix $\mathbf{x}^t\mathbf{x}$. Silvey (1969) claims that the most important issue arising from multicollinearity is the imprecise estimations, rather than seek to define the degree of the problem. Huang (1970) comments that there is no analytical measure of the severity of multicollinearity on $\mathbf{x}$ except for $|\mathbf{x}^t\mathbf{x}|$. Wichers (1975) demonstrates that the third test developed by Farrar and Glauber does not work. Finally, Smith and Campbell (1980), in their criticism to ridge regression, state that the Farrar and Glauber criterion is inadequate.

2.2.2 Variance Inflation Factor (VIF)

The VIF can be interpreted as a tool based on the correlation between explanatory variables (Novales (1988)).

Starting from model (1.1) and supposing $\mathbf{X}$ can be decomposed as $\mathbf{X} = \begin{pmatrix}\mathbf{X}_i & \mathbf{X}_{-i}\end{pmatrix}$, where $\mathbf{X}_i$ represents variable i and $\mathbf{X}_{-i}$ is the matrix that includes the rest of the explanatory variables. Then,

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} \mathbf{X}_i^t\mathbf{X}_i & \mathbf{X}_i^t\mathbf{X}_{-i} \\ \\ \mathbf{X}_{-i}^t\mathbf{X}_i & \mathbf{X}_{-i}^t\mathbf{X}_{-i} \end{pmatrix}.$$

By using the inverse of a partitioned matrix, the important element to obtain $(\mathbf{X}^t\mathbf{X})^{-1}$ is the element $(1,1)$ of the same, that is:

$$\left((\mathbf{X}_i^t\mathbf{X}_i) - \mathbf{X}_i^t\mathbf{X}_{-i}(\mathbf{X}_{-i}^t\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^t\mathbf{X}_i\right)^{-1} = \left(\mathbf{X}_i^t\ \mathbf{M}_i\ \mathbf{X}_i\right)^{-1},$$

where $\mathbf{M}_i = \mathbf{I} - \mathbf{X}_{-i}(\mathbf{X}_{-i}^t\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^t$, which is symmetric and idempotent, and $\mathbf{I}$ is the identity matrix. So,

$$\widehat{Var}(\widehat{\beta}_i) = \frac{\widehat{\sigma}^2}{\mathbf{X}_i^t\ \mathbf{M}_i\ \mathbf{X}_i}.$$

As Giacalone et al. (2018) say, the VIF studies the linear dependence between variable $\mathbf{X}_i$ and the rest of explanatory variables of the original model (1.1), hence the regression which concerns us at this point is the following:

$$\mathbf{X}_i = \mathbf{X}_{-i}\boldsymbol{\alpha} + \mathbf{v}, \tag{2.5}$$

where $\mathbf{v}$ is spherical.

The sum of square residuals from model (2.5), SSR_i, is equal to $\mathbf{X}_i^t\ \mathbf{M}_i\ \mathbf{X}_i$. Thus,

$$\widehat{Var}(\widehat{\beta}_i) = \frac{\widehat{\sigma}^2}{\text{SSR}_i} = \frac{\widehat{\sigma}^2}{\text{SST}_i(1-R_i^2)},$$

where SST_i and R_i^2 are the total sum of squares and the coefficient of determination from model (2.5), respectively. It is important to note that the above expression is verified if the intercept is in the model (when it is verified that SST = SSE + SSR).

In the above expression, it is known that $\widehat{\sigma}^2$ is totally independent from the correlation between explanatory variables, SST depends only on $\mathbf{X}_i$, and R_i^2 is influenced not only by $\mathbf{X}_i$ but also by the rest of explanatory variables $\mathbf{X}_{-i}$. Hence, regarding multicollinearity, the only factor that affects $\widehat{Var}(\widehat{\beta}_i)$ in these terms is R_i^2. Furthermore, the lower value of this variance appears when $R_i^2 = 0$, and this fact occurs when the explanatory variables are linearly

independent each other, so $\widehat{Var}(\widehat{\beta}_i)$ when there is any relationship in $\mathbf{X}_{-i}$, $\widehat{Var}(\widehat{\beta}_i^0)$, is equal to σ^2/SST_i.

Regarding the threshold that the researcher has to take into account when testing the existence of collinearity, Marquardt (1970) states that a value lower than 10 indicates no problematic collinearity, Kennedy (1992). For tighter results, some authors set the VIF threshold to 4 (O'Brien (2007)).

If the VIF is defined as the percentage of the variance that is inflated for each coefficient, so the VIF value shows the ratio between the actual situation and the situation where there is no collinearity, then:

$$\text{VIF}_i = \frac{\widehat{Var}(\widehat{\beta}_i)}{\widehat{Var}(\widehat{\beta}_i^0)} = \frac{1}{1 - R_i^2}. \tag{2.6}$$

Note that expression (2.6) corresponds to r_{ii} from equation (2.3) of Subsection 2.2.1.

Thus, the VIF is related to the R_i^2 (the coefficient of determination of model (2.5)), and if the threshold of the VIF is situated at 10, then the R_i^2 will be equal or higher than 0.9 if worrying near collinearity exists in the model, and it will be equal or higher than 0.75 if the researcher takes the value of 4 as the VIF threshold.

On the other hand, the works by Curto and Pinto (2011) and Salmerón et al. (2017b) developed a corrected version of the traditional VIF. Curto and Pinto (2011) explain the corrected VIF (CVIF) is useful when explanatory variables are not redundant and their importance increases due to the inclusion of another explanatory variable in the regression equation. However, according to Salmerón et al. (2017b), R_0^2 is non-negative and could even be higher than one. This implies that the CVIF will be negative since $1 - R_0^2 < 0$ and $0 < R^2 < 1$. Taking into account the rules of thumb proposed by Curto and Pinto (2011), the final consequence of this fact is that the CVIF can take non-interpretable

negative values. Salmerón et al. (2017b) proposed a modified CVIF (MCVIF) that corrects the above one.

Although the VIF is one of the most commonly-used methods to detect collinearity in a specified model, some inconveniences appear. First, the VIF does not detect the non-essential collinearity because it ignores the role of the intercept of the model (see Salmerón et al. (2018, 2019)). Second, as the VIF is based on the calculation of R_i^2, it is not appropriate when qualitative variables are used in the model, specifically, it is not recommended when variable $\mathbf{X}_i$ is binary.

2.2.3 Tolerance

It is known that the tolerance can be defined as $\text{TOL}_i = 1 - R_i^2$. So from expression (2.6), the tolerance of variable $\mathbf{X}_i$ can be interpreted as the inverse of the VIF_i. Thus, the tolerance of a variable is related to the value of the VIF.

If value 10 is taken as the threshold for the VIFs (see Hair et al. (1995); Kennedy (1992); Neter et al. (1989)), then $\text{TOL}_i < 0.1$. That is, if the tolerance of variable $\mathbf{X}_i$ is lesser than 0.1, this is a problematic variable regarding worrying multicollinearity of the model. In parallel, if a value of 4, following O'Brien (2007), is taken as the threshold for the VIFs (see, for example, Pan and Jackson (2008)), then $\text{TOL}_i < 0.25$, so the researcher can have tougher requirements regarding collinearity problems.

As the tolerance is the inverse of the VIF, it presents the same disadvantages as the original expression: it does not detect the non-essential collinearity and it is not appropriate when qualitative variables are used in the model.

2.2.4 Condition Number (CN)

The CN can be interpreted as a method based on the size of $\mathbf{X}^t\mathbf{X}$ (Novales (1988)). It is known that the CN measures the sensitivity of the solution of a linear equation model to changes in the original data, i.e. it measures the sensitivity of an inverse matrix to changes in it.

As has been shown in the above section, the problem of multicollinearity is caused by the matrix $\mathbf{X}$. With near multicollinearity, this matrix is not singular, but it could be said that it is approximately singular. Hence, one possibility to detect the problem could be based on the size of this matrix. The first idea that emerges is to use the determinant of this matrix. As in Novales (1988), the determinant of a symmetric matrix is equal to the product of the eigenvalues of it. Thus, analysing the eigenvalues of $\mathbf{X}^t\mathbf{X}$, the problem of multicollinearity could be examined by taking the size of this matrix as the starting point. Small eigenvalues will produce a small determinant, and a small determinant will mean that the variables are highly correlated, so the model presents problematic collinearity. It is important to note that the value of the determinant of $\mathbf{X}^t\mathbf{X}$ is sensitive to the units of measure employed for the variables, and it is a considerable disadvantage. Because of this, it could be interesting to examine the individual eigenvalues not the whole determinant and, in particular, their relative values. For example, if the ratio between the higher eigenvalue ($\mu_{\max}$) and the lower one ($\mu_{\min}$) is studied, and if the value of the ratio is small, then it could be concluded that the minimum value is relatively high compared to the maximum eigenvalue, indicating that the collinearity problem will not be problematic. By contrast, if the ratio is high, then the minimum eigenvalue will be relatively small regarding to the maximum one, and multicollinearity will be a major problem.

Bearing that in mind, the well-known CN will be calculated as:

$$\text{CN} = \sqrt{\frac{\mu_{\max}}{\mu_{\min}}}. \tag{2.7}$$

So, with a low condition number, $\mathbf{X}^t\mathbf{X}$ is considered well-conditioned, while with a high condition number, $\mathbf{X}^t\mathbf{X}$ is considered ill-conditioned.

The minimum value of CN will be one, and it appears when all of the explanatory variables are orthogonal to each other, which is the situation where there is no relationship between independent variables. Following Belsley et al. (1980), values lower than 20 imply light collinearity, between 20 and 30, moderate collinearity, and values higher than 30 imply strong collinearity. Thus, a value of 30 is usually taken as the threshold for detecting strong collinearity (see, for example, Midi et al. (2010); Myers (1990); Pesaran (2015)).

In contrast to VIF, since the CN takes into account the constant of the model, it can be used to detect not only essential collinearity, but also non-essential collinearity. In any case, it presents a significant disadvantage, explained in depth in Lazaridis (2007): sometimes the value of the CN delivers inflated results even if the model does not present strong collinearity problems. As this author reveals, this problem is usually, although not necessarily, created by the intercept (non-essential collinearity), thus it is recommended to use centred data where possible and some other tools together with CN to check the existence of strong collinearity problems.

2.2.5 Stewart Index (SI)

Stewart (1987) defined the *collinearity indices*, whose purpose was to detect the existing near collinearity in the econometric model. The Stewart index, usually named as k_i^2 for variable $\mathbf{X}_i$, is able to identify essential and non-essential

collinearity in an econometric model, in contrast to VIF, for example. In this dissertation, the index is going to be referred to as SI_i to avoid any confusion with the ridge factor, usually referred to as k (see Section 2.3.1).

The expression for the SI_i is the following:

$$\mathrm{SI}_i = \frac{\mathbf{X}_i^t \mathbf{X}_i}{\left| \mathbf{X}_i^t \mathbf{X}_i - \mathbf{X}_i^t \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i} \right)^{-1} \mathbf{X}_{-i}^t \mathbf{X}_i \right|},$$

where $\mathbf{X}_{-i}$ represent the matrix $\mathbf{X}$ by deleting variable $\mathbf{X}_i$ and $\mathbf{X}_i$ represents variable i.

It is verified that if $\mathbf{X}_i \mathbf{X}_{-i} = \mathbf{0}$, then SI_i is equal to 1. Furthermore, as $\mathbf{X}_{-i} \mathbf{X}_{-i}$ is a positive-definite matrix, then $\mathrm{SI}_i \geq 1$, so this index is able to capture the existing orthogonality between variable $\mathbf{X}_i$ and the rest of explanatory variables.

For the intercept ($i = 1$), the index is:

$$\mathrm{SI}_1 = \frac{1}{1 - \frac{1}{n} \cdot \overline{\mathbf{X}}_{-1} \cdot \left(\mathbf{X}_{-1}^t \mathbf{X}_{-1} \right)^{-1} \cdot \overline{\mathbf{X}}_{-1}^t},$$

where $\overline{\mathbf{X}}_{-1} = \mathbf{X}_1^t \mathbf{X}_{-1} = \begin{pmatrix} \sum_n X_{n2} & \sum_n X_{n3} & \ldots & \sum_n X_{np} \end{pmatrix}$.

Here, the orthogonality of the constant with the rest of explanatory variables is measured (the non-essential collinearity of the model). If the value of the index is exactly 1, then, $\overline{\mathbf{X}}_{-1} = \mathbf{0}$, i.e. all the explanatory variables are centred, there is no non-essential collinearity, and SI_1 has its minimum value.

For the rest of explanatory variables (in this case, $i = 2, \ldots, p$), the index captures the orthogonality of the analysed variable with the rest of independent variables:

$$\mathrm{SI}_i = \frac{\mathbf{X}_i^t \mathbf{X}_i}{\mathrm{SSR}_i},$$

where SSR_i is the residual sum of squares of the regression (2.5).

As $\mathbf{X}_i^t\mathbf{X}_i = n \cdot (\text{var}(\mathbf{X}_i) + \overline{\mathbf{X}}_i^2) = \text{SST}_i + n \cdot \overline{\mathbf{X}}_i^2$, and $\text{VIF}_i = \text{SST}_i/\text{SSR}_i$, then the index can be expressed as a function of the VIF:

$$\text{SI}_i = \text{VIF}_i + n \cdot \frac{\overline{\mathbf{X}}_i^2}{\text{SSR}_i}.$$

If the explanatory variables are centred, $\overline{\mathbf{X}}_i = 0$, then SI_i has exactly the same value as VIF_i. Thus, the values are related, but they give the same value only when the researcher is using centred variables.

2.2.6 Coefficient of Variation (CV)

For testing the existence of non-essential collinearity problems, Salmerón et al. (2019) demonstrate that the CV is a good measure of the problem. The authors consider that strong non-essential collinearity appears due to a small variance of the explanatory variable which is under analysis. With this purpose, they *determine how small this variance needs to be for collinearity becomes a serious problem*. Starting from the traditional CN, they obtained an expression that links the variance of the variable under analysis with its mean. Following this expression, they give a threshold that indicates when non-essential collinearity becomes problematic, which is a value of CV less than approximately 0.07.

2.2.7 Red indicator

The Red indicator was presented by Kovács et al. (2005) as an alternative to VIF or CN for measuring multicollinearity in a specified model. Basically, this indicator quantifies the average correlation of the dataset. Following this author, values close to one imply near collinearity problems, and values close to zero indicate no strong collinearity.

As García et al. (2015) write, the Red indicator *is related to the redundancy between the variables and, in its simplest expression, the Red indicator is the*

quadratic mean of the elements outside the main diagonal of the correlation matrix of exogenous variables, $\mathbf{R}$*. According to Kovács et al. (2005), the Red indicator is though to measure the redundancy related to the proportion of useful data in the database. The less useful data in the database the greater redundancy and viceversa.* With this, García et al. (2015) recommend the Red indicator to measure the systematic volatility (high correlation between explanatory variables).

For a standardized model, the expression of the Red indicator is the following:

$$\text{Red} = \sqrt{\frac{\sum_{i=2}^{p}\sum_{j=2}^{p}\rho_{ij}^2}{(p-1)\ (p-2)}}, \quad i \neq j,$$

where ρ_{ij} is the correlation between variables i and j, and there are $p-1$ explanatory variables.

The value of the indicator will quantify the existing collinearity and the useful data there are, compared to a database of the same size and with the minimum redundancy, García et al. (2015). If there is no redundancy, the value of the indicator will be zero, and if the maximum redundancy appears, the value will be one[2].

Finally, the Red indicator is related to the VIF when $p = 3$ (two observed explanatory variables plus the intercept): $\text{VIF} = 1/(1 - \text{Red})$, as García et al. (2015) demonstrate. Thus, if the value 10 is taken as the threshold for the VIF, then collinearity problems will appear if $\text{Red} \geq 0.9$. Additionally, since the Red indicator is based on correlations, it presents the same disadvantages as the use of the coefficient of correlation for detecting collinearity problems.

[2]For example, if the value is Red= 0.3, then the proportion of useful data will be 70%, García et al. (2015).

2.2.8 Curto and Pinto indicators

Curto and Pinto (2007) develop two principal indicators to detect collinearity problems. In this section, they are briefly explained.

The first indicator, the Direct Effects Factor (DEF) compares the direct effects of explanatory variables on the dependent variable with the indirect effects resulting from the intercorrelations between explanatory variables. This indicator takes values from zero to one, and values close to one imply the existence of collinearity problems.

In turn, the Inter-Correlation Effect (ICE) is a relative collinearity measure for testing how the estimated parameters are affected by the existing correlation between explanatory variables. We will have as many ICE statistics as explanatory variables in the model, so it is not an overall measure. It also takes values from zero to one, and, as the authors said, *if the value of a reduced number of statistics is very small when compared to the others, the estimated coefficients associated with the corresponding explanatory variables can be strongly affected by the correlation among regressors.*

2.3 How to mitigate collinearity: traditional methodologies

Once the existence of high near multicollinearity is detected, it is time for the researcher to apply a methodology that can mitigate the problem. This section looks in depth at the most commonly-used ones in previous research.

First of all, it is important to note that no matter the origin of the problem, it is an element with negative consequences in the model, as has been shown in

Section 2.1, so it has to be dealt with in a way that ensures accurate conclusions and robust results. This leads to the third question raised at the beginning: Is it possible to mitigate the problem? How? In the literature, there have been emerged some types of "solutions" to the problem[3] (Alin (2010); Farrar and Glauber (1967); Grewal et al. (2004); Wurm and Fisicaro (2014); York (2012)):

- Deleting one or more of the explanatory variables. Here, *multicollinearity constitutes a problem only if it undermines that portion of the independent variable set that is crucial to the analysis in question*, Farrar and Glauber (1967). Otherwise, the problematic and non-crucial variables from the study could be deleted and the results will improve. Pasha and Shah (2004) comment that the procedure of selecting variables for a particular model could not be performed well because of the high correlation between predictor variables. In this line of argumentation, Hoerl and Kennard (1970b) propose the following technique for deleting variables from a specified model with collinearity:

  1. The first variables to be deleted from the model will be those which are stable but have a weak predicting power: the less significant variables are deleted.
  2. If the problem persists, from the remaining variables, those with small coefficients will be eliminated.

- Introducing a priori information. Farrar and Glauber (1967) consider that this is the first step in the treatment of collinearity once the problem has been detected: *[correction of multicollinearity] requires the generation*

[3]The reader has to take into account that there are always some relationships among variables in any empirical modelling, as has been said, and actually that is why there are no solutions but rather methodologies that allow the researcher to deal with strong collinearity problems.

of additional information [...]. It may involve additional primary data collection, the use of extraneous parameter estimates from secondary data sources, or the application of subjective information through constrained regression, or through Bayesian estimation procedures. Fabrycy (1973) suggests that re-specifying the structure of the explanatory variables could "solve" the problem: *Frequently all variables are entered into economic behavioral functions in the same form (e.g. linear in parameters) even though, for some variables, this is in conflict with economic realism*, and the author states that *multicollinearity can in some cases be overcome by adopting nonlinear mathematical functions which comply more closely with economic relationships.*

- Adding new data or new variables to the model which could introduce some degree of independency into it.

In practice, the previous solutions create new circumstances or new models by deleting variables, introducing more information or increasing the sample or the variables used. These procedures may be very difficult to build or even very expensive to implement. Furthermore, these new circumstances could produce other problems in the specified model like heteroscedasticity or endogeneity. Additionally, these solutions may mitigate collinearity under certain conditions. For example, in the case of erratic collinearity, to increase the sample or to include a priori information could be an interesting choice.

Using alternative methodologies to OLS is another procedure considered in earlier research to deal with multicollinearity problems. As distinguished by García et al. (2011), there are two types of techniques that allow the researcher to estimate a model with collinearity problems (Alin (2010); Farrar

and Glauber (1967); Grewal et al. (2004); Wurm and Fisicaro (2014); York (2012)): 1) methods that directly solve the algebraic problem: ridge regression, or LASSO regression, and 2) methods that act on the sample by modifying (or deleting) part of it: principal components analysis, raise regression or the one that concerns this dissertation, which is residualization.

As Schroeder (1990) says, the advantage of these type of techniques, which are biased methods, is that *the theoretical model is not compromised. The disadvantage is that the estimators are no longer unbiased as they are in the commonly used OLS regression procedure. However, if the reduction in the mean square error variance is greater than the magnitude of the bias induced in the estimators, the trade-off seems warranted. [...] The tradeoff is between using an unbiased model, such as OLS, with unstable regression coefficients or using a biased model in an attempt to stabilise the regression coefficients, reduce the error, and render the model more generalisable.*

The following sections will review the main alternative techniques and methodologies to OLS used in earlier literature to deal with strong essential multicollinearity problems. Table 2.1 contains a summary of the characteristics of the results obtained by each method, which are explained in following subsections.

2.3.1 Ridge regression

Ridge regression was introduced by Hoerl and Kennard (1970a,b) and is a common methodology used in the treatment of strong essential collinearity (Alauddin and Nghiem (2010); Alin (2010); Grewal et al. (2004); Kiers and Smilde (2007); Meloun et al. (2002)). It is one of the solutions to collinearity known as shrinkage or regularisation, which basically consists in minimising

Table 2.1: Comparison of methods.

Based on the statistical analysis of the model	**Ridge regression**	**LASSO regression**
Estimated parameters	Original variables modified in some way, depending on the multicollinearity problem. The inference may not be interpreted.	Original variables modified in some way, depending on the multicollinearity problem. The inference may not be interpreted.
Global robustness	The global significance may not be interpreted.	The global significance may not be interpreted.
Based on the numerical analysis of the model	**PCR**	**Raise regression**
Estimated parameters	Different data: components obtained from PCA. Importance of the variables in the model observed by using the VIP values.	Original variables modified in some way, depending on the multicollinearity problem. Individual significance of the estimated parameters by using t statistic.
Global robustness	% cumulative variance explained.	R^2, F statistic and sum of squares with the same values as OLS estimation.

the influence of the less important predictors. Tibshirani (1996) affirms that one of the reasons why the data analyst is often not satisfied with the OLS estimates is prediction accuracy. According to this author, *the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients. By doing so a little bias is scarified to reduce the variance of the predicted values and hence may improve the overall prediction accuracy.* By introducing small changes in the data, models will be unstable, as has been said throughout this chapter, so the prediction accuracy is reduced. Ridge regression can be defined as *a continuous process that shrinks coefficients and hence is more stable: however, it does not set any coefficients to 0 and hence does not give an easily interpretable model*,

Tibshirani (1996).

As is well-known, starting from model (1.1), the way to estimate $\boldsymbol{\beta}$ using OLS leads to the formula $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$.

Basically, this ridge method consists in adding a reasonable amount of bias into the model, which means that the mean square error of prediction decreases. So, the ridge estimator will be:

$$\widehat{\boldsymbol{\beta}}(k) = (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{Y}, \tag{2.8}$$

where $k > 0$ and $\mathbf{I}$ is the identity matrix.

As Holland (2014) says, this addition of the value k, named as ridge factor, allows ridge regression to have enough flexibility to reduce the inflated variances of OLS coefficients that arise from multicollinearity (Li et al. (2010)), and thus increases the reliability of point estimates (Butler and McNertney (1991)). By increasing the diagonal elements of $\mathbf{X}^t\mathbf{X}$, the size of this matrix changes, and the problem of approximate singularity is avoided, García et al. (2017b); Novales (1988). The reliability and stability of the ridge regression estimation are based on the selection of the k parameter, and numerous methodologies have been proposed for this purpose. Hoerl and Kennard (1970a,b) found that k must lie between zero and one, and it should be as small as possible to retain the maximum amount of information. The traditional criterion to select k proposed by Hoerl et al. (1975) is given as:

$$k = \frac{p \cdot \widehat{\sigma}^2_{k=0}}{\sum_{i=0}^{p}[\widehat{\beta}_i(k=0)]^2}, \tag{2.9}$$

where $\widehat{\sigma}^2_{k=0}$ and $\widehat{\beta}_i(k=0)$ are $\widehat{\sigma}^2$ and $\widehat{\beta}_i$ from traditional OLS estimation $(k=0)$.

This value of k, as Hoerl et al. (1975) reveal, emerges with the purpose of stabilising the estimations and obtaining a lower value of the mean square error (MSE) than the OLS estimator.

As Pasha and Shah (2004) state, *in general, there is an "optimum" value of k for any problem, but it is desirable to examine the ridge solution for a range of admissible values of k*. Although the traditional k value used in the literature is the one proposed by Hoerl et al. (1975), as noted above, many authors have proposed different algorithms to obtain the biasing parameter k (see Kibria and Banik (2016)). Nevertheless, these k values do not always mitigate the existing collinearity and sometimes the indications of Marquardt (1970), who argues that the maximum VIF must be lower than 10, are ignored. With the purpose of proposing new k values with good properties, García et al. (2019b) propose three new ridge factors based on the determinant of correlation matrix. See this work and its references for an in-depth look at ridge regression and its properties.

The main disadvantage of ridge regression is that the decomposition of the sum of squares cannot be verified, and hence the calculation of R^2 could be questioned, thus using ridge regression, the global significance and also the inference of the model may not be interpreted (see Jensen and Ramírez (2008); Rodríguez et al. (2019)). Furthermore, the estimators obtained are difficult to interpret because the procedure does not use the original variables or some interpretable variation thereof. As Kidwell and Brown (1982) stated, although ridge regression is presented as a good technique for dealing with multicollinearity problems, it present some weaknesses: *a ridge regression solution can produce results that are different from the OLS solution when the predictors are not orthogonal. The different solutions each suggest a different interpretation of the data. [...] It should be noted that application of the*

ridge solution does not necessarily produce the correct answer. Hence, the interpretation of the results is not a trivial issue. As Deegan Jr (1975) argues, ridge regression is more suitable as an additional methodology than as the principal methodology for interpreting the results. According to Kidwell and Brown (1982), *the use of ridge regression procedure can suggest directions for further investigation and/or further theoretical development that may not be apparent from least squares solution*. Ridge regression penalises the least squares regression with an additional value on the size of the regression weights, but as Kiers and Smilde (2007) reveal, *a mere shrinkage by itself may not solve the problem of poor performance in data*.

Finally, another widespread error when applying the ridge regression, is to calculate the VIF with the expression proposed by Marquardt (1970) that can lead to values of VIF lower than 1, which is inconsistent with the definition of this diagnostic measure (see among others, Ma et al. (2017)). Due to this issue, for the empirical part of this dissertation, the VIF in the case of the application of ridge regression will be obtained from the expression proposed by García et al. (2015), who solve this problem, instead of the widely applied expression proposed by Marquardt (1970).

2.3.2 LASSO regression

The LASSO (Least Absolute Shrinkage and Selection Operator) regression was proposed by Tibshirani (1996). As has been noted in the previous section, this author affirms that one of the reasons why the data analyst is often not satisfied with the OLS estimates is prediction accuracy. Prediction accuracy *can sometimes be improved by shrinking or setting to 0 some coefficients*. Ridge regression shrinks coefficients, *however, it does not set any coefficients to 0*

and hence does not give an easily interpretable model, Tibshirani (1996). By using subset selection, which refers to the technique of finding a small subset of the set of well-functioning explanatory variables in predicting the dependent variable, the models will be more interpretable but, because of the discrete nature of this process, the results can be extremely variable. Because of these deficiencies a new technique appears: the LASSO.

LASSO methodology seems to "delete" some observed explanatory variables of the model. It *shrinks some coefficients and sets others to 0, and hence tries to retain the good features of both subset selection and ridge regression*, Tibshirani (1996). Thus, it could be interpreted as an alternative methodology that solves the algebraic problem, like the ridge regression.

Assuming $p-1$ explanatory variables and standardized data. The LASSO problem, based on the ℓ_1-norm, is defined by:

$$\widehat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_n \left(\mathbf{Y} - \sum_{i=2}^{p} \beta_i \mathbf{x}_i \right)^2 \right\}$$
$$\text{s.t. } \sum_{i=1}^{p} |\beta_i| \leq t,$$

where t is a tuning parameter ($t \geq 0$) and it controls the size of the shrinkage. If $\sum |\widehat{\boldsymbol{\beta}}^{\text{OLS}}|$ is named as t^{OLS}, then values of $t < t^{\text{OLS}}$ *will cause shrinkage of the solutions towards to 0, and some coefficients may be exactly equal to 0*, Tibshirani (1996).

The solution to the previous problem can be gained from the orthonormal design case. Supposing that $\mathbf{x}^t\mathbf{x} = \mathbf{I}$, then:

$$\widehat{\beta}_i = \widehat{\beta}_i^{\text{OLS}} \max\left(0\ ,\ 1 - \frac{n\ \zeta}{|\widehat{\beta}_i^{\text{OLS}}|}\right),$$

where ζ represents the Lagrange multiplier.

As Tibshirani (1996) confirms, it is difficult to obtain the standard errors because the LASSO is a non-linear and non-differentiable function of the response values. In addition, this author says that *ridge regression scales the coefficients by a constant factor, whereas the LASSO translates by a constant factor, truncating at 0.* If the LASSO solution is approximated by the ridge regression, for predictors with $\widehat{\beta}_i = 0$, the estimated variance will also be zero.

Another important problem is that, as Lockhart et al. (2014) state, *the usual constructs like p-values, confidence intervals, etc., do not exist for LASSO estimates*, so it is not possible to make conclusions about the global significance and inference.

As Hans (2009) states, the LASSO regression is a widely used alternative to OLS estimation when regression problems are observed. It is another shrinkage or regularisation solution but it differs from ridge regression: while ridge regression approximates some estimated parameters to zero without excluding any of them, LASSO regression can exclude some of them. LASSO is recommended for models with a low number of predictors with substancial standardized coefficients (there are differences between the values of the estimated parameters), while ridge is recommended when there is no differences between predictors, Zou and Hastie (2005). As Dormann et al. (2013) affirm, *depending on the form of the penalty, the regression coefficients are shrunk and/or selected [...]. Ridge regression performs neither selection nor grouping, while LASSO selects but does not group parameters*, thus LASSO regression and ridge regression may be interpreted, in some way, as similar methodologies.

Finally, as has been said regarding ridge regression, *a mere shrinkage by itself may not solve the problem of poor performance in data*, Kiers and Smilde (2007). In addition, these methods *require the use of marginal statistics to estimate regression coefficients or determine the relative importance of*

individual explanatory variables, and thus offer no refuge from associated biases due to multicollinearity, Graham (2003).

2.3.3 Principal Component Regression (PCR)

PCR was developed by Pearson (1901) and Hotelling (1933). It consists, basically, in converting a set of explanatory variables into a set of orthogonal components, by deleting information from each variable or transforming it (Mittelhammer et al. (1980)).

PCR is a technique that is based on principal component analysis (PCA). Typically, it considers regressing the dependent variable on a set of explanatory variables, indeed it is based on a standard linear regression model. As Geladi and Kowalski (1986) say, PCA is a method that leads the researcher to rewrite a matrix $\mathbf{X}$ of rank p as a sum of p matrices of rank 1. Each new matrix, can be rewritten as a product of two vectors: the score, s, and the loading, d. The final result will be *an operator that projects the columns of* $\mathbf{X}$ *onto a single dimension and an operator that projects the rows of* $\mathbf{X}$ *onto a single dimension*, s and d respectively (Geladi and Kowalski (1986)).

Nonlinear iterative partial least squares (NIPALS) can be used to obtain the vectors s and d for each variable. This procedure calculates s_1 and d_1 from $\mathbf{X}$, then the residual ($\mathbf{e}_1$) is calculated as $\mathbf{X} - s_1\ {d_1}^t$, and this residual can be used to obtain s_2 and d_2 as $\mathbf{e}_2 = \mathbf{e}_1 - s_2\ {d_2}^t$ (Geladi and Kowalski (1986)).

Thus, PCR uses the matrix $\mathbf{S}$ of scores, which is $\mathbf{S} = \mathbf{X} \cdot \mathbf{D}$, where $\mathbf{D}$ is the loadings matrix. The multiple linear regression that starts from the model (1.1) can thus be rewritten as $\mathbf{Y} = \mathbf{S}\boldsymbol{\beta} + \mathbf{u}$. Note that the original matrix $\mathbf{X}$ is replaced by $\mathbf{S}$, which has better properties (orthogonality). So, PCR solves the collinearity problem by substituting the original explanatory variables with

orthogonal components, but it *has the risk that useful (predictive) information will end up in discarded principal components and that some noise will remain in the components used for regression*, Geladi and Kowalski (1986). Works like Mardia et al. (1979) or Draper and Smith (1981) detail the PCR procedure.

One commonly-used method is Kendall method *of regressing the dependent variable on the subset of "significant" components obtained from PCA*, Haitovsky (1969). With regard to this method, Farrar and Glauber (1967) comment that it is dangerous to reject or delete the non-significative components because it is likely the researcher will miss important information from each variable, and even the problem could be exacerbate. Furthermore, these authors also state that the usefulness of this method is limited to the situation in which the significant components may be interpreted directly as economic phenomena. Indeed, Gimenez and Giussani (2018) highlight the difficulty of interpreting the coefficients obtained with PCR.

Partial least squares (PLS) appears as a particular application of PCR. PLS regression was presented by Wold (1966): is built on the properties of the NIPALS algorithm and produces factor scores that are linear combinations of the original independent variables, such that there is no correlation between the new predictor factors. PLS is convenient when there are more predictor variables than observations and when multicollinearity exists in the model. See Geladi and Kowalski (1986) for a complete explanation of PLS procedure.

To observe the importance of each component in the model, the variable importance in projection (VIP) coefficient needs to be studied. It reflects the importance of each explanatory variable in fitting both $\mathbf{X}$ and $\mathbf{Y}$, as $\mathbf{Y}$ is predicted using $\mathbf{X}$. Therefore, VIP enables the classification of the independent variables according to their explanatory power for $\mathbf{Y}$. Variables with VIP scores close to or greater than one are considered to be important in a given

model, making them the most relevant predictors for explaining $\mathbf{Y}$. Variables with VIP scores significantly less than one are less important, making them good candidates for exclusion from the model (Chong and Jun (2005)).

These procedures have some disadvantages. The traditional PCA, which is the basis of PCR and PLS, is a linear projection method, and it may not be capable of efficiently capturing the non-linear features existing in real data, as Liu et al. (2012) or Deng et al. (2013) affirm. Additionally, Yuan et al. (2015) state that *PCA is developed in a deterministic manner, which lacks a probabilistic interpretation for modelling data.* It should be taken into account that the principal components are linear combinations of the original variables and the method produces misleading results regarding the empirical interpretation: the researcher is not measuring the original variables, but artificial ones (Bitetto et al. (2016); Chatfield (1995); Dormann et al. (2013); Graham (2003); Gimenez and Giussani (2018); Vigneau et al. (1997)). Hawkins (1973) argues that, with the introduction of new components, each component is identified with some part of the independent variables, but is there *a guarantee that the dependent variable is dependent on the [components] rather than on the near multicollinearities which have been ignored*? In addition, this author also suggests that these techniques give *no explicit information on the number or composition of the alternative good subsets.* Furthermore, Artigue and Smith (2019) have recently revealed that the principal problem is the reliability of the methodology in making predictions. This idea is also supported in Kiers and Smilde (2007): PCA gives *the poorest recoveries of regression weight in conditions with relatively low noise and collinearity*, and the authors stated that *prediction suffers far less from collinearity than recovery of the regression weights.* Finally, Artigue and Smith (2019) also demonstrate with Monte Carlo

Simulations that the larger the number of potential explanatory variables, the less effective and more likely to be misleading PCR will be.

2.3.4 Raise regression

Raise regression has been presented by García et al. (2011) and fully developed in Salmerón et al. (2017a). It is an alternative methodology to estimate models with multicollinearity, solving it from a geometrical point of view.

Instead of deleting data that may contain prior information, raise regression maintains the available information and modifies the problematic variables. If essential multicollinearity problems have not been completely mitigated after raising one variable, it is possible to raise more variables of the model. This particular procedure is known as successive raising (see García and Ramírez (2017)).

Starting from the linear model (2.1), with two explanatory variables plus the intercept ($p = 3$), the collinearity problem arises because vector $\mathbf{X}_2$ and vector $\mathbf{X}_3$ are very close geometrically, i.e. the angle that determines both vectors, θ_1, is very small (see Figure 2.1).

The raise regression tries to separate both explanatory variables through the auxiliary regression (2.5), whose estimation by OLS leads to the estimated residuals $\mathbf{e}_2$. The raise vector is thus defined as:

$$\widetilde{\mathbf{X}}_2 = \mathbf{X}_2 + \lambda \mathbf{e}_2,$$

where $\lambda > 0$.

The raise model will be obtained by substituting vector $\mathbf{X}_2$ by the raise vector $\widetilde{\mathbf{X}}_2$ in the original model. This is to say:

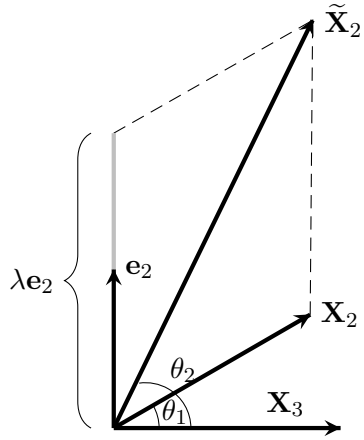

Figure 2.1: The raise method.

$$\mathbf{Y} = \gamma_1 + \gamma_2\widetilde{\mathbf{X}}_2 + \gamma_3\mathbf{X}_3 + \mathbf{w},$$

or in its matrix form:

$$\mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\gamma} + \mathbf{w}.$$

All the global characteristics of the original model are maintained. In summary, the following values do not change:

- The square sums of the residuals of the original model.
- The estimated variance of the random disturbance.
- The coefficient of determination, R^2.
- The global significance test.
- Prediction.

Chapter 3

Residualization: some criticism and methodological preliminaries

To briefly explain the general concept of the method, it might be said that by residualizing one of the explanatory variables, its effect is being isolated from the rest of the variables of the model. Thus, the part of this variable that has no relationship with the rest of independent variables is being included in the model, leading to a new interpretation of the residualized variable.

This and next chapter fully develop the residualization procedure and justifies its application not only for dealing with multicollinearity but also for separating the individual effects of the predictor variables.

It is important to point out that the application of residualization leads to conclusions about a model different to the original even though they have several identical characteristics (such as the variance estimation of the random perturbation, the coefficient of determination or the significance statistics).

To sum up, the main contributions when applying this technique are:

- The new interpretations of the coefficients. The residualized model can answer questions that could not be answered with the initial model. However, it is relevant to note that residualization is not always applicable because the interpretations of the new estimated coefficients are not always possible. This fact will be further analysed in the coming chapters.

- The isolation of the individual effect of the explanatory variables (the fulfillment of the principle *ceteris paribus*).

- The possibility of reducing the degree of near collinearity in the initial model.

The present chapter provides a first overview of the methodology: how it works and what are the results obtained. The estimation and inference of the multiple linear regression model using the residualization procedure will be considered exhaustively in Chapter 4.

The structure of this chapter is as follows: Section 3.1 examines some criticism of the residualization procedure. Section 3.2 introduces the reader to the methodology, explaining the procedure of centring explanatory variables and the link between this and residualization. Sections 3.3 and 3.4 explain the application of the residualization procedure focusing on the mitigation of the existing collinearity problem in a model with two and three standardized explanatory variables, respectively, and finally, Section 3.5 takes an in-depth look at the new interpretations of the coefficients, which will be also shown in coming chapters. Sections 3.3 to 3.5 are based on the work by Salmerón et al. (2016).

3.1 Criticism of residualization

Residualization has been briefly explained by Hill and Adkins (2003), who based their research on works by Kennedy (1982) and Buse (1994).

The first criticism to residualization that deserves to be mentioned is the one in the work by Buse (1994). The author states that the residualization procedure *is a cure for collinearity that is potentially worse than the disease*. The author concludes that there is no guarantee that *insignificant coefficients will become significant after orthogonalization* and he also concludes that the estimated variances may increase. However, the decrease in the estimated variances when the residualization is applied is demonstrated in this dissertation (see Section 4.2.1 for more details). It is true that there is no guarantee that non-significant coefficients (as a result of the presence of strong collinearity) might not become significant after the residualization procedure, but it is demonstrated that the individual significance of unaltered variables of the model change (see Section 4.1.3 for more details). In any case, as Kennedy (1982) states, residualization should be applied with the aim of isolating the individual effect of the explanatory variables, not only to avoid multicollinearity problems. The application of the procedure leads to new interpretations of the residualized variables, which imply the fulfillment of the principle *ceteris paribus*, as has been noted.

On the other hand, Wurm and Fisicaro (2014) state that *residualization of predictor variables is not the hoped-for panacea* to collinearity. As the authors comment, residualization *creates an analysis that is neither simultaneous nor hierarchical in terms of the original variables, but which blends aspects of both. Specifically, residualizing exaggerates the statistical importance of the non-residualized predictor in a region of Redundancy or Suppression, and*

underestimates it in a region of Enhancement (as defined by Friedman and Wall (2005)). Although non-residualized variables change their estimations with regard to the original model, it will be demonstrated in following chapters that these new results correspond to the model without the problematic variable. Hence, the new model does not exaggerate the statistical importance of the non-residualized variables because it operates as if the problematic variable had been deleted, which is the solution proposed in, for example, York (2012). Wurm and Fisicaro (2014) also state that *residualizing replaces the problem of collinearity (to the extent that it is a problem) with one that is less obvious and less well-understood. For these reasons, residualizing sometimes creates conceptual difficulty and leaves the researcher unable to draw any firm conclusions.* In this dissertation, the authors fully explain the new interpretations of the new residualized variables, showing that these new interpretations are sometimes more suitable for the research in progress. Additionally, Wurm and Fisicaro (2014) note that residualization *does not create an improved, purified, or corrected version of the original predictor*, which is not actually true because the residualized variable represents the part of the variable that has no relationship with the rest of independent variables of the model (it is a purified version of the original variable).

Another criticism of the procedure is the one in York (2012). As this work states, collinearity *reduces the amount of information we have about [the problematic variable] isolated from other factors. A point that is frequently missed is that it is the absolute amount of information that matters, not the proportion*; in any case, although this author comments that the absolute amount of information is what is important, if it is distorted due to the presence of multicollinearity, it is appropriate to mitigate the problem, and thus, pay more attention to the proportion of information which is isolated from

the other factors. Furthermore, in this work, York states that *residualization biases the coefficient estimate and standard error of the residualizer, thereby creating a problem rather than solving one.* It is true that residualization is a biased method, but it is controversial to disclose that it will create a problem: the researcher must determine whether is better to use a distorted model, which will provide unstable results, or to use a biased model; and it depends on the degree of the existing multicollinearity in the model.

These earlier works named other solutions (ridge regression, PCA, deleting variables, etc.), also concluding that they might not solve the problem or they might include other problems in the results. The key point that has not been taken into consideration until now is that residualization provides an alternative interpretation for the estimated parameters, apart from the mitigation of collinearity. This could be seen as a limitation since the methodology is not always applicable, but it can be also seen as an opportunity to obtain new interpretations that cannot be obtained from the initial model.

3.2 Centring explanatory variables

The basic idea of centring explanatory variables consists on subtracting a constant from each value they take. This constant is the mean of each explanatory variable. With this procedure, the researcher is redefining the origin of the modified variables. The slope between the modified explanatory variables and the explained variable does not change, but the interpretation of the intercept of the model does. With the modification, the intercept will be the mean of the explained variable (the result when the explanatory variables are equal to zero).

As has been shown above, there is a type of near multicollinearity that

regards the relationship between the constant and the explanatory variables: non-essential collinearity. According to Dalal and Zickar (2012); Iacobucci et al. (2016); Marquardt (1980); Marquardt and Snee (1975); Smith and Campbell (1980); Snee and Marquardt (1984), which offer the idea that the constant is an explanatory variable, centring predictor variables is a good method to mitigate non-essential collinearity, because the researcher is isolating the effect of the constant from the rest of independent variables, i.e. the researcher will mitigate the collinearity that involves the constant from the model.

Salmerón et al. (2019) state that the idea of centring explanatory variables is similar to residualization, so it can be interpreted as a special case of residualization. The authors use the simple linear regression to demonstrate the fact. Let us define the following model:

$$\mathbf{Y} = \beta_1 + \beta_2 \mathbf{X}_2 + \mathbf{u},$$

where $\mathbf{u}$ is spherical and there are n observations. Let us also define an auxiliary regression where the dependent variable is $\mathbf{X}_2$ and the independent variables would be the rest of explanatory variables from the previous model; in this case, as only the constant is taken into account (i.e. a matrix of ones with dimension $n \times 1$, noted as $\mathbf{1}$), the auxiliary regression will be:

$$\mathbf{X}_2 = \alpha \cdot \mathbf{1} + \mathbf{v}.$$

By applying the OLS estimation, $\widehat{\boldsymbol{\alpha}} = (\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\mathbf{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i} = \overline{\mathbf{X}}_2$, and it is verified that the estimated error is $\mathbf{e} = \mathbf{X}_2 - \overline{\mathbf{X}}_2$.

As was briefly explained in Chapter 1, residualization consists in replacing the problematic variables in the model with their estimated residuals (from an auxiliary regression). In the case of simple linear regression, the estimated

residuals are $\mathbf{e} = \mathbf{X}_2 - \overline{\mathbf{X}}_2$, so the final residualized model will be:

$$\begin{aligned}\mathbf{Y} &= \gamma_1 + \gamma_2\mathbf{e} + \mathbf{w} \\ &= \gamma_1 + \gamma_2(\mathbf{X}_2 - \overline{\mathbf{X}}_2) + \mathbf{w}.\end{aligned}$$

With this, it is demonstrated that in the simple linear regression, residualization coincides with centring variable $\mathbf{X}_2$, which mitigates non-essential collinearity.

3.3 Residualization for two standardized explanatory variables ($p = 2$)

Residualization was presented in Novales et al. (2015) for the linear model with two standardized explanatory variables (i.e. the standardized version of model (2.1)):

$$\mathbf{Y} = \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \mathbf{u}. \tag{3.1}$$

The reason of using standardized variables is to obtain expressions easy to interpret from the correlations among the explanatory variables of the model. Therefore, the use of standardized variables makes the constant disappears.

Suppose that $\mathbf{x}_3$ is the variable to be residualized. The auxiliary model will be $\mathbf{x}_3 = \alpha_2\mathbf{x}_2 + \mathbf{v}$, whose estimation by OLS leads to the estimated residuals $\mathbf{e}_3$, which are orthogonal to $\mathbf{x}_2$. Thus the final model (the residualized one) will be: $\mathbf{Y} = \gamma_2\mathbf{x}_2 + \gamma_3\mathbf{e}_3 + \mathbf{w}$.

Table 3.1 presents all the relevant characteristics of the original model and the residualized model. Note that the residualization procedure makes the estimated variances of the parameters diminish, while the global characteristics of the model remain unchanged ($\widehat{\sigma}^2$, R^2 and F_{exp}). Furthermore, the estimated

Table 3.1: Main characteristics of the original model and the residualized model, with two standardized explanatory variables.

Original model	Residualized model
$\mathbf{Y} = \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \mathbf{u}$	$\mathbf{Y} = \gamma_2\mathbf{x}_2 + \gamma_3\mathbf{e}_3 + \mathbf{w}$
$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \frac{\varrho_2 - \rho\varrho_3}{1-\rho^2} \\ \frac{\varrho_3 - \rho\varrho_2}{1-\rho^2} \end{pmatrix}$	$\widehat{\boldsymbol{\gamma}} = \begin{pmatrix} \varrho_2 \\ \frac{\varrho_3 - \rho\varrho_2}{1-\rho^2} \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_2 + \rho\widehat{\beta}_3 \\ \widehat{\beta}_3 \end{pmatrix}$
$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2 \begin{pmatrix} \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} \\ -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{pmatrix}$	$\widehat{Var}(\widehat{\boldsymbol{\gamma}}) = \widehat{\sigma}^2_O \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{1-\rho^2} \end{pmatrix}$
$\widehat{\sigma}^2 = \frac{1-\rho^2-\varrho_2^2-\varrho_3^2+2\rho\varrho_2\varrho_3}{(n-2)(1-\rho^2)}$	$\widehat{\sigma}^2_O = \frac{1-\rho^2-\varrho_2^2-\varrho_3^2+2\rho\varrho_2\varrho_3}{(n-2)(1-\rho^2)} = \widehat{\sigma}^2$
$R^2 = \frac{\varrho_2^2+\varrho_3^2-2\rho\varrho_2\varrho_3}{1-\rho^2}$	$R^2_O = \frac{\varrho_2^2+\varrho_3^2-2\rho\varrho_2\varrho_3}{1-\rho^2} = R^2$
$\widehat{\beta}_i \pm t_{n-2}\left(1-\frac{\alpha}{2}\right)\widehat{\sigma}\sqrt{\frac{1}{1-\rho^2}} \quad \forall i$	$\widehat{\gamma}_2 \pm t_{n-2}\left(1-\frac{\alpha}{2}\right)\widehat{\sigma}$ $\widehat{\gamma}_3 \pm t_{n-2}\left(1-\frac{\alpha}{2}\right)\widehat{\sigma}\sqrt{\frac{1}{1-\rho^2}}$
$F_{exp} = \frac{(n-2)\cdot(\varrho_2^2+\varrho_3^2-2\rho\varrho_2\varrho_3)}{1-\rho^2-\varrho_2^2-\varrho_3^2+2\rho\varrho_2\varrho_3} > F_{1,n-2}(1-\alpha)$	$F_{exp,O} = \frac{(n-2)\cdot(\varrho_2^2+\varrho_3^2-2\rho\varrho_2\varrho_3)}{1-\rho^2-\varrho_2^2-\varrho_3^2+2\rho\varrho_2\varrho_3} = F_{exp} > F_{1,n-2}(1-\alpha)$
$\text{VIF}_i = \frac{1}{1-\rho^2} \quad \forall i$	$\text{VIF}_{i,O} = 1 \quad \forall i$
$\text{CN} = \sqrt{\frac{1+\rho}{1-\rho}}$	$\text{CN}_O = 1$

ρ corresponds to the correlation between variables $\mathbf{x}_2$ and $\mathbf{x}_3$, and ϱ_i corresponds to the correlation between $\mathbf{x}_i$ and $\mathbf{Y}$, for $i = 2, 3$.
α represents the significance level.
The subindex O regards to the residualized model.

parameter for the modified variable also has the same value in the original model and in the residualized model, which is not the case of the other explanatory variable ($\widehat{\beta}_2 \neq \widehat{\gamma}_2$ while $\widehat{\beta}_3 = \widehat{\gamma}_3$).

On the other hand, although this technique allows the researcher to deal with strong near collinearity problems, the new estimation of the variables has a different interpretation. In particular, the researcher is able to address non-analysed questions in the original model, i.e. even when there is no strong

collinearity in the model, the researcher may implement the method to obtain new interpretations of the explanatory variables. This fact is explained in depth in Section 3.5.

3.4 Residualization for three standardized explanatory variables ($p = 3$)

Suppose three standardized predictor variables (in order to obtain similar conclusions as in Table 3.1) and n observations. Assume also that there is high or strong essential collinearity. The steps to follow are:

1. Select the variable that is going to be residualized (the dependent variable in the auxiliary regression).

2. After residualizing the chosen variable, it is necessary to check whether the problem has been mitigated. To this end, this chapter uses the VIF.

3. If the problem persists, it is necessary to select another explanatory variable to be residualized (successive residualization).

3.4.1 Step 1: residualization

Suppose the following model:

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u} = \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \mathbf{u}, \tag{3.2}$$

with:

$$\mathbf{x}^t\mathbf{x} = \begin{pmatrix} 1 & \rho_{23} & \rho_{24} \\ \rho_{23} & 1 & \rho_{34} \\ \rho_{24} & \rho_{34} & 1 \end{pmatrix}, \quad \mathbf{x}^t\mathbf{Y} = \begin{pmatrix} \varrho_2 \\ \varrho_3 \\ \varrho_4 \end{pmatrix}, \tag{3.3}$$

where ρ_{ij} is the coefficient of correlation between variables $\mathbf{x}_i$ and $\mathbf{x}_j$ ($i, j = 2, 3, 4$, $i \neq j$), and ϱ_i is the coefficient of determination between variables $\mathbf{Y}$ and $\mathbf{x}_i$ ($i = 2, 3, 4$).

The OLS estimator of model (3.2) will be:

$$\widehat{\boldsymbol{\beta}} = C \begin{pmatrix} (1-\rho_{34}^2)\varrho_2 - (\rho_{23}-\rho_{24}\rho_{34})\varrho_3 + (\rho_{24}-\rho_{23}\rho_{34})\varrho_4 \\ -(\rho_{23}-\rho_{24}\rho_{34})\varrho_2 + (1-\rho_{24}^2)\varrho_3 - (\rho_{34}-\rho_{23}\rho_{24})\varrho_4 \\ (\rho_{24}-\rho_{23}\rho_{34})\varrho_2 - (\rho_{34}-\rho_{23}\rho_{24})\varrho_3 + (1-\rho_{23}^2)\varrho_4 \end{pmatrix}, \tag{3.4}$$

where $C = \frac{1}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2}$.

Suppose $\mathbf{x}_4$ is the variable to be residualized. The auxiliary regression will be $\mathbf{x}_4 = \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3 + \mathbf{v}$, which implies $\mathbf{x}_4 = \widehat{\alpha}_2\mathbf{x}_2 + \widehat{\alpha}_3\mathbf{x}_3 + \mathbf{e}_4$, where $\mathbf{x}_2$ is orthogonal to $\mathbf{e}_4$ and $\mathbf{x}_3$ is orthogonal to $\mathbf{e}_4$ ($\mathbf{x}_2 \perp \mathbf{e}_4$, $\mathbf{x}_3 \perp \mathbf{e}_4$), and:

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_{24} \\ \rho_{34} \end{pmatrix} = \frac{1}{1-\rho_{23}^2} \begin{pmatrix} \rho_{24}-\rho_{23}\rho_{34} \\ \rho_{34}-\rho_{23}\rho_{24} \end{pmatrix}. \tag{3.5}$$

Then, the OLS estimator of the residualized model:

$$\mathbf{Y} = \mathbf{x}_O\boldsymbol{\gamma} + \mathbf{w} = \gamma_2\mathbf{x}_2 + \gamma_3\mathbf{x}_3 + \gamma_4\mathbf{e}_4 + \mathbf{w}, \tag{3.6}$$

with:

$$\mathbf{x}_O^t\mathbf{x}_O = \begin{pmatrix} 1 & \rho_{23} & 0 \\ \rho_{23} & 1 & 0 \\ 0 & 0 & \mathbf{e}_4^t\mathbf{e}_4 \end{pmatrix}, \quad \mathbf{x}_O^t\mathbf{Y} = \begin{pmatrix} \varrho_2 \\ \varrho_3 \\ \mathbf{e}_4^t\mathbf{Y} \end{pmatrix}, \tag{3.7}$$

will be:

$$\widehat{\boldsymbol{\gamma}} = \begin{pmatrix} \frac{\varrho_2-\rho_{23}\varrho_3}{1-\rho_{23}^2} \\ \frac{\varrho_3-\rho_{23}\varrho_2}{1-\rho_{23}^2} \\ \frac{\mathbf{e}_4^t\mathbf{Y}}{\mathbf{e}_4^t\mathbf{e}_4} \end{pmatrix}. \tag{3.8}$$

Taking into account results of A.1 from Appendix A:

$$\mathbf{e}_4^t\mathbf{Y} = \varrho_4 - \frac{\varrho_2 - \rho_{23}\varrho_3}{1-\rho_{23}^2}\varrho_2 - \frac{\varrho_3 - \rho_{23}\varrho_2}{1-\rho_{23}^2}\varrho_3,$$
$$\mathbf{e}_4^t\mathbf{e}_4 = \frac{1 + 2\rho_{23}\rho_{24}\rho_{34} - \rho_{23}^2 - \rho_{24}^2 - \rho_{34}^2}{1-\rho_{23}^2},$$

it is clear that:

$$\widehat{\boldsymbol{\gamma}} = \begin{pmatrix} \frac{\varrho_2 - \rho_{23}\varrho_3}{1-\rho_{23}^2} \\ \frac{\varrho_3 - \rho_{23}\varrho_2}{1-\rho_{23}^2} \\ \frac{(\rho_{24}-\rho_{23}\rho_{34})\varrho_2 - (\rho_{34}-\rho_{23}\rho_{24})\varrho_3 + (1-\rho_{23}^2)\varrho_4}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2} \end{pmatrix}. \tag{3.9}$$

As in the case of Section 3.3, the estimated parameter for the residualized variable remains unchanged:

$$C \cdot \widehat{\beta}_4 = \widehat{\gamma}_4 = \frac{(\rho_{24} - \rho_{23}\rho_{34})\varrho_2 - (\rho_{34} - \rho_{23}\rho_{24})\varrho_3 + (1-\rho_{23}^2)\varrho_4}{1 + 2\rho_{23}\rho_{24}\rho_{34} - \rho_{23}^2 - \rho_{24}^2 - \rho_{34}^2},$$

while the parameters of the rest of the explanatory variables have changed and they are the same as those in the model $\mathbf{Y} = \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \mathbf{u}$ (see Table 3.1).

Furthermore, taking into account that:

$$\left(\mathbf{x}^t\mathbf{x}\right)^{-1} = C \begin{pmatrix} 1-\rho_{34}^2 & -(\rho_{23}-\rho_{24}\rho_{34}) & \rho_{24}-\rho_{23}\rho_{34} \\ -(\rho_{23}-\rho_{24}\rho_{34}) & 1-\rho_{24}^2 & -(\rho_{34}-\rho_{23}\rho_{24}) \\ \rho_{24}-\rho_{23}\rho_{34} & -(\rho_{34}-\rho_{23}\rho_{24}) & 1-\rho_{23}^2 \end{pmatrix},$$

$$\left(\mathbf{x}_O^t\mathbf{x}_O\right)^{-1} = \begin{pmatrix} \frac{1}{1-\rho_{23}^2} & -\frac{\rho_{23}}{1-\rho_{23}^2} & 0 \\ -\frac{\rho_{23}}{1-\rho_{23}^2} & \frac{1}{1-\rho_{23}^2} & 0 \\ 0 & 0 & \frac{1-\rho_{23}^2}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2} \end{pmatrix},$$

the inference of the parameters from both models (through both confidence intervals and significance tests) does not change for the residualized variable:

$$\left(\mathbf{x}^t\mathbf{x}\right)^{-1}_{(3,3)} = \left(\mathbf{x}_O^t\mathbf{x}_O\right)^{-1}_{(3,3)} = C \cdot (1-\rho_{23}^2) = \frac{1-\rho_{23}^2}{1 + 2\rho_{23}\rho_{24}\rho_{34} - \rho_{23}^2 - \rho_{24}^2 - \rho_{34}^2},$$

where $\left(\mathbf{x}^t\mathbf{x}\right)^{-1}_{(3,3)}$ and $\left(\mathbf{x}_O^t\mathbf{x}_O\right)^{-1}_{(3,3)}$ represent the element $(3,3)$ of each matrix. For the rest of explanatory variables, the inference of the parameters coincides with those in the model $\mathbf{Y} = \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \mathbf{u}$ (see Table 3.1).

On the other hand, the total sum of squares from both models is the same (due to the explained variable is the same in both models), as well as the explained sum of squares (see A.2 from Appendix A), the residual sum of squares, the estimated variance of the random disturbance, the coefficient of determination and the global significance test.

3.4.2 Step 2: check if the problem persists

It is clear that in model (3.6) there is no relationship between variable $\mathbf{e}_4$ and the other two explanatory variables, but what if another relationship exists between the other two variables that makes the problem of strong collinearity persist? To answer this question, this chapter is going to use the VIF. To determine the values of the VIF, it is necessary to obtain the coefficient of determination from the following auxiliary regressions:

$$\begin{aligned}\mathbf{x}_2 &= \alpha_2\mathbf{x}_3 + \alpha_3\mathbf{e}_4 + \mathbf{v},\\ \mathbf{x}_3 &= \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{e}_4 + \mathbf{v},\\ \mathbf{e}_4 &= \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3 + \mathbf{v}.\end{aligned} \tag{3.10}$$

For the first two auxiliary models from (3.10), it is obtained that:

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{e}_4^t\mathbf{e}_4 \end{pmatrix}^{-1} \begin{pmatrix} \rho_{23} \\ 0 \end{pmatrix} = \begin{pmatrix} \rho_{23} \\ 0 \end{pmatrix},$$

thus $\text{SSE} = (\rho_{23}\ 0)\begin{pmatrix} \rho_{23} \\ 0 \end{pmatrix} = \rho_{23}^2$. As the total sum of squares is equal to one, the coefficient of determination has the same value as the explained sum

of squares, i.e. it is equal to ρ_{23}^2. Hence, the VIF will be:

$$\text{VIF}_{i,O} = \frac{1}{1 - \rho_{23}^2}, \quad i = 2, 3. \tag{3.11}$$

As the reader may observe, the obtained results coincide with those in the model $\mathbf{Y} = \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \mathbf{u}$ (see Table 3.1).

Regarding the third auxiliary regression from (3.10), it is clear that its coefficient is zero, due to $\mathbf{e}_4$ being orthogonal to $\mathbf{x}_2$ and $\mathbf{x}_3$ so its VIF is equal to one (the minimum value of the VIF).

Therefore, from (3.11), if $\rho_{23} > 0.9$ then $\text{VIF}_{i,O} > 10$ $(i = 2, 3)$ meaning the problem of strong collinearity persists.

On the other hand, the VIFs from model (3.2) would be obtained after calculating the coefficient of determination in the following auxiliary regression:

$$\mathbf{x}_i = \alpha_2 \mathbf{x}_j + \alpha_3 \mathbf{x}_q + \mathbf{v}, \tag{3.12}$$

with $i, j, q = 2, 3, 4$ $i \neq j,\ i \neq q,\ j \neq q$. As the coefficient of determination of the two first models from (3.10) is the same as that from the regressions $\mathbf{x}_i = \alpha \mathbf{x}_j + \mathbf{v}$ $(i, j = 2, 3\ i \neq j)$, it is verified that it will be less than the one from the regressions in (3.12) because the introduction of new variables in a multiple regression implies a higher coefficient of determination. In the third regression from (3.10) this fact is trivial since the coefficient of determination is always higher or equal to zero.

To sum up, when residualizing the model, it is certain that the VIFs will decrease, with the one from the residualized variable being equal to one (the minimum value).

3.4.3 Step 3: successive residualization

If the problem has not been mitigated by residualizing one variable, it is necessary to residualize another one.

The following auxiliary regression is proposed:

$$\mathbf{x}_3 = \alpha_2 \mathbf{x}_2 + \mathbf{v},$$

thus $\mathbf{x}_3 = \widehat{\alpha}_2 \mathbf{x}_2 + \mathbf{e}_3$ with $\widehat{\alpha}_2 = \rho_{23}$ and $\mathbf{e}_3 \perp \mathbf{x}_2$. As $\mathbf{e}_4 \perp \mathbf{x}_2$ and $\mathbf{e}_4 \perp \mathbf{x}_3$, it is verified that $\mathbf{e}_3 \perp \mathbf{x}_4$.

The double residualized model will be:

$$\mathbf{Y} = \delta_2 \mathbf{x}_2 + \delta_3 \mathbf{e}_3 + \delta_4 \mathbf{e}_4 + \boldsymbol{\varsigma}. \tag{3.13}$$

Being $\mathbf{x}_{OO} = (\mathbf{x}_2 \ \mathbf{e}_3 \ \mathbf{e}_4)$, then:

$$\mathbf{x}_{OO}^t \mathbf{x}_{OO} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mathbf{e}_3^t \mathbf{e}_3 & 0 \\ 0 & 0 & \mathbf{e}_4^t \mathbf{e}_4 \end{pmatrix}, \quad \mathbf{x}_{OO}^t \mathbf{Y} = \begin{pmatrix} \varrho_2 \\ \varrho_3 - \rho_{23}\varrho_2 \\ \mathbf{e}_4^t \mathbf{Y} \end{pmatrix},$$

where $\mathbf{e}_4^t \mathbf{Y}$ and $\mathbf{e}_4^t \mathbf{e}_4$ have been calculated in A.1 from Appendix A, and $\mathbf{e}_3^t \mathbf{e}_3 = (\mathbf{x}_3 - \rho_{23}\mathbf{x}_2)^t (\mathbf{x}_3 - \rho_{23}\mathbf{x}_2) = 1 - \rho_{23}^2$.

By estimating model (3.13) with OLS:

$$\begin{aligned} \widehat{\boldsymbol{\delta}} &= \left(\mathbf{x}_{OO}^t \mathbf{x}_{OO}\right)^{-1} \mathbf{x}_{OO}^t \mathbf{Y} \\ &= \begin{pmatrix} \varrho_2 \\ \frac{\varrho_3 - \rho_{23}\varrho_2}{1-\rho_{23}^2} \\ \frac{(\rho_{24}-\rho_{23}\rho_{34})\varrho_2 - (\rho_{34}-\rho_{23}\rho_{24})\varrho_3 + (1-\rho_{23}^2)\varrho_4}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2} \end{pmatrix}. \end{aligned} \tag{3.14}$$

The estimated parameter of the first residualized variable ($\mathbf{x}_4$) is the same as the one from the original model (3.2), the estimated parameter from the second residualized variable ($\mathbf{x}_3$) is the same as the one from the residualized model

(3.6) and it is also the same as the one from the model $\mathbf{Y} = \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \mathbf{u}$. Also, the estimated parameter of the unchanged variable ($\mathbf{x}_2$) is the same as the one from the model $\mathbf{Y} = \beta_2\mathbf{x}_2 + \mathbf{u}$. And ditto with the inference of that coefficients:

$$\left(\mathbf{x}_{OO}^t\mathbf{x}_{OO}\right)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{1-\rho_{23}^2} & 0 \\ 0 & 0 & \frac{1-\rho_{23}^2}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2} \end{pmatrix}.$$

On the other hand, it is clear that $\mathrm{SST}_{OO} = 1 = \mathrm{SST}$, and as:

$$\begin{aligned}\mathrm{SSE}_{OO} = \widehat{\boldsymbol{\delta}}^t\mathbf{x}_{OO}^t\mathbf{Y} &= \varrho_2^2 + \frac{(\varrho_3 - \rho_{23}\varrho_2)^2}{1-\rho_{23}^2} + \frac{(\mathbf{e}_4^t\mathbf{Y})^2}{\mathbf{e}_4^t\mathbf{e}_4} \\ &= \frac{\varrho_2^2 + \varrho_3^2 - 2\rho_{23}\varrho_2\varrho_3}{1-\rho_{23}^2} + \frac{(\mathbf{e}_4^t\mathbf{Y})^2}{\mathbf{e}_4^t\mathbf{e}_4} = \mathrm{SSE},\end{aligned}$$

it is verified that $\mathrm{SSR}_{OO} = \mathrm{SSR}$, thus, $R_{OO}^2 = R^2$, $\widehat{\sigma}_{OO}^2 = \widehat{\sigma}^2$ and the global significance test is $F_{exp,OO} = F_{exp}$. All the characteristics of the original model still remain unchanged.

3.5 Interpretation of the coefficients: partial and total effects

One of the objetives of residualization is to mitigate multicollinearity problems in a model, without eliminating any explanatory variables. However, by substituting one variable from the corresponding estimated residuals, the understanding of the pertinent estimated parameter has another interpretation. This new interpretation can be very interesting for the researcher, and even when collinearity is not significant, the researcher might want to obtain that new interpretation, which might not be obtained from the original model.

For example, in model (3.6), the parameter γ_4 will be interpreted as the effect on $\mathbf{Y}$ of the part of $\mathbf{x}_4$ unrelated with $\mathbf{x}_2$ and $\mathbf{x}_3$. Similarly, regarding model (3.13), the parameter δ_3 will be interpreted as the effect on $\mathbf{Y}$ of the part of $\mathbf{x}_3$ that has no relationship with $\mathbf{x}_2$ and $\mathbf{x}_4$ (the interpretation of δ_4 will be the same as γ_4).

The interpretation of the modified variable is clear, but what is the interpretation of the unchanged variables? Following the concepts of partial and total effect from Novales (2010):

- What is the total impact on $\mathbf{Y}$ of a unitary variation on $\mathbf{x}_i$, while keeping the rest of the explanatory variables unchanged? Answer: partial effect (multiple regression).

- What is the total impact on $\mathbf{Y}$ of a unitary variation on $\mathbf{x}_i$ if the rest of explanatory variables change, given the observed correlations of the sample? Answer: total effect (simple regression).

Additionally:

- As is shown in Table 3.1, one consequence of residualization is that the estimated parameter of the unchanged variables coincides with the coefficient of correlation between variables $\mathbf{Y}$ and $\mathbf{x}_2$, i.e. it coincides with the one from the model $\mathbf{Y} = \beta_2\mathbf{x}_2 + \mathbf{u}$.

- The estimations of γ_2 and γ_3 in model (3.6) are the same as the estimations from the model $\mathbf{Y} = \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \mathbf{u}$.

- The estimation of δ_2 in model (3.13) is the same as the estimation from the model $\mathbf{Y} = \beta_2\mathbf{x}_2 + \mathbf{u}$.

In light of the above, the partial and total effects regarding the unchanged variable are the same as the ones from the residualized model, which was expected because it is logical to think that the variations of the unchanged variable do not affect the rest due to the existing orthogonality between variables.

Chapter 4

Generalization of the method: residualization for p explanatory variables

As has been stated throughout the earlier chapters, explanatory variables of an econometric model can imply strong near collinearity problems. Even when collinearity diagnostic measures consider that the problem is not of concern, it is possible that the individual effects of the variables may not be separated or displayed clearly. This idea resembles the objective of the Shapley value regression, Shapley (2016), which presents an entirely different strategy for assessing the contribution of predictor variables to the dependent variable and owes its origin to the theory of cooperative games. The value of R^2 obtained by fitting a linear regression model is regarded as the value of a cooperative game played by the independent variables (each variable is a member) against the dependent variable (thus explaining it). The analyst does not have sufficient information to disentangle the contributions made by the individual members,

only their joint contribution R^2 is known. The Shapley value decomposition imputes the most likely contribution of each individual member. On the other hand, Baird and Bieber (2016) proposed an alternative methodology to OLS based on ordered variable regression (OVR), originally presented by Woolf (1951), which fully resolves the issue of related predictors by creating and using variables that are perfectly unrelated.

These antecedents lead to residualization, which is a procedure applied in previous research articles published in major social science journals in many different fields, such as linguistics (Ambridge et al. (2012); Cohen-Goldberg (2012); Jaeger (2010); Kuperman et al. (2008, 2010); Lemhöfer et al. (2008)), environmental issues (Jorgenson (2006); Jorgenson and Burns (2007); Jorgenson and Clark (2009)) or economic development and policies (Bandelj and Mahutga (2010); Bradshaw (1987); Kentor and Kick (2008); Mahutga and Bandelj (2008); Walton and Ragin (1990)). This method has been also applied in previous research under the name of regression with orthogonal variables (see Novales et al. (2015); Salmerón et al. (2016)). However, this method has not been fully developed in prior works and we consider that this lack of specification can lead to different criticisms such as the one in York (2012) or in Wurm and Fisicaro (2014). The key point not taken into consideration until now is that this methodology provides an alternative interpretation for the estimated parameters, apart from the mitigation of collinearity. This could be seen as a limitation since the methodology is not always applicable but it can be also seen as an opportunity to obtain new interpretations which are not possible from the initial model (see Section 3.5 for more details).

The structure of this chapter is as follows: Section 4.1 presents the estimation and main properties of residualization showing that the estimation of the variance of the random disturbance, the global significance test, the

individual significance test of the residualized variable and the goodness of fit obtained by the residualization will be similar to that of the original model. Section 4.2 analyses how residualization mitigates collinearity, demonstrating the decrease in the estimated variance and focusing on the value of the variance inflation factor (VIF) and the condition number (CN) for checking whether the collinearity has been mitigated after the application of residualization. Section 4.3 compares the residualization procedure with OLS and other well-known techniques, such as ridge regression, principal component regression (PCR) or partial least squares regression (PLSR). Finally, Section 4.4 presents the successive residualization procedure for the general case.

This chapter corresponds to the work García et al. (2019c).

4.1 Estimation and properties

Starting from model (1.1), $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, the first step is to define the auxiliary regression (2.5), $\mathbf{X}_i = \mathbf{X}_{-i}\boldsymbol{\alpha} + \mathbf{v}, \quad i = 2, \ldots, p.$

By applying the OLS estimation of the auxiliary regression (2.5), it will be obtained the corresponding estimated residuals, $\mathbf{e}_i$. They will represent the part of variable i that has no relationship with another explanatory variable of the initial model (1.1) since the residuals $\mathbf{e}_i$ are orthogonal to $\mathbf{X}_{-i}$ (that is, $\mathbf{e}_i^t\mathbf{X}_{-i} = \mathbf{0}$, with $\mathbf{0}$ being a zero vector of appropriate dimensions).

In light of the above, residualization procedure consists in replacing variable $\mathbf{X}_i$ with the estimated residuals from model (2.5), $\mathbf{e}_i$, in the original model (1.1). Hence, the residualization procedure uses the following regression[1]:

$$\mathbf{Y} = \mathbf{X}_O\boldsymbol{\gamma} + \mathbf{w}, \tag{4.1}$$

[1]The subindex O regards to the residualized model.

where $\mathbf{X}_O = (\mathbf{X}_{-i}\ \mathbf{e}_i)$.

Once the basic procedure is explained, the results of model (1.1) and model (4.1) are compared in following sections.

4.1.1 Estimation

From $\mathbf{X} = (\mathbf{X}_{-i}\ \mathbf{X}_i)$, the OLS estimator of model (1.1), $\widehat{\boldsymbol{\beta}}$, will be:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y} &= \begin{pmatrix} \mathbf{X}_{-i}^t\mathbf{X}_{-i} & \mathbf{X}_{-i}^t\mathbf{X}_i \\ \mathbf{X}_i^t\mathbf{X}_{-i} & \mathbf{X}_i^t\mathbf{X}_i \end{pmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{X}_{-i}^t\mathbf{Y} \\ \mathbf{X}_i^t\mathbf{Y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^t & C \end{pmatrix} \cdot \begin{pmatrix} \mathbf{X}_{-i}^t\mathbf{Y} \\ \mathbf{X}_i^t\mathbf{Y} \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^t\mathbf{Y} - \widehat{\boldsymbol{\alpha}} \cdot \frac{\mathbf{e}_i^t\mathbf{Y}}{\mathbf{e}_i^t\mathbf{e}_i} \\ \frac{\mathbf{e}_i^t\mathbf{Y}}{\mathbf{e}_i^t\mathbf{e}_i} \end{pmatrix} \\ &= \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{-i} \\ \widehat{\beta}_i \end{pmatrix}, \end{aligned} \tag{4.2}$$

taking into account that:

$$\begin{aligned} C &= \left(\mathbf{X}_i^t\mathbf{X}_i - \mathbf{X}_i^t\mathbf{X}_{-i}\left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^t\mathbf{X}_i\right)^{-1} \\ &= \left(\mathbf{X}_i^t\left(\mathbf{I} - \mathbf{X}_{-i}\left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^t\right)\mathbf{X}_i\right) = \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1}, \\ \mathbf{B} &= -\left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^t\mathbf{X}_i \cdot \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1} = -\widehat{\boldsymbol{\alpha}} \cdot \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1}, \\ \mathbf{A} &= \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} + \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^t\mathbf{X}_i \cdot \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1}\mathbf{X}_i^t\mathbf{X}_{-i}\left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} \\ &= \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} + \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1} \cdot \widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t, \end{aligned}$$

where $\widehat{\boldsymbol{\alpha}}$ and $\mathbf{e}_i^t\mathbf{e}_i$ are, respectively, the OLS estimator and the sum of square residuals from the auxiliary regression (2.4).

Likewise, since $\mathbf{e}_i^t\mathbf{X}_{-i} = \mathbf{0}$, the OLS estimator of model (4.1), $\widehat{\boldsymbol{\gamma}}$, will be:

$$\widehat{\boldsymbol{\gamma}} = \left(\mathbf{X}_O^t\mathbf{X}_O\right)^{-1}\mathbf{X}_O^t\mathbf{Y} = \begin{pmatrix} \mathbf{X}_{-i}^t\mathbf{X}_{-i} & \mathbf{X}_{-i}^t\mathbf{e}_i \\ \mathbf{e}_i^t\mathbf{X}_{-i} & \mathbf{e}_i^t\mathbf{e}_i \end{pmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{X}_{-i}^t\mathbf{Y} \\ \mathbf{e}_i^t\mathbf{Y} \end{pmatrix}$$
$$= \begin{pmatrix} \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^t\mathbf{Y} \\ \frac{\mathbf{e}_i^t\mathbf{Y}}{\mathbf{e}_i^t\mathbf{e}_i} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_{-i} \\ \widehat{\gamma}_i \end{pmatrix}. \tag{4.3}$$

Therefore, it is possible to compare the OLS estimator of the residualized model (4.1), expression (4.3), with the OLS estimator of model (1.1), expression (4.2). The following conclusions are obtained:

- The estimation of the coefficient of the residualized variable does not change in model (4.1), i.e. $\widehat{\beta}_i = \widehat{\gamma}_i$. However, the interpretation of both estimates is different: $\widehat{\gamma}_i$ represents the variation produced in the dependent variable $\mathbf{Y}$, given an increase in $\mathbf{e}_i$, i.e. the part of the independent variable $\mathbf{X}_i$ unrelated with the rest of the independent variables $\mathbf{X}_{-i}$. Hence, due to the new interpretation of the residualized variable, the procedure can be applied to obtain conclusions that otherwise may not be possible.

- The orthogonality between $\mathbf{e}_i$ and $\mathbf{X}_{-i}$ verifies the principle *ceteris paribus*, i.e. when $\mathbf{e}_i$ increases, the other variables remain unchanged.

- The estimation of the non-residualized variables in model (4.1) changes:

$$\widehat{\boldsymbol{\beta}}_{-i} = \widehat{\boldsymbol{\gamma}}_{-i} - \widehat{\boldsymbol{\alpha}} \cdot \frac{\mathbf{e}_i^t\mathbf{Y}}{\mathbf{e}_i^t\mathbf{e}_i}. \tag{4.4}$$

  However, the interpretation is the same as that in model (1.1).

In addition, it is interesting to take into consideration that:

- For convenience purposes, all of the independent variables of the model (1.1) are included in the auxiliary regression (2.5). However, it is possible to include only some of the independent variables, depending on the interest of the researcher (for example, trying to obtain interpretable new variables). In this case, the estimations of the explanatory variables which are not included in the auxiliary regression do not change their value. The constant is included in the auxiliary regression, hence non-essential collinearity is mitigated because the residuals are orthogonal to the constant (see Section 3.2 for more details).

- The estimation of the non-residualized variables of model (4.1) coincides with the estimation obtained from model $\mathbf{Y} = \mathbf{X}_{-i}\boldsymbol{\beta} + \mathbf{u}$, i.e. the estimation and interpretation of the non-residualized variables is the same as that obtained in a regression in which the residualized variable is eliminated. Nevertheless, this coincidence only occurs when all the rest of the explanatory variables of the original model are included in the auxiliary regression. Furthermore, since the two models have different residuals, the inference associated with these coefficients will be different.

Remark 1. An interesting issue is how to select the variable to be residualized. The chapter presents different criteria that can be applied, or a combination thereof, depending on the goal of the research. If the goal is to look for new interpretations, the variable to be residualized will be the one that leads to the new interpretation desired by the researcher since the only interpretation that changes is that of the residualized variable.

Remark 2. It may be also interesting to rank the independent variables of the model (1.1) according to their relevance to avoid residualizing variables considered to be relevant maintaining the original interpretation of these

coefficients. This fact was already proposed in the work by Baird and Bieber (2016), which uses OVR models.

4.1.2 Goodness of fit, estimation of the variance of the random disturbance and joint significance

The estimated residuals of the original model (1.1) will be:

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X} \cdot \widehat{\boldsymbol{\beta}} = \mathbf{Y} - (\mathbf{X}_{-i}\ \mathbf{X}_i) \cdot \begin{pmatrix} \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \mathbf{X}_{-i}^t \mathbf{Y} - \widehat{\boldsymbol{\alpha}} \cdot \frac{\mathbf{e}_i^t \mathbf{Y}}{\mathbf{e}_i^t \mathbf{e}_i} \\ \frac{\mathbf{e}_i^t \mathbf{Y}}{\mathbf{e}_i^t \mathbf{e}_i} \end{pmatrix}$$

$$= \mathbf{Y} - \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \mathbf{X}_{-i}^t \mathbf{Y} + \mathbf{X}_{-i} \widehat{\boldsymbol{\alpha}} \cdot \frac{\mathbf{e}_i^t \mathbf{Y}}{\mathbf{e}_i^t \mathbf{e}_i} - \mathbf{X}_i \frac{\mathbf{e}_i^t \mathbf{Y}}{\mathbf{e}_i^t \mathbf{e}_i}$$

$$= \mathbf{Y} - \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \mathbf{X}_{-i}^t \mathbf{Y} - \mathbf{e}_i \frac{\mathbf{e}_i^t \mathbf{Y}}{\mathbf{e}_i^t \mathbf{e}_i}, \tag{4.5}$$

since $\mathbf{e}_i$ are the residuals of the auxiliary regression (2.5), it is verified that $\mathbf{e}_i = \mathbf{X}_i - \mathbf{X}_{-i}\widehat{\boldsymbol{\alpha}}$.

The estimated residuals of the residualized model (4.1) will be:

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}_O = \mathbf{Y} - \mathbf{X}_O \cdot \widehat{\boldsymbol{\gamma}}$$

$$= \mathbf{Y} - \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \mathbf{X}_{-i}^t \mathbf{Y} - \mathbf{e}_i \frac{\mathbf{e}_i^t \mathbf{Y}}{\mathbf{e}_i^t \mathbf{e}_i}. \tag{4.6}$$

It is evident that expression (4.6) coincides with expression (4.5), i.e. the estimated residuals of the original model (1.1) and the residualized model (4.1) are the same. Therefore, it is possible to conclude the following:

- The sum of square residuals from both models coincides and consequently, both models yield the same estimation of the variance of the random disturbance.
- Since the two models employ the same dependent variable, the total sum of squares will be the same and consequently, the coefficient of determination from both models will also coincide.

- Since the F statistic of the global significance test can be expressed as a function of the coefficient of determination, it is evident that the global significance test from both models will also be the same.
- It is clear that $\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}_O$, i.e. the original model and the residualized model provide the same prediction.

4.1.3 Individual inference

Since the random disturbances are spherical, the individual inference will be given by the main diagonal of matrix $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$, i.e. by (see expression (4.2)):

$$\begin{pmatrix} \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} + \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1} \cdot \widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t & -\widehat{\boldsymbol{\alpha}} \cdot \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1} \\ -\widehat{\boldsymbol{\alpha}}^t \cdot \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1} & \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1} \end{pmatrix}. \tag{4.7}$$

Taking into account the following expression:

$$\left(\mathbf{X}_O^t\mathbf{X}_O\right)^{-1} = \begin{pmatrix} \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{e}_i^t\mathbf{e}_i\right)^{-1} \end{pmatrix}, \tag{4.8}$$

it is evident that the main diagonal of both matrices is different, except for the i element. Since the estimation of the variance of the random disturbance is the same, considering the estimation of the coefficients, it is possible to conclude the following:

- The inference related to the individual significance (Student's t-test) of the unchanged variables differs between models (1.1) and (4.1).
- The inference related to the individual significance (Student's t-test) of the residualized variable coincides in models (1.1) and (4.1).

Consequently, the residualization of the initial model does not affect the estimation of the variance of the random disturbance, the coefficient

of determination, the global significance test or the individual significance test of the residualized variable. It only changes the individual significance of unaltered variables.

Remark 3. Another option to select the variable to be residualized is to choose a variable with a coefficient that is significantly different from zero in the original model since the individual significance test of the residualized variable is maintained in the residualized model.

4.2 Collinearity

In addition to the new interpretation of the coefficient of the residualized variable, another result of interest in the residualized model is the effect on the linear relationship between the independent variables of the initial model. To verify that collinearity is mitigated after the residualization of the initial model, the estimated variances of the estimated coefficients, the VIF and the CN are analysed in the residualized model.

4.2.1 Decrease in estimated variance

Considering that the estimation of the variance of the random disturbance of the original model is the same as that of the residualized model, the estimated variances of the coefficients will be determined by the main diagonal of the matrices $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ and $\left(\mathbf{X}_O^t\mathbf{X}_O\right)^{-1}$, respectively. As noted above, the element corresponding to the residualized variable is the same in both matrices, and thus, the estimated variance will be also the same, i.e. $\widehat{Var}\left(\widehat{\beta}_i\right) = \widehat{Var}\left(\widehat{\gamma}_i\right)$.

For the rest of the variables, given expressions (4.7) and (4.8), it is possible

to obtain that:

$$\widehat{Var}\left(\widehat{\beta}_j\right) = \widehat{\sigma}^2 \cdot \left(w_{jj} + (\mathbf{e}_i^t \mathbf{e}_i)^{-1} \alpha_{jj}\right), \quad \widehat{Var}\left(\widehat{\gamma}_j\right) = \widehat{\sigma}^2 \cdot w_{jj}, \quad j = 1, \ldots, p,\ j \neq i,$$

where w_{jj} and $\alpha_{jj} = \alpha_j^2$ are the elements (j, j) of the matrices $\left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1}$ and $\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t$, respectively. Since $(\mathbf{e}_i^t \mathbf{e}_i)^{-1} \alpha_j^2 \geq 0$, it is verified that $\widehat{Var}\left(\widehat{\beta}_j\right) \geq \widehat{Var}\left(\widehat{\gamma}_j\right)$ for $j = 1, \ldots, p$, with $j \neq i$. In consequence, the estimated variances of the residualized model will be always lower than or equal to those in the original model. This result is relevant since it demonstrates that the residualization implies a decrease in the estimated variances of the estimated coefficients (which are assumed to be inflated due to the presence of collinearity). Note that this result is contrary to the conclusions presented by Buse (1994).

The linear relationship between the coefficients of the model (1.1) given in (4.4) can also be used to reduce the variance of the estimated coefficients only by estimating the model with restricted least-squares. In this case, the residualization could be used to mitigate this particular consequence of the existence of severe collinearity in the multiple linear regression model.

4.2.2 Variance Inflation Factor (VIF)

Each explanatory variable of model (1.1) has an associated VIF given by expression (2.6) for $i = 2, \ldots, p$.

As was said in Section 2.2.2, it is generally accepted that values of VIF higher than 10 indicate severe collinearity.

Being $\mathbf{e}_i$ the dependent variable of the auxiliary regression, its coefficient of determination will be zero and the associated VIF will be one (the minimum value possible). In other case, R_j^2 will be obtained from the following auxiliary regression:

$$\mathbf{X}_j = \mathbf{X}_{O_{-j}}\, \boldsymbol{\xi} + \boldsymbol{\epsilon}, \quad j = 2, \ldots, p,\ j \neq i, \tag{4.9}$$

where $\mathbf{X}_{O_{-j}}$ is the result obtained after eliminating column (variable) j from matrix $\mathbf{X}_O$.

Due to the orthogonality between $\mathbf{e}_i$ and $\mathbf{X}_{-i,-j}$ (matrix $\mathbf{X}$ after eliminating columns (variables) i and j from the same), the residuals of (4.9) coincide with the residuals of the following model[2]:

$$\mathbf{X}_j = \mathbf{X}_{-i,-j}\ \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad i, j = 2, \ldots, p,\ i \neq j. \tag{4.10}$$

Then, models (4.9) and (4.10) have the same coefficient of determination since the dependent variable is the same in both models.

However, the coefficient of determination from model (4.10) will be lower than that of the following model:

$$\mathbf{X}_j = \mathbf{X}_{-j}\boldsymbol{\theta} + \boldsymbol{\omega}, \quad j = 2, \ldots, p,\ j \neq i, \tag{4.11}$$

since this latter model contains an additional independent variable, $\mathbf{X}_i$. Then, the coefficient of determination from model (4.9) is lower than that of model (4.11).

Therefore, since the VIF associated with variable j in the original model (1.1) is obtained from the coefficient of determination of the auxiliary regression given by (4.11) and in the residualized model (4.1) is obtained from the coefficient of determination of model (4.9), it is clear that the VIF is decreased after residualizing the model, i.e. the existing collinearity of the model is diminished.

[2] With $\mathbf{e}_a$ being the residuals of the auxiliary regression (4.9) and $\mathbf{e}_b$ being the residuals of the regression (4.10) and given that $\mathbf{e}_i$ is orthogonal to $\mathbf{X}_j$ and $\mathbf{X}_{-i,-j}$, it is obtained that

$$\begin{aligned}\mathbf{e}_a &= \mathbf{X}_j - (\mathbf{X}_{-i,-j}\ \mathbf{e}_i)\,\widehat{\boldsymbol{\xi}} = \mathbf{X}_j - (\mathbf{X}_{-i,-j}\ \mathbf{e}_i)\cdot \begin{pmatrix} \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{X}_j - \mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_j = \mathbf{X}_j - \mathbf{X}_{-i,-j}\widehat{\boldsymbol{\eta}} = \mathbf{e}_b.\end{aligned}$$

Remark 4. If the goal is to mitigate the existing collinearity in the model, one suggestion may be to residualize the variable with the highest VIF because, after the residualization, the VIF will be equal to one. In this case, all independent variables should be included in the auxiliary regression (2.5) to mitigate the essential and non-essential collinearity in the most efficient way.

4.2.3 Condition Number (CN)

Starting from model (1.1), the CN is given by expression (2.7). Note that the matrix $\mathbf{X}$ should be transformed to be unit length by columns, i.e. data should be divided by the square root of the sum of its square elements (see Belsley (1991)).

The CN associated with the model (4.1) is obtained by using the minimum and maximum eigenvalue of $\mathbf{X}_O^t\mathbf{X}_O$, where:

$$\mathbf{X}_O = \left(\frac{\mathbf{X}_1}{||\mathbf{X}_1||}\cdots\frac{\mathbf{X}_{i-1}}{||\mathbf{X}_{i-1}||}\ \frac{\mathbf{X}_{i+1}}{||\mathbf{X}_{i+1}||}\cdots\frac{\mathbf{X}_p}{||\mathbf{X}_p||}\ \frac{\mathbf{e}_i}{||\mathbf{e}_i||}\right) = \left(\mathbf{X}_{-i}\ \frac{\mathbf{e}_i}{||\mathbf{e}_i||}\right),$$

being $||\mathbf{X}_k|| = \sqrt{\sum_{j=1}^{n} X_{kj}^2}$ for $k = 1,\ldots,i-1,i+1,\ldots,p$ and $||\mathbf{e}_i|| = \sqrt{\sum_{j=1}^{n} e_{ij}^2}$. Then:

$$\mathbf{X}_O^t\mathbf{X}_O = \begin{pmatrix} \mathbf{X}_{-i}^t\mathbf{X}_{-i} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}.$$

Hence, one of the p eigenvalues of $\mathbf{X}_O^t\mathbf{X}_O$ will be equal to one and the rest will coincide with the eigenvalues of matrix $\mathbf{X}_{-i}^t\mathbf{X}_{-i}$. Supposing that the eigenvalue equal to one is the first one, $\mu_{1,O} = 1$, it is verified that:

- If this is the minimum eigenvalue of $\mathbf{X}_O^t\mathbf{X}_O$, the rest of eigenvalues will be equal or higher than one ($\mu_{i,O} \geq 1, i = 2,\ldots,p$) and consequently, its sum will be equal or higher than $p-1$ $\left(\sum_{i=2}^{p}\mu_{i,O} \geq p-1\right)$. However, this sum will be equal to $p-1$ since the trace of $\mathbf{X}_{-i}^t\mathbf{X}_{-i}$ is equal to

$p-1$. Then, all the eigenvalues will be equal to one ($\mu_{i,O}=1$ with $i=1,2,\ldots,p$), i.e. $\mathbf{X}_O^t\mathbf{X}_O$ will be the identity matrix and then, all the variables will be considered orthogonal to each other.

- If this is the maximum eigenvalue of $\mathbf{X}_O^t\mathbf{X}_O$, the rest of eigenvalues will be equal or lesser than one ($\mu_{i,O}\leq 1, i=2,\ldots,p$), and, consequently, its sum will be equal or lesser than $p-1$ $\left(\sum_{i=2}^{p}\mu_{i,O}\leq p-1\right)$. However, this sum is equal to $p-1$. Then, all the eigenvalues will be equal to one ($\mu_i=1$ with $i=1,2,\ldots,p$) and all the variables will be considered orthogonal to each other.

If the eigenvalue equal to one cannot be the minimum or maximum eigenvalue of $\mathbf{X}_O^t\mathbf{X}_O$, they will have to be found on the rest of the eigenvalues of $\mathbf{X}_{-i}^t\mathbf{X}_{-i}$. Thus, the CN of model (4.1) coincides with that of the auxiliary regression (2.5):

$$\text{CN}(\mathbf{X}_O^t\mathbf{X}_O)=\text{CN}(\mathbf{X}_{-i}^t\mathbf{X}_{-i}).$$

On the other hand, according to the Cauchy's Interlace Theorem for Eigenvalues of Hermitian Matrices[3], since $\mathbf{X}_{-i}^t\mathbf{X}_{-i}$ is a submatrix of order $p-1$ of $\mathbf{X}^t\mathbf{X}$, it is verified that:

$$\text{CN}(\mathbf{X}_O^t\mathbf{X}_O)=\text{CN}(\mathbf{X}_{-i}^t\mathbf{X}_{-i})\leq\text{CN}(\mathbf{X}^t\mathbf{X}).$$

Thus, the CN of the residualized model (4.1) has to be lower than or equal to the CN of the original model (1.1).

Remark 5. If the goal is to mitigate the collinearity in the model, one suggestion could be to residualize the variable i whose auxiliary regression (where the

[3]Given a matrix $\mathbf{A}$ with order p and eigenvalues $\xi_1\leq\xi_2\leq\cdots\leq\xi_p$ and given its submatrix $\mathbf{B}$ with order $p-1$ and eigenvalues $\mu_1\leq\mu_2\leq\cdots\leq\mu_{p-1}$, it is verified that $\xi_1\leq\mu_1\leq\xi_2\leq\mu_2\leq\xi_3\leq\cdots\leq\xi_{p-1}\leq\mu_{p-1}\leq\xi_p$.

variable i is the dependent variable) presents the lowest CN since it coincides with the CN of the residualized model that will always be equal to or lower than the CN of the original model.

4.3 Comparison of the residualization method with other existing methods

This section presents a Monte Carlo simulation to compare the residualization methodology with other existing methods such as ridge regression, PCR and PLSR in relation to the mean square error (MSE) and prediction error. Firstly, the obtention of the MSE of the residualization method is presented, as well as the way to compare it with the MSE obtained by OLS. Secondly, the metrics used to measure the prediction capability of each methodology are also presented.

4.3.1 Mean Square Error (MSE)

Note that the original model is different from the residualized model, and for this reason, both models should be analysed separately and the comparison may not be convenient. However, some publications have not considered this divergence (see, for example, York (2012)). Bearing this in mind, given that $\widehat{\boldsymbol{\gamma}}$ is a biased estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{X}_O\mathbf{X}_O)^{-1} \cdot \mathbf{X}_O^t\mathbf{Y} = (\mathbf{X}_O\mathbf{X}_O)^{-1} \cdot \mathbf{X}_O^t\mathbf{X} \cdot \boldsymbol{\beta} + (\mathbf{X}_O\mathbf{X}_O)^{-1} \cdot \mathbf{X}_O^t \cdot \mathbf{u},$$
$$E[\widehat{\boldsymbol{\gamma}}] = (\mathbf{X}_O\mathbf{X}_O)^{-1} \cdot \mathbf{X}_O^t\mathbf{X} \cdot \boldsymbol{\beta} \neq \boldsymbol{\beta} \text{ since } \mathbf{X}_O^t \neq \mathbf{X},$$

it could be interesting to calculate the MSE of the residualization and to compare it with the MSE of the OLS estimator.

Given an estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, its MSE is expressed as:

$$\text{MSE}\left(\widetilde{\boldsymbol{\beta}}\right) = trace\left(var\left(\widetilde{\boldsymbol{\beta}}\right)\right) + \left(E\left[\widetilde{\boldsymbol{\beta}}\right] - \boldsymbol{\beta}\right)^t \left(E\left[\widetilde{\boldsymbol{\beta}}\right] - \boldsymbol{\beta}\right).$$

In the case of the OLS estimator, $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator ($E\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$) and, taking into account expression (4.7), the following is verified:

$$\begin{aligned}\text{MSE}\left(\widehat{\boldsymbol{\beta}}\right) &= trace\left(var\left(\widehat{\boldsymbol{\beta}}\right)\right) \\ &= \sigma^2 \cdot \left[trace\left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} + (\mathbf{e}_i^t\mathbf{e}_i)^{-1} \cdot trace\left(\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t\right) + (\mathbf{e}_i^t\mathbf{e}_i)^{-1}\right]. \quad (4.12)\end{aligned}$$

For the estimator $\widehat{\boldsymbol{\gamma}}$, taking into account expression (4.8), it is verified that[4]:

$$\begin{aligned}\text{MSE}\left(\widehat{\boldsymbol{\gamma}}\right) &= trace\left(var\left(\widehat{\boldsymbol{\gamma}}\right)\right) + \left(E\left[\widehat{\boldsymbol{\gamma}}\right] - \boldsymbol{\beta}\right)^t \left(E\left[\widehat{\boldsymbol{\gamma}}\right] - \boldsymbol{\beta}\right) \\ &= \sigma^2 \cdot \left[trace\left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} + (\mathbf{e}_i^t\mathbf{e}_i)^{-1}\right] + \beta_i \cdot \widehat{\boldsymbol{\alpha}}^t\widehat{\boldsymbol{\alpha}} \cdot \beta_i. \quad (4.13)\end{aligned}$$

From expressions (4.12) and (4.13), it is clear that:

$$\text{MSE}\left(\widehat{\boldsymbol{\gamma}}\right) = \text{MSE}\left(\widehat{\boldsymbol{\beta}}\right) - \sigma^2 \cdot (\mathbf{e}_i^t\mathbf{e}_i)^{-1} \cdot trace\left(\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t\right) + \beta_i^2 \cdot \widehat{\boldsymbol{\alpha}}^t\widehat{\boldsymbol{\alpha}},$$

so $\widehat{\boldsymbol{\gamma}}$ has a lower MSE than $\widehat{\boldsymbol{\beta}}$ if:

$$\beta_i^2 \cdot \widehat{\boldsymbol{\alpha}}^t\widehat{\boldsymbol{\alpha}} < \sigma^2 \cdot (\mathbf{e}_i^t\mathbf{e}_i)^{-1} \cdot trace\left(\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t\right). \quad (4.14)$$

[4]Based on expressions (4.2), (4.3) and (4.4), it is obtained that $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\beta}} + \mathbf{s}$, where: $\mathbf{s} = \begin{pmatrix} \widehat{\boldsymbol{\alpha}} \cdot \frac{\mathbf{e}_i^t\mathbf{Y}}{\mathbf{e}_i^t\mathbf{e}_i} \\ 0 \end{pmatrix}$. Thus, as:

$$\mathbf{e}_i^t\mathbf{Y} = \mathbf{e}_i^t\mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i^t\mathbf{u} = [\mathbf{0}\ \mathbf{e}_i^t\mathbf{X}_i] \cdot \boldsymbol{\beta} + \mathbf{e}_i^t\mathbf{u} = \mathbf{e}_i^t\mathbf{X}_i\beta_i + \mathbf{e}_i^t\mathbf{u} = \mathbf{e}_i^t\mathbf{e}_i\beta_i + \mathbf{e}_i^t\mathbf{u},$$

it is obtained that:

$$E\left[\widehat{\boldsymbol{\gamma}}\right] = E\left[\widehat{\boldsymbol{\beta}}\right] + E\left[\mathbf{s}\right] = \boldsymbol{\beta} + \begin{pmatrix} \widehat{\boldsymbol{\alpha}} \cdot \beta_i \\ 0 \end{pmatrix} \Rightarrow \left(E\left[\widehat{\boldsymbol{\gamma}}\right] - \boldsymbol{\beta}\right)^t \left(E\left[\widehat{\boldsymbol{\gamma}}\right] - \boldsymbol{\beta}\right) = \beta_i^2 \cdot \widehat{\boldsymbol{\alpha}}^t\widehat{\boldsymbol{\alpha}}.$$

4.3.2 Metrics

The root mean square error (RMSE) and the mean absolute error (MAE) will be applied to measure the fit capability of each model while the prediction capability will be measured by the root mean square prediction error (RMSPE) and the mean absolute prediction error (MAPE).

Given a sample with n observations and assuming it is divided into two subsamples: the first with m observations and the second with h observations, verifying that $m + h = n$. Then, the first subsample is applied to measure the fit capability calculating the RMSE and MAE. The second subsample is applied to evaluate the prediction capability obtaining the RMSPE and MAPE. The following expressions are obtained:

$$\text{RMSE} = \sqrt{\frac{1}{m} \cdot \sum_{i=1}^{m} (Y_i - \widehat{Y}_i)^2}, \qquad \text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |Y_i - \widehat{Y}_i|,$$

$$\text{RMSPE} = \sqrt{\frac{1}{h} \cdot \sum_{i=m+1}^{n} (Y_i - \widehat{Y}_i)^2}, \qquad \text{MAPE} = \frac{1}{h} \cdot \sum_{i=m+1}^{n} |Y_i - \widehat{Y}_i|.$$

4.3.3 Simulation

The simulation performed to compare residualization with other existing methods is described below.

Given the model (2.1), $\mathbf{Y} = \beta_1 + \beta_2\, \mathbf{X}_2 + \beta_3\, \mathbf{X}_3 + \mathbf{u}$, the following simulation is performed in order to establish the behavior of condition (4.14):

1. It is considered that $\boldsymbol{\mu}_{2\times 1} = (\mu_1, \mu_2)^t$ with $\mu_1, \mu_2 \in \{-10, -9, -8, \ldots, 10\}$.

2. Additionally, it is also considered that $a_1, a_2 \in \{0, 1, 2, 3, 4\}$ and $b_1, b_2 \in \{0.1, 0.2, 0.3, \ldots, 2\}$, so $c^i_{5\times 1} \sim N(a_i, b_i^2)$ is generated. Thus, given matrix

$\mathbf{C} = [c^1\ c^2]$, a symmetric positive-definite matrix, $\Sigma_{2\times 2} = \mathbf{C}^t\mathbf{C}$, is built.

3. $\mathbf{X}_2$ and $\mathbf{X}_3$ are generated from $N_2\left(\boldsymbol{\mu}_{2\times 1}, \Sigma_{2\times 2}\right)$.

4. The random perturbance, $\mathbf{u}$, is generated as $\mathbf{u} \sim N(0, d^2)$, where $d \in \{1, 2, 3, 4\}$, from which is calculated $\mathbf{Y} = \beta_1 + \beta_2\ \mathbf{X}_2 + \beta_3\ \mathbf{X}_3 + \mathbf{u}$, where $\beta_i \in \{-5, -4, -3, \ldots, 5\}$.

5. A comparison of both models (OLS and residualization) is conducted with different sample sizes, $n \in \{25, 50, 75, 100, 125, 150\}$, such that 60000 simulations are performed in this experiment.

First, once the previous model and the corresponding auxiliary regressions are estimated, condition (4.14) is calculated from the obtained estimations of β_i, σ^2 and $\boldsymbol{\alpha}$. In Table 4.1, two types of situations can be observed: one where essential collinearity does not imply strong collinearity problems (the mean correlation is equal to 0.4877, which leads to a VIF value of 1.31208) and another where essential collinearity implies strong collinearity problems (the maximum and minimum correlations lead to VIF values of approximately 50.2512). It can also be observed that there are two types of situations in relation to non-essential collinearity: one where it is not worrisome (the mean value of the coefficient of variation (CV) is approximately 6, which implies the data have enough variability) and another where non-essential collinearity is worrisome (the minimum values of CV for each variable are close to zero, which implies slight variability of the data and indicates that the data may be considered almost constant and hence related to the intercept).

The first and second rows of Table 4.1 show the percentage of cases in which $\text{MSE}\left(\widehat{\boldsymbol{\gamma}}\right) < \text{MSE}\left(\widehat{\boldsymbol{\beta}}\right)$ (condition (4.14) is verified), considering that variables $\mathbf{X}_2$ and $\mathbf{X}_3$ are residualized, respectively. Note that both results are similar

Table 4.1: Simulation results for MSE.

	n	25	50	75	100	125	150	Mean
Condition (4.14)	Resid. variable $\mathbf{X}_2$	8.13%	6.96%	7.15%	6.69%	6.96%	7.00%	7.159%
	Resid. variable $\mathbf{X}_3$	8.42%	7.15%	7.12%	6.71%	6.77%	6.84%	
min $cor(\mathbf{X}_2, \mathbf{X}_3)$		-0.9862	-0.9747	-0.9883	-0.9973	-0.9934	-0.9915	0.4877
max $cor(\mathbf{X}_2, \mathbf{X}_3)$		0.9999	0.9999	0.9999	0.9999	0.9998	0.9999	
min CV($\mathbf{X}_2$)		0.00958	0.00935	0.00911	0.0111296	0.00628	0.01072	6.6665
max CV($\mathbf{X}_2$)		30222.48	12359.03	3249.84	3426.69	31001.38	1498.77	
min CV($\mathbf{X}_3$)		0.0107	0.0116	0.00791	0.00682	0.0108	0.0114	5.7096
max CV($\mathbf{X}_3$)		9763.29	37893.3	5199.45	2718.18	2655.24	3586.79	

and there are no material differences for different sample sizes. The results show that in only 7.159% of the cases the condition $\text{MSE}(\widehat{\boldsymbol{\gamma}}) < \text{MSE}\left(\widehat{\boldsymbol{\beta}}\right)$ is verified.

Second, Table 4.2 is obtained from 60000 more simulations performed dividing the sample as was described in Subsection 4.3.2 and considering $h = 0.15 \cdot n$. `R`'s package `pls` (Mevik et al. (2019)) was applied to obtain values of PCR and PLSR considering one and two components. For ridge regression, the value of k was selected in order to mitigate the collinearity considering that it is not worrying for values of CN lower than 20, as showed in the work by Salmerón et al. (2018). This idea was also applied in García et al. (2019b) by using the VIF instead of the CN. In this section it was considered more appropriate to use the CN since the VIF ignores the non-essential collinearity, Salmerón et al. (2018).

From the results of the first sample, it is obtained that residualization and OLS lead to the same values of RMSE and MAE due to $\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}_O$ being verified. These values are slightly lower than those of other techniques.

From values of RMSPE and MAPE obtained from the second sample, it is possible to conclude that the residualization method presents the lowest prediction capability. However, the fact that the rest of methods do not improve the results obtained by OLS could indicate that, when the purpose is prediction, the best way to proceed is to do nothing. These results support the idea provided by Gujarati (2004): *if the goal is simply to predict [...], then multicollinearity is not a problem [because] the predictions will still be accurate.*

Table 4.2: Simulation results for RMSE, MAE, RMSPE and MAPE.

Metric	OLS	Resid. var. $\mathbf{X}_2$	Resid. var. $\mathbf{X}_3$	RR	PCR (1c)	PLSR (1c)	PCR (2c)	PLSR (2c)
RMSE	2.635103	2.635103	2.635103	2.637032	8.289639	5.435111	8.810389	6.182757
MAE	1.960508	1.960508	1.960508	1.961239	6.611481	4.300871	8.507369	6.266333
RMSPE	2.725054	19.73536	19.64029	2.721985	8.49428	5.591137	2.725054	2.725054
MAPE	2.087758	17.56701	17.48688	2.084665	6.92049	4.524651	2.087758	2.087758

4.4 Successive residualization

It is possible that the goal of the researcher (to mitigate collinearity or obtain a new interpretation for the estimated coefficients) has not been achieved after residualizing the first variable. In that case, it is necessary to residualize a second variable.

In Chapter 3 (Section 3.4), the successive residualization for three standardised explanatory variables was presented together with general properties of this procedure. The goal of this section is not to obtain the estimated inference of this model (4.15) but to generalize the successive residualization.

For p independent variables, the double residualized model will be:

$$\mathbf{Y} = \mathbf{X}_{OO}\boldsymbol{\delta} + \boldsymbol{\varsigma}, \tag{4.15}$$

where $\mathbf{X}_{OO} = (\mathbf{X}_{-i,-j}\ \mathbf{e}_i\ \mathbf{e}_j)$ with $\mathbf{e}_j$ being the residuals of the auxiliary regression (4.10).

The residuals $\mathbf{e}_i$ and $\mathbf{e}_j$ will be orthogonal since it is verified that:

$$\mathbf{e}_i^t\mathbf{e}_j = \mathbf{e}_i^t\left(\mathbf{X}_j - \mathbf{X}_{-i,-j}\widehat{\boldsymbol{\eta}}\right) = \mathbf{e}_i^t\mathbf{X}_j - \mathbf{e}_i^t\mathbf{X}_{-i,-j}\widehat{\boldsymbol{\eta}} = \mathbf{0}.$$

The previous relationship between the residuals will still hold if more variables are residualized, i.e. the degree of multicollinearity will continue decreasing. Note that if the process is repeated $p-1$ times, all the explanatory variables of the initial model will be orthogonal to each other.

4.5 Overview of the methodology

To sum up the results of this chapter (and the one before it), it has been demonstrated that with residualization it is possible not only to alleviate

multicollinearity problems in the econometric model but also to obtain different interpretations of the modified variables.

To sum up, the characteristics of the methodology are the following:

- Estimations and inference:
  - The coefficient of the residualized variable does not change, but the interpretation of the variable does: it will represent the part of the original variable that has no relationship with the rest of explanatory variables of the model (the principle *ceteris paribus* is strictly fulfilled) if the rest of the independent variables are included in the auxiliary regression.
  - The inference related to the individual significance of the residualized variable is still the same.
  - The coefficients of the non-residualized variables change, however the interpretations are still the same.
  - The inference related to the individual significance of the non-residualized variables is different.
  - The value of the estimated parameters of the non-residualized variables are the same as in the model which does not include the modified variable.
  - If only some of the independent variables are included in the auxiliary regression, then the estimations of the parameters of the explanatory variables not included in it also remain unchanged.
- Global properties:
  - The sum of square residuals of the original model and the residualized model are the same.

  - The estimate of the variance of the random disturbance does not change.
  - The coefficient of determination, R^2, is still the same.
  - The global significance test remains unchanged.
  - The original model and the residualized model provide the same prediction.

- With regard to other methodologies:
  - A Monte Carlo simulation was performed to conclude that ridge regression, principal component regression (PCR) or partial least squares regression (PLSR) present a prediction capability better than residualization but not better than ordinary least squares (OLS). Note that the original model estimated by OLS is different from the residualized model, and for this reason, both models should be analysed separately and the comparison may not be convenient. The fact that the rest of methods do not improve the results obtained by OLS could indicate that, when the purpose is prediction, the best way to proceed is to do nothing.

Chapter 5

Empirical part: environmental applications

The debate between Ehrlich-Holdren and Commoner (Ehrlich and Holdren (1970, 1971, 1972); Commoner et al. (1971)) regarding the factors that influence environmental damage, resulted in the IPAT identity, which states that the environmental impacts of a country (**I**) can be decomposed into the product of three principal factors: population (**P**), affluence (**A**) and technology (**T**), York et al. (2003). The main limitations of the IPAT identity are that the number of factors is limited and the impact of the factors is proportional as all of them affect the environment equally. Moreover, it does not allow hypothesis testing. Due to these problems, the STIRPAT model emerged as the stochastic version of the IPAT identity (STochastic Impacts by Regression on Population, Affluence and Technology) to analyse the influence of these three factors on the environment of a region, Dietz and Rosa (1994, 1997). More complex models have used additional variables such as behaviour (Schulze (2002); Kilbourne and Thyroff (2020)) or alternative specifications separating the technology factor

into parts: energy consumption and technological improvement (Waggoner and Ausubel (2002)) or industry value added and CO_2 intensity (Martínez-Zarzoso et al. (2007); Martínez-Zarzoso and Maruotti (2011)), for example.

Regardless of the selected variables, it is possible to consider that the factors affecting CO_2 emissions may be strongly correlated. Fan et al. (2006) conclude that **P**, **A** and **T** clearly influence environmental damage, although the impact can vary at different levels of development of a country. This idea supports the dependence between the explanatory variables of the STIRPAT methodology (i.e. technical enhancement depends on the structure of the country, the economy, the population and so on). Indeed, Ehrlich and Holdren were aware of the problem of the relationship between variables in the IPAT identity but, as noted by Chertow (2000), expanded their first equation ignoring the interdependence of the variables: they only note the relationship between explanatory variables, but they do not treat the potential problem of multicollinearity.

Despite the likely presence of near collinearity and its consequences, most STIRPAT applications have disregarded this possibility and its analysis.

Henceforth, in order to simplify the reading of the chapter, distinctions between different types of collinearity will only be made if necessary. Additionally, the reader has to take into account that there are always relationships among real variables, thus to also simplify, if the reader comes across expressions such as "there is no collinearity" they refer to strong or troubling collinearity problems.

Tables 5.1 to 5.3 present a compilation of some empirical studies from 1997 to 2020 summarising the data, the variables used and the treatment of collinearity in each paper. Table 5.4 synthesises the different treatments given to collinearity in STIRPAT applications in Tables 5.1 to 5.3. Note that

Table 5.1: Compilation of some empirical studies in chronological order.

Article	Years of study	Countries of study	Dependent variable	Pollutant factors	Collinearity
Gürer and Ban (1997)	1970-1996	173 countries (world)	CO_2 emissions	Population, Affluence, Technology	Not applicable (IPAT)
Roberts and Grimes (1997)	1962-1991	144 countries (world)	CO_2 emissions	Affluence	Non-studied
De Bruyn et al. (1998)	1960-1993	4 countries (world)	CO_2 emissions and other GHG	GDP, Technology, Other social and/or economic variable(s)	Non-studied
Torras and Boyce (1998)	1977-1991	42 countries (world)	Other GHG and water pollutants	Affluence, Urban, Other social and/or economic variable(s)	Non-studied
Bengochea-Morancho et al. (2001)	1981-1995	10 countries (EU)	CO_2 emissions	GDP	Not applicable (only 1 explanatory variable)
Coondoo and Dinda (2002)	1960-1990	88 countries (world)	CO_2 emissions and GDP	GDP, Emissions	Non-studied
Hamilton and Turton (2002)	1982-1997	Whole OCDE countries	CO_2 emissions	Population, Affluence, Technology	Not applicable (IPAT)
Harbaugh et al. (2002)	1971-1992	45 countries (world)	Other GHG	GDP, Other social and/or economic variable(s)	Detected but not solved
Bruvoll and Medin (2003)	1980-1996	Norway	CO_2 emissions and other GHG	Population, Affluence, Technology	Not applicable (IPAT)
Roca and Padilla (2003)	1980-2001	Spain	CO_2 emissions and other GHG	Affluence, Technology	Non-studied
York et al. (2003)	1996 and 1999	146 countries (world)	CO_2 emissions and energy footprint	Population, Affluence, Technology, Urban, Other social and/or economic variable(s)	Non-problematic
Alcántara and Padilla (2005)	1971-2001	43 countries (world)	CO_2 emissions	Population, Affluence, Technology	Not applicable (IPAT)
Fan et al. (2006)	1975-2000	208 (world)	CO_2 emissions	Population, Affluence, Technology, Urban	Detected → PLS

Affluence represents per capita GDP. ***Population*** has variants in some studies (population density or % of some group of population by age or another characteristic). ***Technology*** is broken down into various factors in the major articles, and its measure is a controversial issue; it is usually a measure of energy consumption or similars. ***Urban*** represents the % of people living in urban areas. ***Other social and/or economic variable(s)*** includes factors like democracy, illiteracy, education, trade, transport, etc. ***Emissions*** can represent lagged emissions of CO_2 but also the emissions of another GHG (lagged or not).
Years and countries of each study represent the maximum interval of data for air pollution.

Table 5.2: Compilation of some empirical studies in chronological order (cont.).

Article	Years of study	Countries of study	Dependent variable	Pollutant factors	Collinearity
Kumar (2006)	1971-1992	41 developed and developing countries (world)	Productivity change	Affluence, Technology, Other social and/or economic variable(s)	Non-studied
Martínez-Zarzoso et al. (2007)	1975-1999	23 countries (EU)	CO_2 emissions	Population, Affluence, Technology	Non-problematic by the application of First differences
Jia et al. (2009)	1983-2006	China	Ecological footprint	Population, Affluence, Technology, Urban	Detected → PLS
Lin et al. (2009)	1978-2006	China	Other GHG	Population, Affluence, Technology, Urban	Detected → Ridge regression
Pao and Tsai (2010)	1971-2005	4 (BRIC countries)	CO_2 emissions	GDP, Technology	Non-studied
Gassebner et al. (2011)	1960-2001	120 countries (world)	CO_2 emissions, water pollution (biochemical oxygen demand)	Affluence, Other social and/or economic variable(s)	Non-studied
Martínez-Zarzoso and Maruotti (2011)	1975-2003	88 developing countries (world)	CO_2 emissions	Population, Affluence, Technology, Urban	Non-studied
Büchs and Schnepf (2013)	2006-2009	UK	CO_2 emissions	Other social and/or economic variable(s)	Detected but not solved
Fernández et al. (2015)	2005-2012	6 countries (EU)	CO_2 emissions	GDP, Other social and/or economic variable(s)	Non-studied
Azam and Khan (2016)	1975-2014	4 countries (world)	CO_2 emissions	GDP, Technology, Urban, Other social and/or economic variable(s)	Non-studied
Dong et al. (2016)	1989-2012	China	CO_2 emissions (transport)	Population, Affluence, Technology, Other social and/or economic variable(s)	Detected → Ridge regression
Khan et al. (2016)	2000-2013	9 developed countries (world)	Other GHG	Affluence, Technology, Emissions	Non-problematic
Rafindadi (2016)	1961-2012	Japan	CO_2 emissions	Affluence, Technology, Other social and/or economic variable(s)	Non-studied
Pablo-Romero and De Jesús (2016)	1990-2011	22 countries (Latin America and Caribbean)	Energy	Other social and/or economic variable(s)	Not applicable
Uddin et al. (2016)	1960-2014	Australia	Ecological footprint	Population, Affffluence, Technology, Urban, Emissions	Detected → Ridge regression

Table 5.3: Compilation of some empirical studies in chronological order (cont.).

Article	Years of study	Countries of study	Dependent variable	Pollutant factors	Collinearity
Shuai et al. (2018)	1995-2014	China	CO_2 emissions	Population, Affluence, Technology, Urban, Other social and/or economic variable(s)	Detected → Deleting variables
Chontanawat (2019)	1971-2013	9 countries (AESEAN countries, except Laos)	CO_2 emissions	Population, Affluence, Technology	Not applicable (IPAT)
Hashmi and Alam (2019)	1999-2014	29 OECD countries	CO_2 emissions	Population, Affluence, Other social and/or economic variable(s)	Non-studied
Li et al. (2019a)	1996-2016	China	CO_2 emissions	Population, Affluence, Technology, Urban	Detected → PCR
Li et al. (2019b)	2003-2014	China (regions)	CO_2 emissions	Population, Affluence, Technology	Non-studied
Liu et al. (2019)	2005-2015	China (one region)	CO_2 emissions	Population, Affluence, Technology, Urban	Detected → Ridge regression
Rasool et al. (2019)	1970-2014	Pakistan	CO_2 emissions (transport)	Population, Affluence, Technology	Non-studied
Wen and Shao (2019)	2001-2015	China	CO_2 emissions	Population, Affluence, Technology, Urban, Other social and/or economic variable(s)	Detected → Ridge regression
Xie and Liu (2019)	1997-2016	China (regions)	CO_2 emissions	Affluence, Technology, Urban	Detected → Transformed data
Xu et al. (2019)	2005-2015	China (regions)	Other GHG	Population, Affluence, Technology, Urban, Other social and/or economic variable(s)	Non-detected
Yang and Chen (2019)	2006-2015	China (regions)	SO_2 emissions	Population, Affluence, Technology, Urban, Other social and/or economic variable(s)	Non-studied
Zhang et al. (2019)	1971-2014	China	CO_2 emissions	Population, Affluence, Technology, Urban, Other social and/or economic variable(s)	Detected → Ridge regression
Zhang and Zhao (2019)	1996-2015	China (regions)	CO_2 emissions (energy)	Affluence, Technology, Urban, Other social and/or economic variable(s)	Detected but not solved
Kilbourne and Thyroff (2020)	2011	113 countries (world)	Ecological footprint	Population, Affluence, Technology, Other social and/or economic variable(s)	Non-studied

Table 5.4: Reviews of the different treatments given to collinearity in STIRPAT applications.

Treatment of collinearity	References
Disregarded	Azam and Khan (2016), De Bruyn et al. (1998), Coondoo and Dinda (2002), Fernández et al. (2015), Gassebner et al. (2011), Hashmi and Alam (2019), Kumar (2006), Li et al. (2019b), Martínez-Zarzoso and Maruotti (2011), Pablo-Romero and De Jesús (2016), Pao and Tsai (2010), Rafindadi (2016), Rasool et al. (2019), Roberts and Grimes (1997), Roca and Padilla (2003), Torras and Boyce (1998), Yang and Chen (2019), Kilbourne and Thyroff (2020)
Tested and not detected	Khan et al. (2016), Xu et al. (2019), York et al. (2003)
Detected and not treated	Büchs and Schnepf (2013), Harbaugh et al. (2002), Zhang and Zhao (2019)
Treated by deleting or transforming data	Martínez-Zarzoso et al. (2007), Shuai et al. (2018), Xie and Liu (2019)
Treated by Principal Component Regression (and variants)	Fan et al. (2006), Jia et al. (2009), Li et al. (2019a)
Treated by Ridge regression	Dong et al. (2016), Lin et al. (2009), Liu et al. (2019), Roy et al. (2017), Uddin et al. (2016), Wen and Shao (2019), Zhang et al. (2019)

collinearity is commonly neglected in the vast majority of the studies and efforts to address collinearity in STIRPAT models are usually limited to eliminating variables, the application of first differences or, more commonly, the application of partial least squares (PLS) or ridge regression as alternatives to ordinary least squares (OLS) estimation. Regarding the alternative methodologies to OLS, the scientific literature applies these methods to analyse the influence of environmental driving forces although Wei (2011) recommended ridge regression and PLS when the goal is prediction and the estimated parameters are not

interpreted as causal effects. On the other hand, the two methods differ and they are not comparable each other: while ridge regression is a biased method that tries to decrease the mean square error (MSE), PLS modifies the original variables of the data, transforming them into orthogonal components whose interpretation is questionable, as has been expressed in Subsection 2.3.3. In light of the foregoing, the traditional method used in STIRPAT applications that is going to be analysed in this chapter together with other techniques is ridge regression.

The main goal of biased methods is to decrease the mean square error of prediction by introducing a reasonable amount of bias into the model. Although ridge regression has been widely applied to estimate models with collinearity, it presents some disadvantages as explained in Section 2.3.1.

Throughout this chapter, the STIRPAT model is studied using different datasets: Section 5.1, whose starting point was the work by García et al. (2017a), analyses the model for 124 countries around the world, Section 5.2 uses data from the UE-28 countries focusing on four similar countries (see the corresponding section for more information) and Section 5.3 focuses on China, the most polluting country in the world. Regarding the methodology applied in each section: Section 5.1 compares residualization with three additional methodologies which lead the researcher to deal with worrying near collinearity problems: ridge regression, LASSO regression and raise regression; Section 5.2 uses residualization to mitigate collinearity problems, but this procedure is applied in three different ways as the reader will note; Section 5.3 applies residualization to show the reader the use of the method with empirical purposes; finally, Section 5.4 provides a summary of the chapter. The first example (Section 5.1) is based on the work by García et al. (2020)[1], the second

[1]The work by García et al. (2020) presents updated data.

(Section 5.2) is based on the research by Apergis and García (2019), and the third (Section 5.3) is one of the examples used in García et al. (2019c).

5.1 Model 1: the STIRPAT model in the world. Multicollinearity and residualization

In order to compare the results of the application of different methodologies, the following STIRPAT model is analysed:

$$\ln \mathbf{I} = \beta_1 + \beta_2 \ln \mathbf{P} + \beta_3 \ln \mathbf{A} + \beta_4 \ln \mathbf{T_1} + \beta_5 \ln \mathbf{T_2} + \mathbf{u}, \tag{5.1}$$

where $\mathbf{u}$ is the random disturbance, which is supposed to be spherical.

The dataset is obtained from the World Bank website[2]. It includes data on 124 countries ($n = 124$) for the year 2014. Information regarding the four variables is shown in Table 5.5. Note that the traditional component $\mathbf{T}$ has been separated into two factors according to Martínez-Zarzoso et al. (2007) and Martínez-Zarzoso and Maruotti (2011): $\mathbf{T_1}$, which measures the industry value added, and $\mathbf{T_2}$, which measures the CO_2 intensity.

Regarding the expected sign of the different variables (see Table 5.5), starting with population ($\mathbf{P}$), there is strong empirical evidence that it is a relevant factor in explaining the environmental impact of a country but there are different theories about its sign: followers of Malthus (1973) propose a positive sign due to the pressure that the population puts on resources, whereas followers of Boserup (1981) propose a negative relationship because population growth leads to technological innovation, diminishing the negative impact on the environment, Sherbinin et al. (2007) and Uddin et al. (2016). Note that this same idea can be extended to the interpretation of industry

[2] `https://databank.worldbank.org`

Table 5.5: Variables of STIRPAT model. The case of 124 world countries.

Notation	Variable	Unit	Expected sign
I	CO_2 emissions	kg per 2010 US$ of GDP	-
P	total of population	absolute value	Negative / Positive
A	GDP per capita	constant 2010 US$	Negative
$\mathbf{T_1}$	industry value added	constant 2010 US$	Negative / Positive
$\mathbf{T_2}$	CO_2 intensity	kg per kg of oil-equivalent energy use	Positive

value added ($\mathbf{T_1}$): the traditional interpretation would be that more industry would imply more pollution, but with the latest technology, this would not necessarily be the case. It is important to remark that a positive sign of $\mathbf{P}$ does not go hand-in-hand with a positive sign of $\mathbf{T}_1$. In the case of variable $\mathbf{A}$, if the researcher is studying a group of countries with different characteristics (which is the case of this example), the GDP will show the level of development or wealth of each one; the higher GDP, the greater possibilities of devoting resources to climate targets, and the expected sign for the parameter will be negative. Finally, regarding variable $\mathbf{T_2}$, the expected sign of its parameter will be positive because it is a variable that is directly related to emissions: higher CO_2 intensity implies more CO_2 emissions into the atmosphere.

By paying attention to the expected signs regarding the correlation matrix (5.2), in the case of $\mathbf{P}$ the idea of Malthus (1973) is supported, and in the case of $\mathbf{T}_1$ the perception of Boserup (1981) is sustained. In any case, according to the theory, both signs are acceptable.

$$\begin{pmatrix} & \ln \mathbf{I} & \ln \mathbf{P} & \ln \mathbf{A} & \ln \mathbf{T}_1 & \ln \mathbf{T}_2 \\ \ln \mathbf{I} & 1.000 & & & & \\ \ln \mathbf{P} & 0.065 & 1.000 & & & \\ \ln \mathbf{A} & -0.337 & -0.257 & 1.000 & & \\ \ln \mathbf{T}_1 & -0.128 & 0.659 & 0.530 & 1.000 & \\ \ln \mathbf{T}_2 & 0.466 & -0.079 & 0.441 & 0.289 & 1.000 \end{pmatrix}. \tag{5.2}$$

After this first introduction to the model, it has been validated. Results of heteroscedasticity and multicollinearity tests are explained below.

In relation to heteroscedasticity, the White test has been applied concluding in not rejecting the null hypothesis of homoscedasticity (p value higher than 0.05).

In the case of collinearity, by observing the correlation matrix (5.2), it is possible to see that the correlation coefficients are generally lower than 0.7. This fact indicates that there is not a strong correlation between pairs of variables. However, it is possible to note that population ($\ln \mathbf{P}$) and industry ($\ln \mathbf{T}_1$) are the most closely related variables. It would be logical to consider that the country's population affects the value added of the industry as larger population implies more production (in all sectors). In addition, it could be considered that affluence may be also correlated to industry value added and CO_2 intensity.

Although the correlation matrix gives the reader a first approximation about the existence of collinearity in the model and about the relationships between pairs of variables, a more in-depth analysis is required.

To test the presence of collinearity in the model, the variance inflator factor (VIF) is obtained (see Appendix B). From Table 5.7 presented in Subsection 5.1.2, it can be observed that VIFs for all variables (except $\ln \mathbf{T_2}$) are greater than 10. This implies that population, affluence and the industry value added are all related to each other. By obtaining the coefficients of variation (CV) of each variable of the model ($\text{CV}(\ln \mathbf{P}) = 0.097$, $\text{CV}(\ln \mathbf{A}) = 0.158$, $\text{CV}(\ln \mathbf{T}_1) = 0.079$, $\text{CV}(\ln \mathbf{T}_2) = 1.000$), it is clear that there are no strong non-essential collinearity problems.

As stated at the beginning of the chapter, this section treats the existing essential collinearity of the STIRPAT model by using three different methodologies together with residualization, and compares them to make conclusions about the residualization procedure. Therefore, before analysing the results, each methodology is briefly explained for this particular example.

5.1.1 Methodologies

In this section, residualization, raise regression, ridge regression and LASSO regression are adapted for their applicability to the STIRPAT model (5.1). After applying these methodologies, the mitigation of collinearity needs to be analysed. In the case of residualization, raise and ridge regression, this fact is verified, while it is not the case for LASSO regression. Regarding LASSO regression, it is not possible to corroborate directly whether the problem has been mitigated, but only to make assumptions about the fact (see Section 2.3.2), based on prior literature. Research in the future may develop tools to measure the existing collinearity when using LASSO regression. After that, the MSE is quantified for ridge, raise and residualization methods but not for LASSO regression. LASSO regression is excluded from the calculation of the MSE because this is done to compare these methods in terms other than mitigation of multicollinearity, and since multicollinearity may not be checked after the application of LASSO it is therefore excluded. Ridge and residualization procedures are already compared in terms of MSE in Chapter 4, but here the raise regression MSE is also included in the comparison and, in addition, the MSE is calculated by using real data.

Appendix B presents VIFs and MSEs of residualization, raise and ridge regression, together with traditional OLS. Results of the methods explained below are presented in Tables 5.6 and 5.7.

5.1.1.1 Residualization

Although the variable $\ln \mathbf{T_1}$ has the highest VIF, the variable that will be residualized is $\ln \mathbf{A}$ because the methodology provides an alternative interpretation of the residualized variable. Therefore, it is important to

highlight that the application of the method is only possible if the effect of the residualized variable can be interpreted. In this case, residualizing the factor affluence ($\ln \mathbf{A}$) makes sense from an interpretative point of view as its effect is analysed regardless of the population, the value added of the industry sector and CO_2 intensity. Indeed, the following auxiliary regression is specified:

$$\ln \mathbf{A} = \alpha_1 + \alpha_2 \ln \mathbf{P} + \alpha_3 \ln \mathbf{T_1} + \alpha_4 \ln \mathbf{T_2} + \mathbf{v}. \tag{5.3}$$

From this model, the obtained estimated residuals $\mathbf{e_A}$ will substitute variable $\ln \mathbf{A}$ in the original model (5.1) to obtain the following residualized model:

$$\ln \mathbf{I} = \gamma_1 + \gamma_2 \ln \mathbf{P} + \gamma_3 \mathbf{e_A} + \gamma_4 \ln \mathbf{T_1} + \gamma_5 \ln \mathbf{T_2} + \mathbf{w}. \tag{5.4}$$

Owing to the properties of OLS, $\mathbf{e_A}$ is orthogonal to the other independent variables. Hence, the principle *ceteris paribus* will be strictly fulfilled.

5.1.1.2 Raise regression

Considering model (5.1) as the starting point, and modifying variable $\ln \mathbf{A}$ to make comparisons with the previous method (residualization), $\ln \mathbf{A}$ is replaced by the new variable:

$$\widetilde{\mathbf{A}} = \ln \mathbf{A} + \lambda \mathbf{e_A}, \tag{5.5}$$

where $\mathbf{e_A}$ is the estimated residuals vector from model (5.3). The value of λ is selected to obtain VIFs lower than 10 ($\lambda = 0.719$). Note that for $\lambda = 0$, the OLS estimations of the initial model (5.1) are recovered.

In light of the foregoing, the raise regression for the STIRPAT model will be:

$$\ln \mathbf{I} = \gamma_1 + \gamma_2 \ln \mathbf{P} + \gamma_3 \widetilde{\mathbf{A}} + \gamma_4 \ln \mathbf{T_1} + \gamma_5 \ln \mathbf{T_2} + \mathbf{w}. \tag{5.6}$$

5.1.1.3 Ridge regression

Ridge regression estimates the parameters using the expression (2.8), $\widehat{\boldsymbol{\beta}}(k) = (\mathbf{X}^t\mathbf{X} + k\mathrm{I})^{-1}\mathbf{X}^t\mathbf{Y}$.

Note that when $k = 0$, the initial model (5.1) is estimated by the traditional OLS method. Usually, the criterion for choosing k when ridge regression is applied in the STIRPAT model is to consider a step length within the interval $(0, 1)$. For example, Dong et al. (2016) adopt a set size of 0.005, which yields a value of $k = 0.02$, and Lin et al. (2009) and Uddin et al. (2016) obtain $k = 0.05$ for step lengths of 0.01 and 0.05. In this case, the k value proposed by Hoerl et al. (1975) (see expression (2.9)), $k = 0.093$, is considered, and also the k value that makes the VIFs for all variables less than 10: $k = 0.029$.

The results of this method were obtained by using the library `lmridge` of `R` (Imdad and Aslam (2018)).

5.1.1.4 LASSO regression

In this particular case, the results of the LASSO regression are presented in Tables 5.6 and 5.7 and they were obtained by using the library `HDCI` of `R` (Liu et al. (2017)).

5.1.2 Comparison of the methods

Paying attention to Tables 5.6 and 5.7, regarding the obtained signs of the estimated parameters with each method, in all cases they are consistent with theory and expectations.

The estimation of model (5.1) by OLS indicates that affluence ($\ln \mathbf{A}$) and population ($\ln \mathbf{P}$) have a negative impact on CO_2 emissions per unit of GDP while industry value added ($\ln \mathbf{T_1}$) and CO_2 intensity ($\ln \mathbf{T_2}$) have a positive

Table 5.6: Results of the OLS estimation of the initial model, residualization, raise regression, ridge regressions and LASSO.

		OLS Model (5.1) ($\lambda = 0$; $k = 0$)	RESIDUALIZ. Model (5.4)	RAISE Model (5.6) ($\lambda = 0.719$)	RIDGE ($k = 0.029$)	RIDGE ($k = 0.093$)	LASSO ($\zeta = 0.008$)
Intercept	Estimator	2.957 ***	0.808	2.058 ***	2.949	2.931	2.957
	(s.d.)	(0.596)	(0.526)	(0.551)	n/a	n/a	n/a
P	Estimator	-0.521***	0.284 ***	-0.184 *	-0.518	-0.512	-0.521
	(s.d.)	(0.111)	(0.036)	(0.071)	n/a	n/a	n/a
A	Estimator	-0.853 ***			-0.850	-0.843	-0.853
	(s.d.)	(0.111)			n/a	n/a	n/a
$\mathbf{e_A}$	Estimator		-0.853 ***				
	(s.d.)		(0.111)				
$\widetilde{\mathbf{A}}$	Estimator			-0.496 ***			
	(s.d.)			(0.065)			
$\mathbf{T_1}$	Estimator	0.489 ***	-0.289 ***	0.164 *	0.487	0.480	0.489
	(s.d.)	(0.106)	(0.032)	(0.067)	n/a	n/a	n/a
$\mathbf{T_2}$	Estimator	0.875 ***	0.880 ***	0.877 ***	0. 874	0.873	0.875
	(s.d.)	(0.734)	(0.074)	(0.074)	n/a	n/a	n/a

***, * Statistically significant at 0.001 (99.9% confidence level) and at 0.05 (95% confidence level), respectively.

Table 5.7: Results of residualization, raise regression, ridge regression and LASSO estimator (VIF values and other characteristics).

VIFs	OLS Model (5.1) ($\lambda = 0$; $k = 0$)	RESIDUALIZ. Model (5.4)	RAISE Model (5.6) ($\lambda = 0.719$)	RIDGE ($k = 0.029$)	RIDGE ($k = 0.093$)	LASSO ($\zeta = 0.008$)
$\mathbf{P}$	19.317	2.057	7.898	7.719	3.715	n/a
$\mathbf{A}$	14.937			6.178	3.144	n/a
$\mathbf{e_A}$		1.000				
$\widetilde{\mathbf{A}}$			5.716			
$\mathbf{T}_1$	25.186	2.230	9.999	9.900	4.620	n/a
$\mathbf{T}_2$	1.269	1.269	1.269	1.244	1.212	n/a
Other characteristics						
MSE	0.397	2.484	0.751	0.385	0.499	n/a
R^2	0.648	0.648	0.648	n/a	n/a	n/a
F statistic	54.7 ***	54.7 ***	54.7 ***	n/a	n/a	n/a

*** Statistically significant at 0.001 (99.9% confidence level).

impact on CO_2 emissions per unit of GDP. Regarding the relevance of the factors in the OLS estimation (i.e. the ranking of the absolute values of the correspondent parameters), it can be concluded that the most important (the highest parameter in absolute value) is CO_2 intensity ($\ln \mathbf{T_2}$), followed by affluence ($\ln \mathbf{A}$), population ($\ln \mathbf{P}$) and industry value added ($\ln \mathbf{T_1}$). Another interesting result is that the effect of population ($\ln \mathbf{P}$) and affluence ($\ln \mathbf{A}$) is almost offset by the effect of industry value added ($\ln \mathbf{T_1}$) and CO_2 intensity ($\ln \mathbf{T_2}$), respectively: while a 1% increase in population will diminish emissions by 0.521%, a 1% increase in industry value added will increase emissions by 0.489%, then the effect is almost balanced, and the same conclusion could be reached for affluence ($\ln \mathbf{A}$) and CO_2 intensity ($\ln \mathbf{T}_2$).

Note that ridge, LASSO and OLS estimators present very similar values. Ridge regression mitigate collinearity (see Table 5.7) practically keeping the same values of the estimators obtained in the initial model that were considered to be unstable. However, with the results obtained it is not possible to reach any conclusions about the individual significance of the estimated parameters nor the global significance of the model in the case of ridge and LASSO regressions, hence the study of the influence and importance of each parameter on the dependent variable makes no sense. Furthermore, in the case of LASSO, it is not possible to verify whether multicollinearity problems have been mitigated.

On the other hand, the raise regression for $\lambda = 0.719$ provides estimators that differ from the OLS and ridge estimators. Although the signs of the parameters of each independent variable are the same as in the OLS estimation and the most important factor continues to be CO_2 intensity ($\ln \mathbf{T_2}$), followed by affluence ($\ln \mathbf{A}$), the impact of population ($\ln \mathbf{P}$) on emissions is much lower (-0.184), but it is also almost offset by the influence of industry value added

($\ln \mathbf{T_1}$) on emissions (0.164). However, the effect of affluence and CO_2 intensity is not offset in raise regression.

When residualization is applied, the most relevant variable continues to be CO_2 intensity ($\ln \mathbf{T_2}$), the second most relevant variable is the residualized variable $\mathbf{e_A}$, followed by industry value added ($\ln \mathbf{T_1}$) and population ($\ln \mathbf{P}$). Note that the signs of population ($\ln \mathbf{P}$) and industry ($\ln \mathbf{T_1}$) have also changed. Although both signs are acceptable regarding theory and expectations, residualization is the only methodology that obtains these signs. This methodology provides an alternative interpretation by using the estimated residuals $\mathbf{e_A}$ instead of the original variable $\ln \mathbf{A}$: the part of the affluence that is not related to the population of the country, the value added of the industrial sector and the CO_2 intensity is analysed. The new variable is fully uncorrelated with the economic structure and ecological efficiency, with $\mathbf{e_A}$ representing the real impact of the wealth of each country on the environment. Hence, following the idea of Boserup (1981) expressed previously, an interesting question for future research arises: when the explanatory variables of this model are not related to the wealth of the countries, does the level of technology play a key role in determining the influence of the rest of the variables?

Finally, as stated earlier, ridge regression has mechanically mitigated the collinearity and, in addition, it provides the smallest MSE, but the influence of factor k on the interpretation of the estimators is uncertain. Furthermore, according to Jensen and Ramírez (2008); Rodríguez et al. (2019), the t statistic of the individual significance tests, the R^2 and the F statistic of the global significance test are not included in Table 5.7 and, according to Lockhart et al. (2014), the same applies for LASSO.

Raise regression also mitigates collinearity: VIFs less than 10. The MSE obtained by the raise regression is higher than that obtained by ridge regression but in this case the estimated coefficients are geometrically interpretable (the same variable, separated from the others geometrically, is analysed) and the calculation of R^2 and the F statistic is possible (see references from Subsection 2.3.4).

As regards residualization methodology, it has clearly mitigated collinearity (VIFs less than 4) but the MSE is the highest.

Additionally, the reader may note an important relationship between raise regression and residualization: starting from expression (5.5), if λ tends to infinity, variable $\widetilde{\mathbf{A}}$ will tend to $\mathbf{e_A}$, and model (5.6) will turn into model (5.4). This fact is a very interesting issue for future research.

To conclude, besides the mitigation of strong essential collinearity problems, residualization provides an alternative interpretation of the original variables, as was explained above. The researcher has to choose first between introducing more or less bias into the model. If it is worth sacrificing the unbiased estimations in support of reducing the variance of the predicted values and improving the overall prediction accuracy, then the researcher has to choose between obtaining interesting and interpretable results as well as mitigating collinearity problems (residualization) or using traditional methodologies such as ridge regression that obtain findings that are difficult to interpret. From the perspective of this dissertation, the application of residualization is very substantial for empirical purposes because of the different and direct interpretations of the variables together with the mitigation of multicollinearity problems. According to Schroeder (1990), the advantage of the use of a biased method is that the theoretical model is not compromised and this biased

method has to stabilise the regression coefficients, reduce the error and render the model more generalisable, if not, it is better to use the initial model, so the idea supported by Schroeder (1990) seems to be closer to the use of residualization than to the use of ridge regression.

5.2 Model 2: the STIRPATE model in the European Union. Different uses of the residualization procedure

Nowadays, environmental policies are a significant cornerstone of a developed economy. In connection with this fact, the concept of "environmental risk" emerges, defined as the probability of damages to any community, due to the vulnerability of its environmental components exposed to human activities. For Greenpeace[3] or NASA[4], the solution comes from the energy sector with the use of renewable energy sources. In the context of the European Union (EU), the Europe 2020 strategy is a policy for the years 2010-2020 which includes environmental objectives regarding the climate change and energy targets. Its principal actions can be summarised into two main methods: diminishing the emissions to the atmosphere and increasing the energy efficiency of the countries. Although the strategy ends this year 2020, for this study the framework is relevant because it concerns one part of the studied years (1995-2014). For future research of interest, it is important to remark that the EU has extended the targets to 2030.

In light of the foregoing, it is clear that environmental efficiency and energy targets go together, and this fact is even more notable for the case of the EU.

[3] https://es.greenpeace.org/es/trabajamos-en/cambio-climatico/
[4] https://climate.nasa.gov

For the STIRPAT model defined in previous section, one important variable is technology. This variable has been divided (Section 5.1) into two: industry value added and CO_2 intensity. The last one could be interpreted as a measure of the efficiency of a country: when an economy is environmentally efficient, the CO_2 intensity, i.e. the number of kilograms of CO_2 emissions per kilograms of oil-equivalent energy use, will be lower. The first objective of this section is to define a variable that directly measures the environmental efficiency of a country. To that end, Data Envelopment Analysis (DEA) is used to determine this variable.

On the other hand, although this chapter focuses on the particular case of four countries (Portugal, Spain, Italy and Greece) due to their economic characteristics, it is important to remark that the efficiency scores are obtained for the EU as a whole to further study this particular variable and because of the characteristics of the methodology (it raises a relative measure). The four chosen countries are known as PIGS, which is an acronym originally referred, usually derogatorily, to the economies of the Southern European countries. The term was often used in reference to the growing debt and economic vulnerability of the Southern EU countries, and it was popularised during the European sovereign debt crisis.

Once the efficiency is obtained, the STIRPAT model is redefined as the STIRPATE model, using environmental efficiency scores instead of CO_2 intensity. It is going to be applied for the four Southern members of the EU, as it has been remarked. Predictably, the authors have found strong collinearity in the model and the residualization procedure has been implemented.

5.2.1 Data Envelopment Analysis (DEA) and energy efficiency sustainable index

The first objective is to measure the energy efficiency sustainable index. To that end, the authors employ the DEA approach, proposed by Charnes et al. (1978). It is a well-established non-parametric frontier approach that assesses and measures the relative efficiency of a set of comparable entities (called Decision Making Units or DMUs) featured with multiple factors grouped into two categories: inputs and outputs. Classical DEA models rely on the assumption that inputs have to be minimised and outputs have to be maximised (Vencheh et al. (2005)). Thus, in the standard DEA model, decreases in outputs are not allowed, only inputs are allowed to decrease and, similarly, increases in inputs are not allowed while only outputs are allowed to increase (Seiford and Zhu (2002)). However, the production process can also generate undesirable outputs (pollutants).

There are several approaches for incorporating undesirable outputs in the DEA modelling approach. These models can also be classified into two groups: the ones that take an indirect perspective and the ones that take a direct approach. As Scheel (2001) argues, indirect approaches transform the values of the undesirable outputs through a monotone decreasing function, such that the transformed data can be included as desirable outputs in the technology set; direct approaches can use the original output data set, but modify the assumptions about the structure of the technology set in order to treat the undesirable outputs appropriately. As Scheel (2001) remarks, the indirect approaches assume that the transformed data have their own meaning; for instance, if we transform the undesirable output mortality rate, we can then study the desirable output survival rate. In contrast, the direct

approach employs the original output set, but changes the assumptions adopted. The direct approach, suggested by Färe et al. (1989), replaces the strong disposability of outputs with the assumption that outputs are weakly disposable, while only the subvector of desirable outputs is strongly disposable. The direct approach is preferable, meaning that it is not necessary for researchers to make any changes to the main dataset, while it is not necessary to reinterpret the results obtained in terms of the "new" variables (e.g. mortality and survival rates).

The analysis in this chapter makes use of the DEA method, focusing on the direct approach, to calculate the energy efficiency sustainability index among the EU-28 members. It considers one of the models developed by Zhou and Ang (2008), who measure the energy efficiency performances of 21 OECD countries. The reason of using this particular model is due to the fact the analysis focuses on the technical efficiency of energy consumption. The technical efficiency is defined as the ability of a DMU to obtain maximum outputs (or minimum inputs) from a given set of inputs (or outputs), Robaina-Alves et al. (2015); Moutinho et al. (2017).

The principal advantage of using the DEA method is its flexibility in incorporating factors which a priori are not comparable (both inputs and outputs). That makes the results easily interpretable. As Balk et al. (2017) illustrate, the DEA method searches for the most favourable weight when evaluating a production unit, by constructing a virtual aggregate input to output productivity ratio, each constructed as a linear combination of observed values.

Assume that the set of DMUs consists of DMU_k, $k = 1, 2, \ldots, K$. Let $\mathbf{x}_{nk} = (\mathbf{x}_{1k}, \mathbf{x}_{2k}, \ldots, \mathbf{x}_{Nk})$, $\mathbf{e}_{lk} = (\mathbf{e}_{1k}, \mathbf{e}_{2k}, \ldots, \mathbf{e}_{Lk})$, $\mathbf{y}_{mk} = (\mathbf{y}_{1k}, \mathbf{y}_{2k}, \ldots, \mathbf{y}_{Mk})$ and $\mathbf{u}_{jk} = (\mathbf{u}_{1k}, \mathbf{u}_{2k}, \ldots, \mathbf{u}_{Jk})$ are the vectors of non-energy inputs, energy

inputs, desirable outputs and undesirable outputs, respectively. The efficiency score of DMU_i can be obtained by solving model (5.7) below:

$$
\begin{aligned}
&\min \theta_i \\
&\quad \text{s.t.} \\
&\qquad \sum_{k=1}^{K} z_k \ \mathbf{x}_{nk} \leq \mathbf{x}_{ni} \ , \quad n = 1, \ldots, N \\
&\qquad \sum_{k=1}^{K} z_k \ \mathbf{e}_{lk} \leq \theta_i \ \mathbf{e}_{li} \ , \quad l = 1, \ldots, L \\
&\qquad \sum_{k=1}^{K} z_k \ \mathbf{y}_{nk} \geq \mathbf{y}_{mi} \ , \quad m = 1, \ldots, M \\
&\qquad \sum_{k=1}^{K} z_k \ \mathbf{u}_{nk} = \mathbf{u}_{ji} \ , \quad j = 1, \ldots, J \\
&\qquad z_k \geq 0 \quad , \quad k = 1, 2, \ldots, K
\end{aligned} \tag{5.7}
$$

It can be seen that [model (5.7)] attempts to proportionally contract the amouns of energy inputs as much as possible for a given level of non-energy inputs, desirable and undesirable outputs. It provides an aggregated and standardized index for measuring energy efficiency performance (Zhou and Ang (2008)). The higher the value, the better the situation for each DMU. The maximum possible value is one, which implies that the DMU is relatively efficient, regarding the rest of DMUs. In contrast, if the value of the index is zero (the minimum possible value), it implies that the DMU is relatively inefficient.

It is important to remark that the DEA approach has certain limitations, despite the attractiveness of its application. More specifically, the flexibility in weight explained above can lead to implausible results, inconsistent with any prior knowledge of the production process (Balk et al. (2017)); in that

sense the results must be analysed carefully and comparing them with the theoretical framework and previous research. In this case, the problem is not encountered and the results are consistent. In addition, it does not allow the comparison of different results "externally"; the results of the analysis can be only compared "internally", i.e. it is not possible to compare the findings with any other dataset which would offer other different scores. DEA measures the relative efficiency of DMUs that perform similar types of functions and have identical goals and objectives; for instance, if we analyse a particular group of countries, we may not compare these results with any other groups, even in the case of introducing only one additional country. Apart from this minor inconvenience, the use of DEA provides the flexibility of the application: it is not necessary to explicitly specify a priori a production function that explains how the inputs and outputs of the production units are linked to each other (Cecchini et al. (2018)). Furthermore, DEA has emerged in recent years as a highly sophisticated method for assessing efficiency measures, and particularly, environmental efficiency across countries and economic sectors (Robaina-Alves et al. (2015)).

Once the methodology is clarified, the index is obtained. The data used to obtain it are grouped into the following categories: two types of inputs (non-energy and energy inputs) as well as two types of outputs (desirable and undesirable outputs). The data are available on the World Bank website[5]. The factors are measured as follows:

- Non-energy inputs:
  - Labour force (total, people ages 15 and older).
  - Gross capital formation (% of GDP).

[5] `https://databank.worldbank.org`

- Energy inputs:
  - Group 1: Energy use (kg of oil equivalent per capita).
  - Group 2:
    - Fossil fuel energy consumption (% of total).
    - Renewable energy consumption (% of total).
- Desirable output: GDP per capita, PPP (constant 2011 international $).
- Undesirable output: CO_2 emissions (kt).

In the case of energy inputs, we have two groups of variables: first, we obtain the efficiency scores using group 1 (with only one energy input: energy use) and then using group 2 (with two energy inputs: energy consumption distinguishing between fossil and non-fossil energies). The final variable is the average of these two energy efficiency scores. The reason for building the variable as above is to balance the energy efficiency results for dealing with the weight flexibility problem previously mentioned.

The analysis will provide the results for the energy efficiency sustainable index (**E**) across the different members of EU-28 (excluding Malta because of the unavailability of data) for each year, from 1995 to 2014 (last data available for CO_2 emissions). Variable **E** is called as sustainable or environmental energy efficiency because of the incorporation of the undesirable output, CO_2 emissions, which allows to obtain the energy efficiency scores taking into account the ecological effect of the economy on the environment. The results are reported in Tables 5.8 and 5.9. In addition, Figures 5.1 and 5.2 display the average of the period for each country and the average of each year for the total EU-28 (excluding Malta), respectively.

Table 5.8: Energy Efficiency Scores for EU-28 (excluding Malta). 1995-2004.

	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Austria	0.7117	0.7149	0.6987	0.6747	0.6701	0.6845	0.6467	0.667	0.6535	0.658
Belgium	1	1	1	1	1	1	1	1	1	1
Bulgaria	0.2101	1	0.2205	0.2308	0.223	0.2287	0.2259	0.2432	0.2557	0.2609
Croatia	0.39	0.432	0.4187	0.4143	0.3937	0.4204	0.4062	0.4131	0.429	0.4122
Cyprus	0.7389	0.6831	0.7565	0.8878	0.8313	0.8869	0.7092	0.6707	0.6717	0.6765
Czech Republic	0.6603	0.5845	0.7993	0.6124	0.5214	0.8296	0.7749	0.8624	0.8591	0.7423
Denmark	0.7738	0.8109	0.7692	0.7309	0.7401	0.7578	0.717	0.7455	0.738	0.7696
Estonia	0.4903	0.5237	0.5733	0.4932	0.4597	0.4977	0.4999	0.564	0.6696	0.7035
Finland	0.5864	0.6693	0.6952	0.6482	0.648	0.6079	0.63	0.7749	0.9276	0.9446
France	0.823	0.81	0.8236	0.8301	0.8389	0.8428	0.833	0.8357	0.8473	0.843
Germany	1	1	1	1	1	1	1	1	1	1
Greece	0.5481	0.5737	0.5609	0.5577	0.5491	0.543	0.5404	0.552	0.5896	0.5756
Hungary	0.3758	0.383	0.3855	0.4064	0.3989	0.4219	0.4096	0.4227	0.4442	0.4397
Ireland	0.8018	0.7398	0.9105	0.9674	0.9982	1	0.9242	0.7803	0.7637	0.7668
Italy	0.8417	0.8406	0.8396	0.8405	0.8389	0.839	0.837	0.8426	0.8519	0.8523
Latvia	0.2469	0.2714	0.2865	0.3082	0.3231	0.3481	0.3387	0.3655	0.3935	0.3874
Lithuania	0.2525	0.2673	0.2827	0.2926	0.3011	0.3342	0.314	0.3302	0.3681	0.3593
Luxembourg	1	1	1	1	1	1	1	1	1	1
Netherlands	1	1	0.9564	0.982	0.999	0.9429	0.8571	0.9107	0.9364	0.9946
Poland	0.4974	0.4959	0.5012	0.4911	0.4972	0.4966	0.4966	0.4985	0.5086	0.5226
Portugal	0.5908	0.6486	0.6244	0.611	0.5745	0.5887	0.576	0.558	0.5845	0.5428
Romania	0.3174	0.3207	0.3158	0.327	0.3465	0.3484	0.356	0.361	0.3817	0.3986
Slovak Republic	0.2744	0.2998	0.3038	0.3447	0.3179	0.3286	0.2965	0.3079	0.3341	0.3325
Slovenia	0.3923	0.4139	0.4048	0.417	0.4153	0.4328	0.4174	0.4302	0.4512	0.4305
Spain	0.6751	0.6976	0.6813	0.6896	0.6885	0.6893	0.6902	0.6963	0.718	0.7082
Sweden	0.7532	0.7574	0.7591	0.7379	0.746	0.7697	0.7378	0.7466	0.7693	0.7717
United Kingdom	0.8912	0.8866	0.9119	0.9088	0.9562	0.9573	0.9589	0.9548	0.9823	1

Table 5.9: Energy Efficiency Scores for EU-28 (excluding Malta). 2005-2014.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Austria	0.6598	0.6813	0.6768	0.6872	0.7047	0.6712	0.6653	0.6723	0.6235	0.6102
Belgium	1	1	0.9259	1	1	1	1	0.922	0.8331	0.7809
Bulgaria	0.2676	0.2757	0.2887	0.313	0.336	0.3278	0.3123	0.3223	0.3217	0.3057
Croatia	0.418	0.4316	0.4345	0.4637	0.4528	0.4391	0.4369	0.4461	0.4165	0.4155
Cyprus	0.6778	0.675	0.6662	0.6725	0.6789	0.6706	0.665	0.6629	0.6487	0.6421
Czech Republic	0.9677	1	1	1	0.959	0.825	0.9495	0.7885	0.5341	0.5829
Denmark	0.7785	0.7499	0.7254	0.7401	0.7555	0.7184	0.7131	0.7521	0.6823	0.7099
Estonia	0.7777	0.7346	0.9578	0.8877	0.7783	1	1	1	1	1
Finland	0.6847	1	0.8222	0.7756	0.7847	0.8595	0.8372	0.6748	0.6173	0.5764
France	0.849	0.8527	0.8595	0.8601	0.8641	0.8604	0.8643	0.863	0.8565	0.8484
Germany	1	1	1	1	1	1	1	1	1	1
Greece	0.6043	0.5893	0.6524	0.6406	0.6841	0.5794	0.5183	0.4887	0.4852	0.478
Hungary	0.4385	0.4476	0.4502	0.4668	0.4716	0.4504	0.4546	0.4672	0.4643	0.4652
Ireland	0.7745	0.7692	0.7934	0.7576	0.7542	0.7443	0.7491	0.7595	0.7537	0.7649
Italy	0.863	0.8467	0.9249	0.8391	0.838	0.8269	0.8265	0.8151	0.8026	0.7872
Latvia	0.4144	0.4418	0.4701	0.4824	0.4457	0.4025	0.4413	0.4578	0.4371	0.441
Lithuania	0.3785	0.401	0.4174	0.4403	0.4254	0.4351	0.4424	0.4566	0.4587	0.4655
Luxembourg	1	1	1	1	1	1	1	1	1	1
Netherlands	1	0.9842	0.9771	1	1	1	1	1	1	1
Poland	0.5308	0.5435	0.5648	0.5633	0.5875	0.5869	0.6092	0.5939	0.5963	0.6172
Portugal	0.5288	0.5468	0.5529	0.5745	0.576	0.5866	0.5709	0.5721	0.5298	0.5223
Romania	0.4112	0.4246	0.4489	0.489	0.5123	0.4896	0.4773	0.4901	0.5103	0.5071
Slovak Republic	0.3438	0.3679	0.4035	0.4296	0.4411	0.4325	0.4432	0.4656	0.4404	0.4558
Slovenia	0.436	0.4465	0.4643	0.4762	0.4803	0.4619	0.4596	0.4552	0.4283	0.441
Spain	0.7245	0.7254	0.734	0.7399	0.7501	0.7376	0.7163	0.7037	0.6997	0.6925
Sweden	0.7732	0.7985	0.7896	0.7877	0.7987	0.7737	0.7655	0.7596	0.7453	0.7514
United Kingdom	1	1	1	1	1	1	1	1	1	1

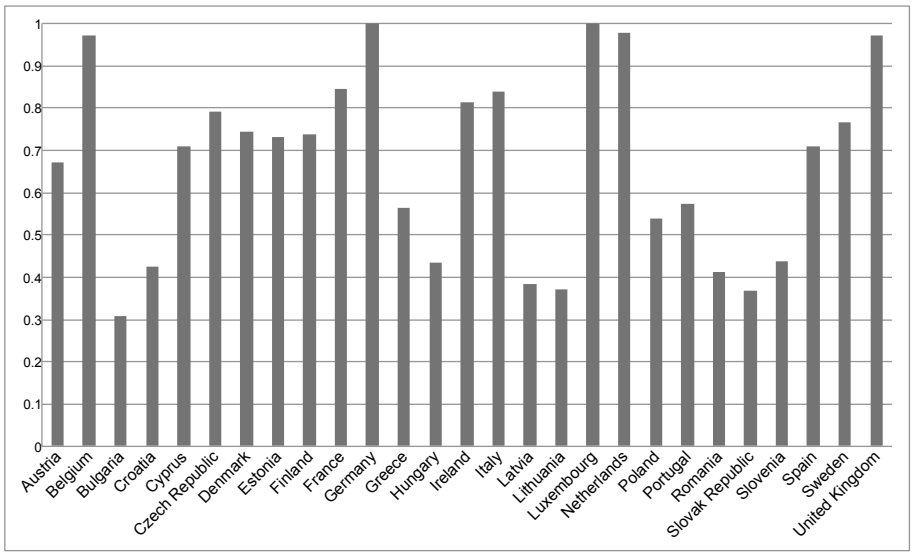

Figure 5.1: Average of Energy Efficiency Scores for each member: EU-28 (excluding Malta), 1995-2014.

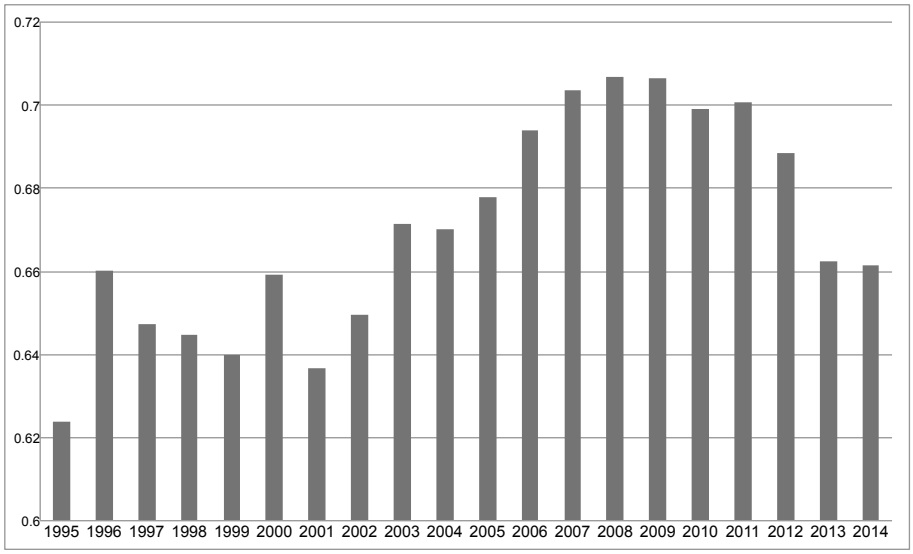

Figure 5.2: Average of Energy Efficiency Scores for each year: EU-28 (excluding Malta), 1995-2014.

The results of this section indicate that the most energy efficient countries are Germany and Luxembourg, followed by the Netherlands, the UK and Belgium. By contrast, the least energy efficient member state is Bulgaria, followed by Slovak Republic, Lithuania and Latvia. Therefore, it is obvious that there are two groups of countries with regard to the efficiency of their energy sector, so it may be thought that there is a nexus between energy efficiency and income. In other words, while the most efficient countries are in the group of the countries with higher GDP per capita on average, the inefficient energy economies are in the group with smaller GDP per capita. With a lower GDP per capita, the country has fewer available resources to invest in new energy technologies and environmentally friendly technologies. The findings also indicated that the energy efficiency scores did not change dramatically through the time span under consideration (they took values between 0.6 and 0.7), but it could be seen that the most efficient years were those of 2007, 2008 and 2009, while the worst values were at the beginning of the time period. The controversial issue here is the most efficient years, because we do not have an increasing trend regarding the efficiency scores: the results present an increasing trend until 2009, when the values started to decrease again. A potential explanation could be the economic (both financial and sovereign debt) crisis in Europe during those years.

5.2.2 The STIRPATE model for Portugal, Spain, Italy and Greece

The dataset, except **E**, which has been calculated in previous subsection, is obtained from the World Bank website[6], and it includes data on each particular

[6] `https://databank.worldbank.org`

country of the EU (Portugal, Spain, Italy and Greece) for the period 1995-2014. The information regarding the variables used in this example is shown in Table 5.10.

The STIRPAT model that is going to be used in this section is the following:

$$\mathbf{I} = \beta_1 + \beta_2\mathbf{P} + \beta_3\mathbf{A} + \beta_4\mathbf{T} + \beta_5\mathbf{E} + \mathbf{u}, \tag{5.8}$$

where $\mathbf{u}$ is the random disturbance, which is supposed to be spherical.

Because of the introduction of variable **E**, the model has been renamed as the STIRPATE to differentiate it from the original STIRPAT model. As Kilbourne and Thyroff (2020) suggest, *future research on the STIRPAT should consider expanding the model to add renewable energy to the equation*, and this new variable **E** includes not only the use of renewable energy but also other interesting variables that lead to the reaching of conclusions about a complex environmental variable.

Regarding the expected signs of each variable, those of variables **P** and **T** are controversial, and this idea has already been expressed in Section 5.1.

In the case of affluence (variable **A**), in Section 5.1 above it was argued that if the researcher is studying a group of countries with different characteristics, the GDP will show the level of development or wealth of each one; the higher the GDP, the greater the possibilities of devoting resources to climate targets and the expected sign for the parameter will be negative. But in the case of the European Union and, particularly, in the case of Portugal, Spain, Italy and Greece, this fact cannot be applied. Because of their similarities, the GDP does not show the level of development of each one (they have similar level of development), but shows the real production of the country analysed: higher production implies more pollution going into the atmosphere; the attitudes in each country are similarly environmentally friendly, indeed this group

Table 5.10: Variables of STIRPATE model. The case of Portugal, Spain, Italy and Greece.

Notation	Variable	Unit	Expected sign
I	CO_2 emissions	kilotonnes	-
P	population growth	annual %	Negative / Positive
A	GDP	billions of constant 2010 US$	Positive
T	industry value added	billions of constant 2010 US$	Negative / Positive
E	environmental energy efficiency	-	Negative

of countries invest similar resources on the preservation of the environment. Hence, the expected sign of the parameter will be positive. Additionally, in this example variable **A** is not per capita GDP but the absolute value of the GDP. The use of GDP instead of GDP per capita as variable **A** aims to overcome the following: from an interpretative point of view, a very important issue of the economy is ignored when using the GDP per capita since the distribution of income and the level of development of each region are disregarded when all people are considered equal in terms of earnings, thus it would be better to use GDP instead of GDP per capita. This fact is explained in depth in Section 5.3 below.

Finally, the efficiency scores (variable **E**) obtained in Subsection 5.2.1 are expected to have a negative parameter: environmental efficiency may be interpreted as a variable that will influence negatively to emissions (more environmental energy efficiency implies less pollution). This example uses variable **E** instead of $\mathbf{T}_2$ from Section 5.1 above to provide a new important variable for the STIRPAT model, redefining it as the STIRPATE model. The minimum posible value (zero) represents the less environmentally energy efficient countries, and the higher possible value of this variable (one) represents the higher environmentally energy efficient countries, hence it is clear that the obtained efficiency is related to the CO_2 intensity (kg per kg of oil-equivalent energy use). As was said in Subection 5.2.1, the higher efficient countries, the less polluting countries, therefore $\mathbf{T}_2$ from Section 5.1 above could be interpreted as the opposite of variable **E**. This idea confirms the negative expected sign of the corresponding parameter of variable **E**.

The results of the estimation of model (5.8) for the four countries are presented in Tables 5.11 to 5.14. Additionally, with regard to the validation of

the model of each country:

- In the case of Portugal, the model does not present heteroscedasticity: the White test concludes in not rejecting the null hypothesis of homoscedasticity (p value higher than 0.05).

  Regarding collinearity, there is non-essential collinearity for variables **A** and **E** (CV(**A**) = 0.060 and CV(**E**) = 0.054), hence, these variables are centred: $\mathbf{A}' = \mathbf{A} - \overline{\mathbf{A}}$ and $\mathbf{E}' = \mathbf{E} - \overline{\mathbf{E}}$, where $\overline{\mathbf{A}}$ and $\overline{\mathbf{E}}$ represent the mean of each particular variable. Paying attention to the VIF values in the case of Portugal, the reader will note that the model presents worrying essential collinearity, and the problematic variables are, in this case, **P** and **T**.

- In the case of Spain, the model does not present heteroscedasticity as well: the White test concludes in not rejecting the null hypothesis of homoscedasticity (p value higher than 0.05).

  Regarding collinearity, there is non-essential collinearity for variable **E** (CV(**E**) = 0.030), hence, this variable is centred: $\mathbf{E}' = \mathbf{E} - \overline{\mathbf{E}}$. Paying attention to the VIF values, it is clear that the model for Spain presents strong essential collinearity, and the problematic variable is **T**.

- In the case of Italy, there is no heteroscedasticity problems: the White test concludes in not rejecting the null hypothesis of homoscedasticity (p value higher than 0.05).

  Regarding collinearity, there is non-essential collinearity for variables **A**, **T** and **E** (CV(**A**) = 0.045, CV(**T**) = 0.059 and CV(**E**) = 0.031), hence, these three variables are centred: $\mathbf{A}' = \mathbf{A} - \overline{\mathbf{A}}$, $\mathbf{T}' = \mathbf{T} - \overline{\mathbf{T}}$ and

Table 5.11: Results of model (5.8). Portugal.

		ORIGINAL MODEL	VIF$_i$
Intercept	Estimator (s.d.)	69.06 (18400)	-
P	Estimator (s.d.)	10500 * (4082)	11.379
A′	Estimator (s.d.)	-1301000000 (795500000)	5.360
T	Estimator (s.d.)	1101 * (391.3)	11.646
E′	Estimator (s.d.)	-110900 *** (22120)	2.451
CN		2.751	
Determ. corr. matrix		0.038	
R^2		0.942	
F statistic		61.03 ***	

***, * Statistically significant at 0.001 (99.9% confidence level) and at 0.05 (95% confidence level), respectively.

$\mathbf{E}' = \mathbf{E} - \overline{\mathbf{E}}$. Paying attention to the VIF values, there is no strong essential collinearity: all values are lower than 10.

- In the case of Greece, the model does not present heteroscedasticity: the White test concludes in not rejecting the null hypothesis of homoscedasticity (p value higher than 0.05).

  Regarding collinearity, there is no non-essential collinearity. Paying attention to the VIF values, the model presents essential collinearity, and here there is only one problematic variable: **A**.

As stated earlier, in the cases of Portugal, Spain and Greece the model presents worrying essential collinearity, which is not the case for Italy. Therefore,

Table 5.12: Results of model (5.8). Spain.

		ORIGINAL MODEL	VIF_i
Intercept	Estimator (s.d.)	-30180 (39850)	-
P	Estimator (s.d.)	21010 * (88295)	9.197
A	Estimator (s.d.)	1237000000 (2362000000)	9.870
T	Estimator (s.d.)	802.6 ** (242.4)	23.337
$\mathbf{E}'$	Estimator (s.d.)	-477000 ** (151700)	2.783
CN		87.921	
Determ. corr. matrix		0.011	
R^2		0.967	
F statistic		109.5 ***	

***, **, * Statistically significant at 0.001 (99.9% confidence level), at 0.01 (99% confidence level) and at 0.05 (95% confidence level), respectively.

Table 5.13: Results of model (5.8). Italy.

		ORIGINAL MODEL	VIF_i
Intercept	Estimator (s.d.)	441300 *** (6374)	-
P	Estimator (s.d.)	-40140 * (14440)	1.221
$\mathbf{A}'$	Estimator (s.d.)	-1219000000 (5164000000)	3.975
$\mathbf{T}'$	Estimator (s.d.)	1145 ** (365.3)	6.303
$\mathbf{E}'$	Estimator (s.d.)	171400 (267000)	2.754
CN		4.875	
Determ. corr. matrix		0.092	
R^2		0.853	
F statistic		21.74 ***	

***, **, * Statistically significant at 0.001 (99.9% confidence level), at 0.01 (99% confidence level) and at 0.05 (95% confidence level), respectively.

Table 5.14: Results of model (5.8). Greece.

		ORIGINAL MODEL	VIF_i
Intercept	Estimator (s.d.)	50630 *** (8265)	-
P	Estimator (s.d.)	11160 ** (3300)	4.161
A	Estimator (s.d.)	831400000 (636200000)	10.742
T	Estimator (s.d.)	636.4 ** (200.9)	7.339
E	Estimator (s.d.)	-23590 (26080)	4.993
CN		56.641	
Determ. corr. matrix		0.033	
R^2		0.934	
F statistic		52.74 ***	

***, ** Statistically significant at 0.001 (99.9% confidence level) and at 0.01 (99% confidence level), respectively.

residualization is going to be applied for the cases of Portugal, Spain and Greece. In particular, the procedure is going to be applied in three different ways:

- For the case of Portugal, the typical procedure is applied: variable **T** (which is the one with highest VIF) will be residualized from the rest of explanatory variables of the model, using the following auxiliary regression:

$$\mathbf{T} = \alpha_1 + \alpha_2\mathbf{P} + \alpha_3\mathbf{A}' + \alpha_4\mathbf{E}' + \mathbf{v}, \tag{5.9}$$

  whose OLS estimation leads to residuals $\mathbf{e_T}$.

- For the case of Spain, only one of the explanatory variables will be used in the residualization procedure to isolate the effect of **T** (which is the

one with highest VIF) from it, using the following auxiliary regression:

$$\mathbf{T} = \alpha_1 + \alpha_2 \mathbf{A} + \mathbf{v}, \tag{5.10}$$

whose OLS estimation leads to residuals $\mathbf{e_T}$.

- For the case of Greece, one external variable will be used to isolate the effect of variable **A** (which is the variable with highest VIF), using the following auxiliary regression:

$$\mathbf{A} = \alpha_1 + \alpha_2 \, \mathbf{Population} + \mathbf{v}, \tag{5.11}$$

where **Population** is the absolute value of people living in Greece for each year[7]. Its OLS estimation leads to residuals $\mathbf{e_A}$.

Therefore, model (5.8) for each country will be modified in the following ways:

$$\text{PORTUGAL} \ \rightarrow \ \mathbf{I} = \gamma_1 + \gamma_2 \mathbf{P} + \gamma_3 \mathbf{A}' + \gamma_4 \mathbf{e_T} + \gamma_5 \mathbf{E}' + \mathbf{w}. \tag{5.12}$$

$$\text{SPAIN} \ \rightarrow \ \mathbf{I} = \gamma_1 + \gamma_2 \mathbf{P} + \gamma_3 \mathbf{A} + \gamma_4 \mathbf{e_T} + \gamma_5 \mathbf{E}' + \mathbf{w}. \tag{5.13}$$

$$\text{GREECE} \ \rightarrow \ \mathbf{I} = \gamma_1 + \gamma_2 \mathbf{P} + \gamma_3 \mathbf{e_A} + \gamma_4 \mathbf{T} + \gamma_5 \mathbf{E} + \mathbf{w}. \tag{5.14}$$

The reason of using the residualization procedure in three different ways is to show the reader the possibilities of the methodology. Apart from mitigating strong essential collinearity, the use of variations of the procedure leads to different interpretations of the variables, depending on the goals and interests of the researcher. For this example:

- In the case of Portugal, the original method is applied. All the rest of the explanatory variables have been included in the auxiliary regression. This

[7]Note that, in this section, variable **P** is the population growth.

makes sense from an interpretative point of view because technology will be isolated from GDP (the level of wealth), the population growth and the environmental energy efficiency (which could be related to technological improvements); hence, variable $\mathbf{e_T}$ will represent the real and isolated influence of the level of the industry on the environment.

- In the case of Spain, variable **T** is isolated only from the effect of the GDP, to show the influence on the environment of the part of industry not related to the wealth of the country.

- In the case of Greece, an interesting alternative is applied. Although the use of an external variable in the auxiliary regression does not ensure the mitigation of essential collinearity in the model (the problem appears among the explanatory variables used in the original model), in this case the total population and population growth are closely interrelated, so if the variable **P** (population growth) is related to some other explanatory variable, it could be interpreted that the total population is also related to it. This fact is demonstrated by applying residualization: in the auxiliary regression, the variable "total population" is used instead of **P** (population growth), and the residuals obtained from the auxiliary regression allow mitigation of the problem in the model (see Table 5.17).

Tables 5.15, 5.16 and 5.17 display the results of the estimation of models (5.12), (5.13) and (5.14), respectively.

Now, regarding strong collinearity problems, the reader will see that they are mitigated: the VIFs indicate there are no problematic variables, the determinant of the correlation matrix as well, and the condition number is lower than 30 in all cases.

Table 5.15: Results of model (5.12): Portugal.

		RESIDUALIZATION Model (5.12)	VIF_i
Intercept	Estimator (s.d.)	51810 *** (523.5)	-
P	Estimator (s.d.)	21270 *** (1419)	1.375
$\mathbf{A}'$	Estimator (s.d.)	561400000 (441200000)	1.649
$\mathbf{e_T}$	Estimator (s.d.)	1101 * (391.3)	1.000
$\mathbf{E}'$	Estimator (s.d.)	-134200 *** (20510)	2.108
CN		2.407	
Determ. corr. matrix		0.447	
R^2		0.942	
F statistic		61.03 ***	

***, * Statistically significant at 0.001 (99.9% confidence level) and at 0.05 (95% confidence level), respectively.

Regarding the final results of each country:

- In the case of Portugal (Table 5.15):
  - For population, the corresponding parameter is positive and individually significant. It has been said that both signs (positive or negative) are consistent with theory. The conclusion that could be reached here is that the traditional perspective of Malthus (1973) (expressed in Section 5.1), who proposes that the population exerts pressure on the environment, is more consistent with the results obtained than the theory of Boserup (1981) (also expressed in Section 5.1).
  - Regarding affluence, the corresponding parameter is positive (as

Table 5.16: Results of model (5.13): Spain.

		RESIDUALIZATION Model (5.13)	VIF_i
Intercept	Estimator (s.d.)	-60380 (44420)	-
P	Estimator (s.d.)	21010 * (8295)	9.197
A	Estimator (s.d.)	11460000000 *** (1672000000)	4.944
$\mathbf{e_T}$	Estimator (s.d.)	802.6 ** (242.4)	6.463
$\mathbf{E}'$	Estimator (s.d.)	-477000 ** (151700)	2.783
CN		7.433	
Determ. corr. matrix		0.039	
R^2		0.967	
F statistic		109.5 ***	

***, **, * Statistically significant at 0.001 (99.9% confidence level), at 0.01 (99% confidence level) and at 0.05 (95% confidence level), respectively.

expected) but is non-significant, hence the conclusion that could be reached here is that affluence is not a relevant factor.

– The parameter of $\mathbf{e_T}$ is positive and individually significant. As for variable **P**, both signs are acceptable, and a positive sign agrees with the perspective of Malthus (1973).

– Finally, the corresponding parameter for variable **E** is negative, which is consistent with expectations and with the correlation matrix

Table 5.17: Results of model (5.14): Greece.

		RESIDUALIZATION Model (5.14)	VIF_i
Intercept	Estimator (s.d.)	34810 ** (11720)	-
P	Estimator (s.d.)	9541 *** (2282)	2.061
$\mathbf{e_A}$	Estimator (s.d.)	-1159000000 (764700000)	6.257
T	Estimator (s.d.)	999.1 *** (135.6)	3.486
E	Estimator (s.d.)	13520 (19220)	2.792
CN		20.605	
Determ. corr. matrix		0.056	
R^2		0.936	
F statistic		54.73 ***	

***, ** Statistically significant at 0.001 (99.9% confidence level) and at 0.01 (99% confidence level), respectively.

(5.15), and it is individually significant.

$$\begin{pmatrix} & \mathbf{I} & \mathbf{P} & \mathbf{A'} & \mathbf{T} & \mathbf{E'} \\ \mathbf{I} & 1.000 & & & & \\ \mathbf{P} & 0.765 & 1.000 & & & \\ \mathbf{A'} & 0.206 & -0.239 & 1.000 & & \\ \mathbf{T} & 0.893 & 0.704 & 0.449 & 1.000 & \\ \mathbf{E'} & -0.094 & 0.512 & -0.621 & -0.047 & 1.000 \end{pmatrix}. \quad (5.15)$$

- In the case of Spain (Table 5.16), the results are very similar to those of Portugal:
  - For population, the corresponding parameter is positive and individually significant, so the traditional perspective of Malthus

(1973) is supported.

- Regarding affluence, the corresponding parameter is positive, as it was expected, and it is individually significant.
- The parameter of $\mathbf{e_T}$ is positive and individually significant.
- Finally, the parameter of variable **E** is negative (as expected) and individually significant as well, although in the case of Spain it is not consistent with the correlation matrix (5.16). However, although the expected sign (paying attention only to the correlation matrix) was positive, it makes no sense to argue that environmental efficiency has a positive impact on the environment, therefore the theoretical interpretation of the parameter carries more weight.

$$\begin{pmatrix} & \mathbf{I} & \mathbf{P} & \mathbf{A} & \mathbf{T} & \mathbf{E'} \\ \mathbf{I} & 1.000 & & & & \\ \mathbf{P} & 0.923 & 1.000 & & & \\ \mathbf{A} & 0.715 & 0.538 & 1.000 & & \\ \mathbf{T} & 0.950 & 0.865 & 0.850 & 1.000 & \\ \mathbf{E'} & 0.551 & 0.527 & 0.791 & 0.727 & 1.000 \end{pmatrix}. \tag{5.16}$$

- In the case of Italy (Table 5.13):
  - For population, the corresponding parameter is negative and individually significant. As it has been said, both signs are accepted. The conclusion that could be reached here is that for Italy, population affects pollution negatively, which supports the idea of Boserup (1981) expressed in Section 5.1; thus, Italy could be interpreted as more technologically innovative. This sign is consistent not only with expectations but also with the correlation

matrix (5.17).

$$\begin{pmatrix} & \mathbf{I} & \mathbf{P} & \mathbf{A}' & \mathbf{T}' & \mathbf{E}' \\ \mathbf{I} & 1.000 & & & & \\ \mathbf{P} & -0.431 & 1.000 & & & \\ \mathbf{A}' & 0.629 & 0.113 & 1.000 & & \\ \mathbf{T}' & 0.862 & -0.127 & 0.832 & 1.000 & \\ \mathbf{E}' & 0.754 & -0.213 & 0.577 & 0.784 & 1.000 \end{pmatrix}. \tag{5.17}$$

- Regarding affluence, the corresponding parameter is negative, which is inconsistent with expectations, but it is non-significant, hence, the conclusion is that affluence is not a relevant factor.

- The parameter of **T** is positive and individually significant, as for the previous countries.

- The estimated parameter of variable **E** has a positive sign, which is consistent with the correlation matrix (5.17), but not with the theoretical interpretation. However, the corresponding parameter of efficiency is not individually significant, thus the inconsistency with theory and expectations is supported by the results of the model.

• In the case of Greece (Table 5.17):

- For population, as in Portugal and Spain, the corresponding parameter is positive and individually significant, supporting the theory of Malthus (1973). This sign is also consistent with the

correlation matrix (5.18).

$$\begin{pmatrix} & \mathbf{I} & \mathbf{P} & \mathbf{A} & \mathbf{T} & \mathbf{E} \\ \mathbf{I} & 1.000 & & & & \\ \mathbf{P} & 0.656 & 1.000 & & & \\ \mathbf{A} & 0.712 & 0.088 & 1.000 & & \\ \mathbf{T} & 0.918 & 0.432 & 0.840 & 1.000 & \\ \mathbf{E} & 0.724 & 0.563 & 0.710 & 0.693 & 1.000 \end{pmatrix}. \tag{5.18}$$

- Regarding affluence, for the case of Greece the results are similar to Italy: the parameter is negative and it is not individually significant, hence the inconsistency of this value is supported by the results of the model.
- The parameter of **T** is positive and individually significant, as for the rest of countries.
- Finally, the result for variable **E** is similar to those of Italy: the parameter has a positive sign, which is consistent with the correlation matrix (5.18) but inconsistent with the theoretical interpretation, however, the parameter is not individually significant, thus the inconsistency with theory and expectations is supported by the results of the model.

Previously, in subsection 5.2.1, it was concluded that there is a nexus between energy efficiency and income: the most efficient countries are in the group of the countries with higher GDP per capita on average and the inefficient energy economies are in the group with smaller GDP per capita. Nevertheless, the results of the STIRPATE models studied have not found a high relationship between efficiency and GDP per capita. The correlation between variables **A**

and **E** is lower than 0.8 (in absolute value) for the four countries in the study (see expressions (5.15), (5.16), (5.17) and (5.18)). In any case, other relationships have been detected among the explanatory variables and residualization has been applied for mitigating essential collinearity problems in the models of Portugal, Spain and Greece in three different ways, as explained above. After modifying these three models, interesting conclusions about the signs and the importance of the parametes of STIRPATE models have been achieved. As a whole, taking into account the above results, it can be concluded that the STIRPATE model is successful.

5.3 Model 3: the STIRPAT model in China. New interpretations of the variables

The third model is also based on the STIRPAT model, using data from China (1990-2014), the most polluting country in the world, as revealed by the World Bank, with a CO_2 emissions value of 10291926.878 kilotonnes (kt) in 2014. The dataset has been extracted from the World Bank website[8] and information regarding the variables is presented in Table 5.18.

With regard to the expected signs of the variables, **P** and **T** are controversial, and the idea has been expressed in Section 5.1. The variable that is worth to mention is GDP (both GDP per capita, $\mathbf{A}_{pc}$, and total GDP, **A**). In Section 5.1, the expected sign was negative, while in Section 5.2, the expected sign was positive. In this example, the second perspective is more appropriate: it has to be taken into account that in this section only one country is observed and this particular country is China. By observing data from China without considering more countries of the world, it could be interpreted that higher GDP, higher

[8] `https://databank.worldbank.org`

Table 5.18: Variables of STIRPAT model. The case of China.

Notation	Variable	Unit	Expected sign
I	CO_2 emissions	kilotonnes	-
P	total of population	billions of people	Negative / Positive
$\mathbf{A}_{pc}$	GDP per capita	trillions of constant 2010 US$	Positive
A	GDP	trillions of constant 2010 US$	Positive
T	industrialization	% of GDP	Negative / Positive

domestic production, implies more pollution going into the atmosphere, thus, the expected sign for this parameter will be positive.

The traditional specification of the STIRPAT model is:

$$\mathbf{I} = \beta_1 + \beta_2\ \mathbf{P} + \beta_3\ \mathbf{A}_{pc} + \beta_4\ \mathbf{T} + \mathbf{u}, \tag{5.19}$$

where $\mathbf{u}$ is the random disturbance, which is supposed to be spherical.

Although the starting point is model (5.19), the following specification is proposed:

$$\mathbf{I} = \gamma_1 + \gamma_2\ \mathbf{P} + \gamma_3\ \mathbf{e_A} + \gamma_4\ \mathbf{T} + \mathbf{w}, \tag{5.20}$$

where $\mathbf{e_A}$ are the residuals of the following auxiliary regression:

$$\mathbf{A} = \alpha_1 + \alpha_2\ \mathbf{P} + \alpha_3\ \mathbf{T} + \mathbf{v}. \tag{5.21}$$

The use of model (5.20) instead of model (5.19) intends to overcome the following disadvantages:

- Traditionally, per capita GDP has been used to avoid the existing dependency between the GDP and the population. However, as the reader will see in the following correlation matrix, the linear relationship between per capita GDP and population is higher than the relationship between GDP and population, i.e. the linear relationship is not mitigated but increased.

$$\begin{pmatrix} & \mathbf{I} & \mathbf{P} & \mathbf{A}_{pc} & \mathbf{A} & \mathbf{T} \\ \mathbf{I} & 1.0000 & & & & \\ \mathbf{P} & 0.8896 & 1.0000 & & & \\ \mathbf{A}_{pc} & 0.9910 & 0.9111 & 1.0000 & & \\ \mathbf{A} & 0.9905 & 0.9081 & 0.9999 & 1.0000 & \\ \mathbf{T} & -0.5464 & -0.6194 & -0.6296 & -0.6320 & 1.0000 \end{pmatrix}. \tag{5.22}$$

- In STIRPAT studies, $\mathbf{A}_{pc}$ is usually taken as the variable that represents affluence. However, the use of $\mathbf{A}_{pc}$ presents a disadvantage. From an interpretative point of view, a very important issue of the economy is ignored when using the per capita GDP (the ratio between GDP and population) since the distribution of income and the level of development of each region of the country are disregarded when all people are considered equal in terms of earnings, as has been noted in Section 5.2. An increase in the GDP per capita does not necessarily mean the country is more developed; it can also indicate that the richest people in the country have increased their income.

Furthermore, variable $\mathbf{T}$ is also included in the auxiliary regression (5.21) in order to isolate variable $\mathbf{A}$ from the industry sector as well.

First, the models have been validated. In relation to heteroscedasticity, the White test concludes in not rejecting the null hypothesis of homoscedasticity (p value higher than 0.05).

The VIF values from model (5.19) are: $\text{VIF}_{\mathbf{P}} = 6.010$, $\text{VIF}_{\mathbf{A}_{pc}} = 6.137$ and $\text{VIF}_{\mathbf{T}} = 1.691$, hence in terms of essential multicollinearity, this model does not present worrying problems. Therefore, the residualization procedure in this example is applied for empirical purposes. In any case, non-essential collinearity appears since $\text{CV}(\mathbf{P}) = 0.053$ and $\text{CV}(\mathbf{T}) = 0.026$ and the CN has a high value (see Table 5.19).

In order to mitigate the existing non-essential collinearity in model (5.20), the variables population and technology have been centred. Thus, model (5.20) is modified as follows:

$$\mathbf{I} = \gamma_1 + \gamma_2\ \mathbf{P'} + \gamma_3\ \mathbf{e_A} + \gamma_4\ \mathbf{T'} + \mathbf{w}, \tag{5.23}$$

where $\mathbf{P'} = \mathbf{P} - \overline{\mathbf{P}}$ and $\mathbf{T'} = \mathbf{T} - \overline{\mathbf{T}}$.

Furthermore, because of the use of residualization, the relationship between GDP and population is suppressed. In this case, the relationship between GDP and industrialization (variable $\mathbf{T}$) is also deleted. Indeed, $\mathbf{e_A}$ coincides with the part of GDP that has no relationship with population and industrialization. If $\mathbf{A}_{pc}$ could be interpreted as a tool for measuring the enrichment of the people and not the enrichment of the country, $\mathbf{e_A}$ would be interpreted as a tool that measures whether the countries, and not the people, are richer in economic terms that are unrelated to industry.

The results obtained by using OLS estimation of models (5.19) and (5.23) are shown in Table (5.19).

With model (5.23), it is verified that the degree of the existing near multicollinearity (essential and non-essential) is not worrisome. The values of VIF are lower than 4 ($\text{VIF}_{\mathbf{P'}} = 1.622$, $\text{VIF}_{\mathbf{e_A}} = 1.000$ and $\text{VIF}_{\mathbf{T'}} = 1.622$).

Taking into account the results obtained, the reader will observe the following:

- In model (5.19), the intercept has a parameter that is significantly different from zero and has a negative value, i.e. if population and GDP were null, the CO_2 emissions would be negative. This situation is corrected with the model (5.23).

- In model (5.19), the estimated parameter for population is not significantly different from zero; by contrast, in model (5.23), the parameter is significant, and it has a positive value, i.e. when the population increases, the CO_2 emissions also increase. This is in line with the economic theory and the correlation matrix.

- In models (5.19) and (5.23), the GDP parameter (obtained from $\mathbf{A}_{pc}$ and $\mathbf{e}_{\mathbf{A}}$, respectively) is significantly different from zero and has a positive value. However, the interpretations of the two estimated parameters are different. While in model (5.19) it can be concluded that the increase in the enrichment of the people (supposing all people are equal in terms of earnings) implies an increase in the CO_2 emissions, in model (5.23), it can be concluded that the increase in the wealth of the country when the production of goods and services is unrelated to industrialization entails an increase in CO_2 emissions.

- In model (5.19), the estimated parameter for industrialization is significant and has a positive value, which is contrary to the sign expected by observing the correlation matrix. However, according to the theory both signs were acceptable. In any case, in model (5.23), this parameter is not significantly different from zero, so it can be concluded that technology is not a relevant factor for this study.

5.4 Discussion

Although the dependence between the main explanatory factors of environmental damage is evident, it is usually neglected in the scientific literature. In this dissertation, residualization has been proposed as an alternative to be used not only with the goal of mitigating collinearity problems, but also with the objective of analysing the causal effects of the driving forces affecting collinearity together with new interpretations of the variables affected by the procedure. Throughout the chapter, the main goal has been to clarify the role of collinearity in environmental models and to show how results are

Table 5.19: Results of STIRPAT models (5.19) and (5.23).

		OLS Original model (5.19)	RESIDUALIZATION Transformed model (5.23)
Intercept	Estimator	-10287191 *	5405875 ***
	(s.d.)	(3667784)	(53166)
$\mathbf{P}$	Estimator	-1790259	
	(s.d.)	(1861596)	
$\mathbf{P}'$	Estimator		35773988 ***
	(s.d.)		(1004251)
$\mathbf{A}_{pc}$	Estimator	1837647 ***	
	(s.d.)	(77813)	
$\mathbf{e_A}$	Estimator		1300875 ***
	(s.d.)		(57278)
$\mathbf{T}$	Estimator	409211 ***	
	(s.d.)	(80784)	
$\mathbf{T}'$	Estimator		24250
	(s.d.)		(82153)
CN		74.890	2.139
R^2		0.9924	0.9918
F statistic		918.2 ***	851.2 ***

***, * Statistically significant at 0.001 (99.9% confidence level) and at 0.05 (95% confidence level), respectively.

influenced by the methodology of estimation selected. Since environmental research is used for policimaking, the results of this chapter mean that the use of alternative methodologies such as residualization may allow to obtain policy recommendations based on firm statistical results and not subject to statistical instability when serious collinearity appears. This idea can be extended to many different fields.

The first empirical analysis (Section 5.1) has addressed collinearity with different methodologies to analyse how they affect the estimations. Although ridge regression mitigates collinearity problems mechanically and presents

the smallest MSE, this method presents various important disadvantages (see Section 2.3.1). In turn, residualization maintains the initial properties of the model (experimental F, sum of squares and R^2) and it mitigates strong collinearity problems, although it has the highest MSE. As stated in the corresponding section, the use of biased methods has some advantages, so the researcher has to decide if it is worth sacrificing the unbiased estimations in support of reducing the variance of the predicted values and improving the overall prediction accuracy. Once the decision to sacrifice the unbiased estimations is made, the researcher has to choose between obtaining interesting and interpretable results as well as mitigating collinearity problems (residualization) or using traditional methodologies such as ridge regression that obtain findings that are difficult to interpret.

The second empirical analysis (Section 5.2) studies two important issues: the evolution of the countries of the European Union regarding the environmental energy efficiency and whether this efficiency could be interpreted as a relevant variable in the traditional STIRPAT model. With this two objectives, the efficiency scores were obtained and a renewed version of the STIRPAT has been studied: the STIRPATE model. The efficiency results for the whole European Union indicate that there is a nexus between energy efficiency and income. Although this relationship is evident in Subsection 5.2.1, the results of the STIRPATE model in Subsection 5.2.2 did not find a high relationship between efficiency and GDP per capita in the cases of Portugal, Spain, Italy and Greece. The correlation between variables **A** and **E** is lower than 0.8 for the four countries in the study. In any case, residualization is applied for mitigating essential collinearity problems in the models of Portugal, Spain and Greece in three different ways, as explained earlier, to show the reader the possibilities of the application of the residualization procedure. After modifying these

three STIRPATE models, interesting conclusions about the performance of the STIRPATE model have been achieved. As a whole, regarding the importance of variable **E**, the results obtained support the success of the inclusion of this variable in the STIRPAT model instead of CO_2 intensity (the fourth variable from Section 5.1).

Finally, the third empirical analysis (Section 5.3) uses the traditional STIRPAT model to show the reader the advantage of using residualized variables. In this case, there are no strong essential collinearity problems, so the methodology is applied for empirical purposes. By doing so, it is clear that the application of residualization leads to conclusions about the model that differ from the original even though both models (the original and the residualized) have several identical characteristics. The residualized model can answer questions that could not be answered with the initial model, apart from reducing the degree of collinearity, but it has to be taken into account that residualization is not always applicable because the interpretations of the new estimated coefficients are not always simple.

Chapter 6

Conclusions

6.1 Discussion and global conclusions

The dissertation presented here aims to clarify the role of multicollinearity in an econometric model and proposes residualization as a good methodology not only to deal with the problem but also to achieve another type of interpretation of the explanatory variables from the model under consideration.

Chapter 1 gave the reader an initial introduction to the problem and offered a brief explanation of the methodology being presented. Chapter 2 then looked in depth at the multicollinearity problem and the traditional methodologies used in this field. Chapters 3 and 4 explain the methodology further; Chapter 3 focuses its attention on earlier works in the field: criticism of the method and methodological preliminaries, while Chapter 4 presents the generalization of the method, together with the justification and properties of the residualization procedure. These two chapters and in particular Chapter 4, are the main contribution of this Thesis. Finally, Chapter 5 presents the empirical part: three specific models on environmental economics, which present strong collinearity

problems.

As stated in the Introduction (Chapter 1), the main goal in an econometric model is to estimate the parameters which accompany the explanatory variables. When any relationship exists between explanatory variables, it may be said that it exists collinearity or multicollinearity in the model. In general, it could be said that multicollinearity represents a big problem when the main goal of the researcher is to study the impacts of some group of explanatory variables on the selected independent variable but, if the goal is simply to predict the explained variable from a set of variables, then it is not significant. Of course, it is important to remark that the importance of multicollinearity depends on the specific model and the specific study the researcher wants to perform, and the proper identification of the problem is a very important initial step.

Chapter 2 of this dissertation further analysed the problem of multicollinearity: concept and types of collinearity, causes and consequences of this, detection of the problem and, finally, traditional solutions to collinearity.

It has been pointed out that collinearity always exists in an econometric model because the researcher is modelling a reality in which, generally, some type of relationship always appears, so the analyst in practice always has some degree of collinearity. The controversial issue is to detect whether this fact represents a real problem or whether it does not affect the research.

Two principal types of multicollinearity have been distinguished: perfect and near, claiming that near or imperfect collinearity is the most complex and difficult to manage because it allows the researcher to estimate the model, in contrast to perfect multicollinearity, but it leads to unstable estimations. In addition, near multicollinearity, regarding the relationships among explanatory variables, may be split into two types: non-essential and essential collinearity.

The principal consequences of near multicollinearity are: inflated variances of the estimators, tendency to consider estimated parameters as non-significant, high R^2 and high sensitivity of the estimations. Basically, these characteristics mean the researcher cannot separate the individual effects of the independent variables and the results are distorted. Thus, it is very important to detect the existing multicollinearity, and to check what type of collinearity appears in the study in question in order to apply the best solution. Section 2.2 of Chapter 2 outlined some methods to check the existence of the problem, such as variance inflation factor (VIF), among others. Once it is verified there is strong collinearity, the researcher must make decisions about the path to take. To sum up, the analyst may delete some variables if the study allows, may create new circumstances in which collinearity is mitigated or may use alternative methodologies that allow the problem to be mitigated. This last path is the one that concerns us in this dissertation. Some methods have been explained in Section 2.3, such as ridge regression or raise regression. The most commonly-used one is the well-known ridge regression, but this methodology presents some deficiencies, as has been pointed out throughout this Thesis.

On the other hand, it was anticipated at the outset that the main goal of the dissertation is to present residualization as an alternative to the traditional methodologies in dealing with multicollinearity. This methodology has been previously applied in different fields, but has not been developed explicitly. The lack of specification and the consequent misunderstanding have led to some criticism of it in the literature, as was commented in Chapter 1 and clarified in Chapter 3. Residualization has been explained throughout this Thesis: its antecedents and the method, properties and application (see Chapters 3 to 5). With the principal objective of mitigating collinearity, it has been

demonstrated that with residualization it is possible not only to alleviate the problem but also to obtain different interpretations of the modified variables, so this procedure allows the researcher to apply the method with different purposes: to answer questions regarding the interpretation of the coefficients that cannot be performed by the original model. In the generalization of the procedure given in Chapter 4, the properties of the methodology were studied further, leading to the following conclusions (see Section 4.5):

- Estimations and inference:
  - The coefficient of the residualized variable does not change, but the interpretation of the variable does: it will represent the part of the original variable that has no relationship with the rest of explanatory variables of the model (the principle *ceteris paribus* is strictly fulfilled) if the rest of the independent variables are included in the auxiliary regression.
  - The inference related to the individual significance of the residualized variable is still the same.
  - The coefficients of the non-residualized variables change, however the interpretations are still the same.
  - The inference related to the individual significance of the non-residualized variables is different.
  - The value of the estimated parameters of the non-residualized variables are the same as in the model which does not include the modified variable.
  - If only some of the independent variables are included in the auxiliary regression, then the estimations of the parameters of the explanatory

variables not included in it also remain unchanged.

- Global properties:
  - The sum of square residuals of the original model and the residualized model are the same.
  - The estimate of the variance of the random disturbance does not change.
  - The coefficient of determination, R^2, is still the same.
  - The global significance test remains unchanged.
  - The original model and the residualized model provide the same prediction.

In light of the foregoing, it can be concluded that residualization leads to conclusions about a model which is different to the original even though both have several identical characteristics. It means, as a whole, the researcher is estimating the initial model but deleting redundant information regarding to the set of explanatory variables; looking at it in detail, it has some new variables (the residualized variables) that have a different interpretation. Furthermore, collinearity problems are mitigated and, in the majority of the cases, the application of residualization leads to better results in terms of individual significance and consistency with theory (the signs of the estimated parameters). These facts are shown in the empirical part (Chapter 5), specifically, in Section 5.1, which compares residualization with three additional methodologies: ridge regression, LASSO regression and raise regression. Section 5.2 mainly uses residualization to mitigate collinearity problems, but the residualization procedure is applied in three different ways to show the reader the possibilities of the methodology: using all the explanatory variables, using only one of

the explanatory variables and using an external variable. Finally, Section 5.3 applies the methodology to show the reader its use with empirical purposes, i.e. the application of the method for obtaining new interpretations of the variables.

All the examples used in Chapter 5 are based on a traditional environmental model: the STIRPAT model, with a renewed version presented in Section 5.2: the STIRPATE. It is well-known that in social sciences relationships among explanatory variables are always present and, in environmental studies, this fact has important implications as the research is used in policymaking. This chapter allows the reader to reaffirm the idea about residualization: it leads to good properties, characteristics and consistency with theory and expectations about the results.

In conclusion, it is evident that strong collinearity is a real problem that has to be mitigated. In this field, new methods are arising, as residualization, with better properties than other traditional methodologies, such as ridge regression. In social sciences, traditional variables are closely related to each other and this fact has to be taken into account for future research in any type of field.

6.2 General implications

For the analysis of causal effects, residualization has been proposed as a good methodology to deal with multicollinearity in a specific econometric model. It maintains the initial global properties while mitigating strong collinearity problems, to analyse the causal effects of the driving forces affecting collinearity and to study different interpretations of the explanatory variables.

As was noted, the method has been meaningfully studied: the properties, characteristics and uses of this have been further developed and explained. The method has also been compared with other traditional methodologies, reaching conclusions about important implications regarding the use and advantages of the method: throughout this Thesis, the method is presented as a very good alternative to traditional methodologies such as ridge regression.

The methodology has been applied to actual data to show the reader the real implications and useful properties of the same. The implementation of the best methodology is crucial to observe and make applicable conclusions of the results obtained. As explained in the Introduction (Chapter 1) of this dissertation, taking into account the residualization procedure and its properties, it is clear that this method allows the researcher to deal with multicollinearity problems and, furthermore, it also introduces another interpretation of the (modified) variables.

In conclusion, even when the goal of the study is to predict (where it has been concluded previously that collinearity is not significant), it is highly recommended to mitigate the problem because the researcher needs to be very sure of the continuity of the relationships between explanatory variables in the future because, if the relationship changes, the forecast based on the initial model may be unreliable as well.

6.3 Limitations and future lines of research

Residualization has been presented as a good alternative in dealing with collinearity problems, but it has a noteworthy limitation. The principal weakness of residualization is, in turn, its main strength: as it has been pointed out throughout this Thesis, the new interpretation of the modified

variables is crucial for implementing the method, but not all variables could be modified. The choice of the appropriate variable is the more difficult part of the application of the method because if the researcher selects a variable whose isolated part makes no sense from an interpretative point of view, the study will not be able to reach any successful conclusions.

On the other hand, the empirical examples used have essentially been environmental models. As the title of this Thesis anticipates, one important point here is the use of ecological data. In future research, it would be interesting to apply the method to a different range of disciplines, in order to show the reader the broad applicability of the method. As an empirical issue to continue, regarding the STIRPAT or STIRPATE, Kilbourne and Thyroff (2020) suggest that future research may focus on the education in the region on the effect on the environment. Therefore, an interesting field to explore might be an educational version of the STIRPAT or the STIRPATE, by including new variables such as actitudes of the population towards pollution based on their educational background. Furthermore, the same authors stated that *future research may also consider including country typologies as a variable in the model*. Finally, as has been suggested in Section 5.1, an interesting question for future research could be to undertake an in-depth exploration of the level of technology: does it play a key role in determining the influence of the rest of the variables in environmental models? A starting point in this field is the work by García et al. (2019a), which studies the relationship between technological readiness and environmental efficiency in the EU-28 context.

Furthermore, as was noted in Section 5.1.1, regarding LASSO regression it is not possible to corroborate directly whether the potential collinearity has been mitigated in the model under consideration. With regard to check the potential collinearity that may exist after applying any type of methodology, it

would be interesting to develop more tools to measure and detect this problem.

Finally, as an important methodological line to continue working on is the application of the method to another type of explanatory variable, with more complex interpretation, such as dummy variables or interactions between variables. For example, with the use of Moderated Regression Analysis (MRA), which analyses how the effect of one of the explanatory variables is moderated by another independent variable by adding an interaction term between these two variables, the researcher may introduce strong collinearity in the model. Although it has been briefly studied (see García et al. (2016a) and its references), this "artificial" introduction of strong collinearity problems is interesting for in-depth research. Furthermore, it would be interesting to further study any possible link between residualization and other methodologies, such as raise regression.

Appendix A

Notes to Chapter 3.

A.1 Measurement of $\mathbf{e}_4^t\mathbf{Y}$ and $\mathbf{e}_4^t\mathbf{e}_4$

First, starting from expression (3.5), it is clear that:

$$\begin{aligned}\mathbf{e}_4^t\mathbf{Y} &= (\mathbf{x}_4 - \widehat{\mathbf{x}}_4)^t\mathbf{Y} = (\mathbf{x}_4 - \widehat{\alpha}_2\mathbf{x}_2 - \widehat{\alpha}_3\mathbf{x}_3)^t\mathbf{Y} = \mathbf{x}_4^t\mathbf{Y} - \widehat{\alpha}_2\mathbf{x}_2^t\mathbf{Y} - \widehat{\alpha}_3\mathbf{x}_3^t\mathbf{Y} \\ &= \varrho_4 - \frac{\varrho_2 - \rho_{23}\varrho_3}{1-\rho_{23}^2}\varrho_2 - \frac{\varrho_3 - \rho_{23}\varrho_2}{1-\rho_{23}^2}\varrho_3.\end{aligned}$$

On the other hand:

$$\begin{aligned}\mathbf{e}_4^t\mathbf{e}_4 &= (\mathbf{x}_4 - \widehat{\mathbf{x}}_4)^t(\mathbf{x}_4 - \widehat{\mathbf{x}}_4) = 1 - 2\widehat{\alpha}_2\rho_{24} - 2\widehat{\alpha}_3\rho_{34} + 2\widehat{\alpha}_2\widehat{\alpha}_3\rho_{23} + \widehat{\alpha}_2^2 + \widehat{\alpha}_3^2 \\ &= 1 - \frac{2\rho_{24}^2 - 2\rho_{23}\rho_{24}\rho_{34}}{1-\rho_{23}^2} - \frac{2\rho_{34}^2 - 2\rho_{23}\rho_{24}\rho_{34}}{1-\rho_{23}^2} \\ &\quad + \frac{2\rho_{24}\rho_{34} - 2\rho_{23}\rho_{24}^2 - 2\rho_{23}\rho_{34}^2 + 2\rho_{23}^2\rho_{24}\rho_{34}}{(1-\rho_{23}^2)^2} \\ &\quad + \frac{\rho_{24}^2 + \rho_{23}^2\rho_{34}^2 - 2\rho_{23}\rho_{24}\rho_{34} + \rho_{34}^2 + \rho_{23}^2\rho_{24}^2 - 2\rho_{23}\rho_{24}\rho_{34}}{(1-\rho_{23}^2)^2} \\ &= \frac{1 + \rho_{23}^4 - 2\rho_{23}^2 - \rho_{24}^2 - \rho_{34}^2 + \rho_{23}^2\rho_{24}^2 + \rho_{23}^2\rho_{34}^2 + 2\rho_{23}\rho_{24}\rho_{34} - 2\rho_{23}^3\rho_{24}\rho_{34}}{(1-\rho_{23}^2)^2} \\ &= \frac{\left(1 + 2\rho_{23}\rho_{24}\rho_{34} - \rho_{23}^2 - \rho_{24}^2 - \rho_{34}^2\right)\left(1-\rho_{23}^2\right)}{(1-\rho_{23}^2)^2} \\ &= \frac{1 + 2\rho_{23}\rho_{24}\rho_{34} - \rho_{23}^2 - \rho_{24}^2 - \rho_{34}^2}{1-\rho_{23}^2}.\end{aligned}$$

A.2 Demonstration that the residual sum of squares coincides in models (3.2) and (3.6)

Starting from model (3.2), given (3.3) and (3.4), it is clear that:

$$\begin{aligned}\text{SCE} = &\frac{(1-\rho_{34}^2)\varrho_2^2+(1-\rho_{24}^2)\varrho_3^2+(1-\rho_{23}^2)\varrho_4^2-2(\rho_{23}-\rho_{24}\rho_{34})\varrho_2\varrho_3}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2}\\ &+\frac{2(\rho_{24}-\rho_{23}\rho_{34})\varrho_2\varrho_4-2(\rho_{34}-\rho_{23}\rho_{24})\varrho_3\varrho_4}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2}.\end{aligned} \tag{A.1}$$

On the other hand, given (3.7) and (3.8), in model (3.6) it is verified that:

$$\begin{aligned}\text{SSE}_O = &\frac{\varrho_2-\rho_{23}\varrho_3}{1-\rho_{23}^2}\varrho_2+\frac{\varrho_3-\rho_{23}\varrho_2}{1-\rho_{23}^2}\varrho_3+\frac{(\mathbf{e}_4^t\mathbf{Y})^2}{\mathbf{e}_4^t\mathbf{e}_4}\\ =&\frac{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(\varrho_2^2-2\rho_{23}\varrho_2\varrho_3+\varrho_3^2)}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)}\\ &+\frac{(1-\rho_{23}^2)\varrho_4^2+(\rho_{24}-\rho_{23}\rho_{34})^2\varrho_2^2+(\rho_{34}-\rho_{23}\rho_{24})^2\varrho_3^2}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)}\\ &+\frac{2(\rho_{34}-\rho_{23}\rho_{24})(\rho_{24}-\rho_{23}\rho_{34})\varrho_2\varrho_3}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)}\\ &-\frac{2(1-\rho_{23}^2)(\rho_{24}-\rho_{23}\rho_{34})\varrho_2\varrho_4+2(1-\rho_{23}^2)(\rho_{34}-\rho_{23}\rho_{24})\varrho_3\varrho_4}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)},\end{aligned} \tag{A.2}$$

where it has been taken into account that:

$$\frac{(\mathbf{e}_4^t\mathbf{Y})^2}{\mathbf{e}_4^t\mathbf{e}_4}=\frac{\left((1-\rho_{23}^2)\varrho_4-(\rho_{24}-\rho_{23}\rho_{34})\varrho_2-(\rho_{34}-\rho_{23}\rho_{24})\varrho_3\right)^2}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)}.$$

Combining and expanding the first two terms of (A.2), the result is the following:

$$\begin{aligned}\text{SSE}_O = &\frac{\varrho_2^2(1-\rho_{23}^2-\rho_{34}^2+\rho_{23}^2\rho_{34}^2)+\varrho_3^2(1-\rho_{23}^2-\rho_{24}^2+\rho_{23}^2\rho_{24}^2)}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)}\\ &+\frac{2\varrho_2\varrho_3(-\rho_{23}-\rho_{23}^2\rho_{24}\rho_{34}+\rho_{23}^3+\rho_{34}\rho_{24})}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)}\\ &+\frac{(1-\rho_{23}^2)\varrho_4^2-2(\rho_{24}-\rho_{23}\rho_{34})\varrho_2\varrho_4-2(\rho_{34}-\rho_{23}\rho_{24})\varrho_3\varrho_4}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2}\\ =&\frac{\varrho_2^2(1-\rho_{23}^2)(1-\rho_{34}^2)+\varrho_3^2(1-\rho_{23}^2)(1-\rho_{24}^2)+2\varrho_2\varrho_3(1-\rho_{23}^2)(\rho_{23}-\rho_{24}\rho_{34})}{(1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2)(1-\rho_{23}^2)}\\ &+\frac{(1-\rho_{23}^2)\varrho_4^2-2(\rho_{24}-\rho_{23}\rho_{34})\varrho_2\varrho_4-2(\rho_{34}-\rho_{23}\rho_{24})\varrho_3\varrho_4}{1+2\rho_{23}\rho_{24}\rho_{34}-\rho_{23}^2-\rho_{24}^2-\rho_{34}^2}.\end{aligned} \tag{A.3}$$

As the reader can appreciate, expressions (A.1) and (A.3) coincide, thus $\text{SCE}=\text{SCE}_O$.

Appendix B

Notes to Chapter 5.

B.1 Notation

The following notation is used to present the variance inflator factor (VIF) and the mean square error (MSE) of the different methodologies used in Section 5.1.

VIF_i: Variance inflation factor of each independent variable i.

R_i^2: Coefficient of determination of the ordinary least squares (OLS) regressions of each independent variable i on the rest of the explanatory variables of the model (OLS and residualization).

R_i^2 (λ): Coefficient of determination of the OLS regressions of each independent variable i on the rest of the explanatory variables of the model (using the raised variable $\widetilde{\mathbf{A}}$).

R_i^2 (k): Coefficient of determination of the ridge regressions of each independent variable i on the rest of the explanatory variables of the model. This coefficient is calculated following García et al. (2016b).

MSE ($\widehat{\boldsymbol{\beta}}$): Mean square error of the OLS regression.

MSE ($\widehat{\boldsymbol{\beta}}^{O}$): Mean square error of the residualization.

MSE $\left(\widehat{\boldsymbol{\beta}}\ (\lambda)\right)$: Mean square error of the raise regression.

MSE $\left(\widehat{\boldsymbol{\beta}}\ (k)\right)$: Mean square error of the ridge regression.

B.2 VIF and MSE for different methodologies

Table B.1 presents the corresponding expressions of VIFs of OLS, residualization, raise and ridge regressions, and Table B.2 presents the corresponding MSEs.

Table B.1: Detection of collinearity: variance inflation factor (VIF).

Method	$\text{VIF}_{\mathbf{P}}$	$\text{VIF}_{\mathbf{A}}$	$\text{VIF}_{\mathbf{T_1}}$	$\text{VIF}_{\mathbf{T_2}}$
OLS	$\frac{1}{1-R^2_{\mathbf{P}}}$	$\frac{1}{1-R^2_{\mathbf{A}}}$	$\frac{1}{1-R^2_{\mathbf{T_1}}}$	$\frac{1}{1-R^2_{\mathbf{T_2}}}$
Residualization	$\frac{1}{1-R^2_{\mathbf{P}}}$	$\frac{1}{1-R^2_{\mathbf{e_A}}}$	$\frac{1}{1-R^2_{\mathbf{T_1}}}$	$\frac{1}{1-R^2_{\mathbf{T_2}}}$
Raise regression	$\frac{1}{1-R^2_{\mathbf{P}}(\lambda)}$	$\frac{1}{1-R^2_{\mathbf{A}}(\lambda)}$	$\frac{1}{1-R^2_{\mathbf{T_1}}(\lambda)}$	$\frac{1}{1-R^2_{\mathbf{T_2}}(\lambda)}$
Ridge regression	$\frac{1}{1-R^2_{\mathbf{P}}(k)}$	$\frac{1}{1-R^2_{\mathbf{A}}(k)}$	$\frac{1}{1-R^2_{\mathbf{T_1}}(k)}$	$\frac{1}{1-R^2_{\mathbf{T_2}}(k)}$

Table B.2: Mean square error (MSE).

Method	MSE
OLS	$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \sigma^2 tr((\mathbf{X}^t\mathbf{X})^{-1})$.
Raise regression	$\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right) = \sigma^2 tr((\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}})^{-1}) + \boldsymbol{\beta}^t(M_\lambda^{-1} - \mathbf{I})^t(M_\lambda^{-1} - \mathbf{I})\boldsymbol{\beta}$, where $\widetilde{\mathbf{X}} = \mathbf{X} \cdot M_\lambda$ and $M_\lambda = \begin{pmatrix} 1 & 0 & -\lambda\widehat{\alpha}_0 & 0 & 0 \\ 0 & 1 & -\lambda\widehat{\alpha}_1 & 0 & 0 \\ 0 & 0 & (1+\lambda) & 0 & 0 \\ 0 & 0 & -\lambda\widehat{\alpha}_2 & 1 & 0 \\ 0 & 0 & -\lambda\widehat{\alpha}_3 & 0 & 1 \end{pmatrix}$.
Residualization	$\text{MSE}(\widehat{\boldsymbol{\beta}}^O) = \sigma^2 tr((\mathbf{X}_e^t\mathbf{X}_e)^{-1}) + \boldsymbol{\beta}^t(N^{-1} - \mathbf{I})^t(N^{-1} - \mathbf{I})\boldsymbol{\beta}$, where $\mathbf{X_e} = \mathbf{X} \cdot N$ and $N = \begin{pmatrix} 1 & 0 & -\widehat{\alpha}_0 & 0 & 0 \\ 0 & 1 & -\widehat{\alpha}_1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\widehat{\alpha}_2 & 1 & 0 \\ 0 & 0 & -\widehat{\alpha}_3 & 0 & 1 \end{pmatrix}$.
Ridge Regression	$\text{MSE}\left(\widehat{\boldsymbol{\beta}}(k)\right) = \sigma^2 \sum_{i=1}^{p} \frac{\mu_i}{(\mu_i+k)^2} + \beta^t(Z_k - \mathbf{I})^t(Z_k - \mathbf{I})\boldsymbol{\beta}$, where $Z_k = (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}$ and μ_i are the eigenvalues of matrix $\mathbf{X}^t\mathbf{X}$.

Bibliography

M. Alauddin and H.S. Nghiem. Do instructional attributes pose multicollinearity problems? An empirical exploration. *Economic Analysis and Policy*, 40(3):351–361, 2010.

V. Alcántara and E. Padilla. Análisis de las emisiones de CO_2 y sus factores explicativos en las diferentes áreas del mundo. *Revista de economía crítica*, 4:17–37, 2005.

A. Alin. Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):370–374, 2010.

B. Ambridge, J.M. Pine, and C.F. Rowland. Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition*, 123(2):260–279, 2012.

N. Apergis and C. García. Environmentalism in the UE-28 context: The impact of governance quality on environmental energy efficiency. *Environmental Science and Pollution Research*, 26:37012–37025, 2019. DOI: `https://doi.org/10.1007/s11356-019-06600-1`.

H. Artigue and G. Smith. The principal problem with principal components regression. *Cogent Mathematics & Statistics*, 6:1622190, 2019.

M. Azam and A.Q. Khan. Testing the environmental kuznets curve hypothesis: A comparative empirical study for low, lower middle, upper middle and high income countries. *Renewable and Sustainable Energy Reviews*, 63:556–567, 2016.

G.L. Baird and S.L. Bieber. The Goldilocks Dilemma: Impacts of Multicollinearity. A Comparison of Simple Linear Regression, Multiple Regression, and Ordered Variable Regression Models. *Journal of Modern Applied Statistical Methods*, 15(1):18, 2016.

B.M. Balk, M.B.M. De Koster, C. Kaps, and J.L. Zofío. An Evaluation of Cross-Efficiency Methods, Applied to Measuring Warehouse Performance. Working paper (draft, December 2017), 2017.

N. Bandelj and M.C. Mahutga. How socio-economic change shapes income inequality in post-socialist Europe. *Social Forces*, 88(5):2133–2161, 2010.

D.A. Belsley. *Conditioning diagnostics: Collinearity and weak data in regression.* John Wiley, New York, 1991.

D.A. Belsley. Conditioning diagnostics. *Encyclopedia of Statistical Sciences*, 2, 2004.

D.A. Belsley and V.C. Klema. Detecting and assessing the problems caused by multicollinearity: A use of the singular-value decomposition. NBER Working Paper Series, Working Paper No. 66, 1974.

D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity.* John Wiley & Sons, New York, 1980.

A. Bengochea-Morancho, F. Higón-Tamarit, and I. Martínez-Zarzoso. Economic Growth and CO_2 Emissions in the European Union. *Environmental and Resource Economics*, 19(2):165–172, 2001.

A. Bitetto, A. Mangone, R.M. Mininni, and L.C. Giannossa. A nonlinear principal component analysis to study archeometric data. *Journal of Chemometrics*, 30(7):405–415, 2016.

E. Boserup. Population and technological change: A study of long-term trends. University of Chicago Press Chicago Ill. United States, 1981.

Y.W. Bradshaw. Urbanization and underdevelopment: A global study of modernization, urban bias, and economic dependency. *American Sociological Review*, 52(2):224–239, 1987.

A. Bruvoll and H. Medin. Factors Behind the Environmental Kuznets Curve: A Decomposition of the Changes in Air Pollution. *Environmental and Resource Economics*, 24(1):27–48, 2003.

M. Büchs and S.V. Schnepf. Who emits most? Associations between socio-economic factors and UK household's home energy, transport, indirect and total CO_2 emissions. *Ecological Economics*, 90(1):114–123, 2013.

A. Buse. Brickmaking and the collinear arts: a cautionary tale. *Canadian Journal of Economics*, pages 408–414, 1994.

M.R. Butler and E.M. McNertney. Estimating educational production functions: The problem of multicollinearity. *The Social Science Journal*, 28(4):489–499, 1991.

L. Cecchini, S. Venanzi, A. Pierri, and M. Chiorri. Environmental efficiency analysis and estimation of CO_2 abatement costs in dairy cattle farms in

Umbria (Italy): a SBM-DEA model with undesirable output. *Journal of Cleaner Production*, 197:895–907, 2018.

A. Charnes, W.W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444, 1978.

C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 158(3): 419–444, 1995.

S. Chatterjee and A.S. Hadi. *Sensitivity anallysis in linear regression*. John Wiley, New York, 1988.

P. Chennamaneni, R. Echambadi, J.D. Hess, and N. Syam. How Do You Properly Diagnose Harmful Collinearity in Moderated Regressions? *Retrieved June*, 1(2011), 2011.

M.R. Chertow. The IPAT equation and its variants. *Journal of Industrial Ecology*, 4(4):13–29, 2000.

I.G. Chong and C.H. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1):103–112, 2005.

J. Chontanawat. Driving forces of energy-related CO2 emissions based on expanded IPAT decomposition analysis: evidence from ASEAN and four selected countries. *Energies*, 12(4):764, 2019.

A.M. Cohen-Goldberg. Phonological competition within the word: Evidence from the phoneme similarity effect in spoken production. *Journal of Memory and Language*, 67(1):184–198, 2012.

B. Commoner, M. Corr, and P.J. Stamler. The causes of pollution. *Environment: Science and Policy for Sustainable Development*, 13(3):2–19, 1971.

D. Coondoo and S. Dinda. Causality between income and emission: a country group-specific econometric analysis. *Ecological Economics*, 40(3):351–367, 2002.

J.D. Curto and J.C. Pinto. New Multicollinearity Indicators in Linear Regression Models. *International Statistical Review*, 75(1):114–121, 2007.

J.D. Curto and J.C. Pinto. The corrected VIF (CVIF). *Journal of Applied Statistics*, 38(7):1499–1507, 2011.

D.K. Dalal and M.J. Zickar. Some Common Myths About Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression. *Organizational Research Methods*, 15(3):339–362, 2012.

S.M. De Bruyn, J.C.J.M. Van Den Bergh, and J.B. Opschoor. Economic growth and emissions: reconsidering the empirical basis of environmental Kuznets curves. *Ecological Economics*, 25(2):161–175, 1998.

J. Deegan Jr. The process of political development: An illustrative use of a strategy for regression in the presence of multicollinearity. *Sociological Methods & Research*, 3(4):384–415, 1975.

X. Deng, X. Tian, and S. Chen. Modified kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis. *Chemometrics and Intelligent Laboratory Systems*, 127: 195–209, 2013.

T. Dietz and E.A. Rosa. Rethinking the environmental impacts of population, affluence and technology. *Human Ecology Review*, 1:277–300, 1994.

T. Dietz and E.A. Rosa. Effects of population and affluence on CO_2 emissions. *Proceedings of the National Academy of Sciences of the USA*, 94(1):175–179, 1997.

J. Dong, C. Deng, R. Li, and J. Huang. Moving Low-Carbon Transportation in Xinjiang: Evidence from STIRPAT and Rigid Regression Models. *Sustainability*, 9(1):24, 2016.

C.F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J.R. García, B. Gruber, B. Lafourcade, P.J. Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 1981.

P.R. Ehrlich and J.P. Holdren. The people problem. *Saturday Reveiw*, 4(42): 42–43, 1970.

P.R. Ehrlich and J.P. Holdren. The Impact of Population Growth. *Science*, 171(3977):1212–1217, 1971.

P.R. Ehrlich and J.P. Holdren. A bulletin dialogue on the "Closing Circle": Critique: One dimensional ecology. *Bulletin of the Atomic Scientists*, 28(5): 16–27, 1972.

M.Z. Fabrycy. Economic Theory and Multicollinearity. *The American Economist*, 17(1):81–84, 1973.

Y. Fan, L.C. Liu, and Y.M. Wei. Analyzing impact factors of CO_2 emissions using the STIRPPAT model. *Environmental Impact Assessment Review*, 26 (4):377–395, 2006.

R. Färe, S. Grosskopf, C.A.K. Lovell, and C. Pasurka. Multilateral Productivity Comparisons When Some Outputs are Undesirable: A Nonparametric Approach. *The Review of Economics and Statistics*, 71(1):90–98, 1989.

D.E. Farrar and R.R. Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

Y. Fernández, M.A. Fernández, D. González, and B. Olmedillas. El efecto regulador de los Planes Nacionales de Asignación sobre las emisiones de CO_2. *Revista de Economía Mundial*, 40:47–66, 2015.

J. Fox. *Linear statistical models and related methods with applications to social research.* John Wiley, New York, 1984.

L. Friedman and M. Wall. Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2):127–136, 2005.

C. García, C. B. García, R. Salmerón, and J. García. Regresión con variables ortogonales y regresión alzada en el modelo STIRPAT. *Estudios de Economía Aplicada*, 35(3):717–734, 2017a. URL: `https://bit.ly/36GkqWV`.

C. García, C. B. García, and R. Salmerón. Environmental Efficiency and Technological Readiness. An evidence from EU-28. *Estudios de Economía Aplicada*, 37(1):24–34, 2019a. URL: `https://bit.ly/2zLBsXx`.

C. García, R. Salmerón, and C. B. García. Choice of the ridge factor from the correlation matrix determinant. *Journal of Statistical Computation and Simulation*, 89(2):211–231, 2019b. DOI: `https://doi.org/10.1080/00949655.2018.1543423`.

C. García, C. B. García, and R. Salmerón. Confronting collinearity in environmental regression models: an evidence from world data. *Statistical Methods and Applications*, Accepted, 2020.

C. B. García, J. García, and J. Soto. The raise method: An alternative procedure to estimate the parameters in presence of collinearity. *Quality & Quantity*, 45(2):403–423, 2011.

C. B. García, J. García, M. M. López-Martín, and R. Salmerón. Collinearity: Revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*, 42(3):648–661, 2015.

C. B. García, R. Salmerón, , C. García, and J. García. Raise regression mitigating collinearity in moderated regression. In *The 5th Advanced Research in Scientific Areas*, 2016a.

C. B. García, R. Salmerón, J. García, and M. M. López-Martín. On the selection of the ridge and raise factors. *Indian Journal of Science and Technology*, 10(13):1–8, 2017b.

C. B. García, R. Salmerón, C. García, and J. García. Residualization: justification, properties and application. *Journal of Applied Statistics*, 2019c. DOI: `https://doi.org/10.1080/02664763.2019.1701638`.

J. García and D.E. Ramírez. The successive raising estimator and its relation with the ridge estimator. *Communications in Statistics-Theory and Methods*, 46(22):11123–11142, 2017.

J. García, R. Salmerón, M. M. López-Martín, and C. B. García. Revisiting the Condition Number and Red indicator in Ridge Regression. In *Proceedings of the International Work Conference on Time Series*, 2015.

J. García, R. Salmerón, C. B. García, and M. M. López-Martín. Standardization of Variables and Collinearity Diagnostic in Ridge Regression. *International Statistical Review*, 84(2):245–266, 2016b.

M. Gassebner, M.J. Lamla, and J.E. Sturm. Determinants of pollution: what do we really know? *Oxford Economic Papers*, 63(3):568–595, 2011.

P. Geladi and B.R. Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

M. Giacalone, D. Panarello, and R. Mattera. Multicollinearity in regression: an efficiency comparison between L p-norm and least squares estimators. *Quality & Quantity*, 52(4):1831–1859, 2018.

Y. Gimenez and G. Giussani. Searching for the core variables in principal components analysis. *Brazilian Journal of Probability and Statistics*, 32(4): 730–754, 2018.

M.H. Graham. Confronting multicollinearity in ecological multiple regression. *Ecology*, 84(11):2809–2815, 2003.

R. Grewal, J.A. Cote, and H. Baumgartner. Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science*, 23(4):519–529, 2004.

D. Gujarati. *Basic Econometrics*. McGraw Hill, 4th edition, 2003.

D.N. Gujarati. *Basic Econometrics*. McGraw-Hill, 4th edition, 2004.

D.N. Gujarati. *Essentials of Econometrics*. McGraw-Hill, 4th edition, 2010.

N. Gürer and J. Ban. Factors Affecting Energy-related CO_2 Emissions: Past Levels and Present Trends. *OPEC Energy Review*, 21(4):309–350, 1997.

J.F.J. Hair, R.E. Anderson, R.L. Tatham, and W.C. Black. *Multivariate Data Analysis*. MacMillan, New York, NE (USA), third edition, 1995.

Y. Haitovsky. Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics*, 51(4):486–489, 1969.

C. Hamilton and H. Turton. Determinants of emissions growth in OECD countries. *Energy Policy*, 30(1):63–71, 2002.

C. Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

W.T. Harbaugh, A. Levinson, and D.M. Wilson. Reexamining the Empirical Evidence for an Environmental Kuznets Curve. *The Review of Economics and Statistics*, 84(3):541–551, 2002.

R. Hashmi and K. Alam. Dynamic relationship among environmental regulation, innovation, CO2 emissions, population, and economic growth in OECD countries: A panel investigation. *Journal of Cleaner Production*, 231, 2019.

D.M. Hawkins. On the investigation of alternative regressions by principal component analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(3):275–286, 1973.

R.C. Hill and L.C. Adkins. *Collinearity. In: A Companion to Theoretical Econometrics (ed. Badi H. Baltagi)*. Blackwell Publishing Ltd, Malden, MA (USA), 2003.

A.E. Hoerl and R.W. Kennard. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82, 1970a.

A.E. Hoerl and R.W. Kennard. Ridge Regression: Biased Estimation for Northogonal Problems. *Technometrics*, 12(1):55–67, 1970b.

A.E. Hoerl, R.W. Kennard, and K.F. Baldwin. Ridge Regression: Some simulation. *Communications in Statistics-Theory and Methods*, 4(2):105–123, 1975.

L.M. Holland. *Evaluation of estimators for ill-posed statistical problems subject to multicollinearity*. PhD thesis, University of Waikato, New Zealand, 2014.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

D.S. Huang. *Regression and econometric methods*. John Wiley & Sons, New York, 1970.

D. Iacobucci, M.J. Schneider, D.L. Popovich, and G.A. Bakamitsos. Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behavior research methods*, 48(4):1308–1317, 2016.

M.U. Imdad and M. Aslam. *lmridge: Linear Ridge Regression with Ridge Penalty and Ridge Statistics*, 2018. URL: `https://CRAN.R-project.org/package=lmridge`. R package version 1.2.

T.F. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62, 2010.

D.R. Jensen and D.E. Ramírez. Anomalies in the Foundations of Ridge Regression. *International Statistical Review*, 76(1):89–105, 2008.

J. Jia, H. Deng, J. Duan, and J. Zhao. Analysis of the major drivers of the ecological footprint using the STIRPAT model and the PLS method. A case study in Henan Province, China. *Ecological Economics*, 68(11):2818–2824, 2009.

J. Johnston. *Econometric Methods*. McGraw-Hill, New York, 2nd edition, 1972.

J. Johnston and J. Dinardo. *Métodos de Econometría*. Vicens-Vives, Barcelona, 2001.

A.K. Jorgenson. Global warming and the neglected greenhouse gas: A cross-national study of the social causes of methane emissions intensity, 1995. *Social Forces*, 84(3):1779–1798, 2006.

A.K. Jorgenson and T.J. Burns. The political-economic causes of change in the ecological footprints of nations, 1991–2001: a quantitative investigation. *Social Science Research*, 36(2):834–853, 2007.

A.K. Jorgenson and B. Clark. The economy, military, and ecologically unequal exchange relationships in comparative perspective: a panel study of the ecological footprints of nations, 1975-2000. *Social Problems*, 56(4):621–646, 2009.

P.E. Kennedy. Eliminating problems caused by multicollinearity: A warning. *The Journal of Economic Education*, 13(1):62–64, 1982.

P.E. Kennedy. *A guide to Econometrics*. MIT Press, 3rd edition, 1992.

J. Kentor and E. Kick. Bringing the military back in: Military expenditures and economic growth 1990 to 2003. *Journal of World-Systems Research*, 14 (2):142–172, 2008.

S.A.R. Khan, K. Zaman, and Y. Zhang. The relationship between energy-resource depletion, climate change, health resources and the environmental Kuznets curve: Evidence from the panel of selected developed countries. *Renewable and Sustainable Energy Reviews*, 62:468–477, 2016.

B.M.G. Kibria and S. Banik. Some Ridge Regression Estimators and Their Performances. *Journal of Modern Applied Statistical Methods*, 15(1):206–238, 2016.

J.S. Kidwell and L.H. Brown. Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and the Family*, 44(2): 287–299, 1982.

H.A. Kiers and A.K. Smilde. A comparison of various methods for multivariate regression with highly collinear variables. *Statistical Methods and Applications*, 16(2):193–228, 2007.

W.E. Kilbourne and A. Thyroff. STIRPAT for marketing: An introduction, expansion, and suggestions for future use. *Journal of Business Research*, 108:351–361, 2020.

P. Kovács, T. Petres, and L. Tóth. A New Measure of Multicollinearity in Linear Regression Models. *International Statistical Review*, 73(3):405–412, 2005.

S. Kumar. Environmentally sensitive productivity growth: A global analysis using Malmquist-Luenberger index. *Ecological Economics*, 56(2):280–293, 2006.

T.K. Kumar. Multicollinearity in regression analysis. *The Review of Economics and Statistics*, 57(3):365–366, 1975.

V. Kuperman, R. Bertram, and R.H. Baayen. Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23(7-8):1089–1132, 2008.

V. Kuperman, R. Bertram, and R.H. Baayen. Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, 62(2): 83–97, 2010.

J. Lauridsen and J. Mur. Multicollinearity in cross-sectional regressions. *Journal of Geographical Systems*, 8(4):317–333, 2006.

A. Lazaridis. A note regarding the condition number: the case of spurious and latent multicollinearity. *Quality & Quantity*, 41(1):123–135, 2007.

A. Lazaridis. *Dynamic Systems in Management Science: Design, Estimation and Control*. Springer, 2015.

E.E. Leamer. Multicollinearity: a Bayesian interpretation. *The review of economics and statistics*, 55:371–380, 1973.

T.R. Leighton. *Introductory econometrics: theory and applications*. Longman, New York, 1985.

K. Lemhöfer, T. Dijkstra, H. Schriefers, R.H. Baayen, J. Grainger, and P. Zwitserlood. Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):12, 2008.

Y.F. Li, M. Xie, and T.N. Goh. Adaptive ridge regression system for software cost estimating on multi-collinear datasets. *Journal of Systems and Software*, 83(11):2332–2343, 2010.

Z. Li, Y. Li, and S. Shao. Analysis of Influencing Factors and Trend Forecast of Carbon Emission from Energy Consumption in China Based on Expanded STIRPAT Model. *Energies*, 12(16):3054, 2019a.

Z. Li, S. Shao, X. Shi, Y. Sun, and X. Zhang. Structural transformation of manufacturing, natural resource dependence, and carbon emissions reduction: Evidence of a threshold effect from China. *Journal of cleaner production*, 206:920–927, 2019b.

S. Lin, D. Zhao, and D. Marinova. Analysis of the environmental impact of China based on STIRPAT model. *Environmental Impact Assessment Review*, 29(6):341–347, 2009.

H. Liu, X. Xu, and J.J. Li. *HDCI: High Dimensional Confidence Interval Based on Lasso and Bootstrap*, 2017. URL: `https://CRAN.R-project.org/package=HDCI`. R package version 1.0-2.

S. Liu, B. Peng, Q. Liu, and C. Fan. Economic-related CO2 emissions analysis of Ordos Basin based on a refined STIRPAT model. *Greenhouse Gases: Science and Technology*, 9:1064–1080, 2019.

X. Liu, K. Li, M. McAfee, and J. Deng. Application of nonlinear PCA for fault detection in polymer extrusion processes. *Neural computing and applications*, 21(6):1141–1148, 2012.

R. Lockhart, J. Taylor, R.J. Tibshirani, and R. Tibshirani. A significance test for the LASSO. *Annals of statistics*, 42(2):413, 2014.

M. Ma, R. Yan, and W. Cai. An extended STIRPAT model-based methodology for evaluating the driving forces affecting carbon emissions in existing public building sector: evidence from China in 2000-2015. *Natural Hazards*, 89(2): 741–756, 2017.

M.C. Mahutga and N. Bandelj. Foreign investment and income inequality: The natural experiment of Central and Eastern Europe. *International Journal of Comparative Sociology*, 49(6):429–454, 2008.

T.R. Malthus. *Essay on the principle of population*. JM Dent, 1973.

K.V. Mardia, J.T. Kent, and J.M. Bibby. Multivariate analysis. Probability and mathematical statistics. Academic Press Inc, London, 1979.

D.W. Marquardt. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970.

D.W. Marquardt. A Critique of Some Ridge Regression Methods: Comment. *Journal of the American Statistical Association*, 75(369):87–91, 1980.

D.W. Marquardt and S.R. Snee. Ridge Regression in Practice. *Journal of the American Statistical Association*, 29(1):3–20, 1975.

I. Martínez-Zarzoso and A. Maruotti. The impact of urbanization on CO_2 emissions: Evidence from developing countries. *Ecological Economics*, 70(7): 1344–1353, 2011.

I. Martínez-Zarzoso, A. Bengochea-Morancho, and R. Morales-Lage. The impact of population on CO_2 emissions: evidence from European countries. *Environmental and Resource Economics*, 38(4):497–512, 2007.

M. Meloun, J. Militkỳ, M. Hill, and R.G. Brereton. Crucial problems in regression modelling and their solutions. *Analyst*, 127(4):433–450, 2002.

B.H. Mevik, R. Wehrens, and K.H. Liland. *pls: Partial Least Squares and Principal Component Regression*, 2019. URL: `https://CRAN.R-project.org/package=pls`. R package version 2.7-1.

H. Midi, S.K. Sarkar, and S. Rana. Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3):253–267, 2010.

R.C. Mittelhammer, D.L. Young, D. Tasanasanta, and J.T. Donnelly. Mitigating the effects of multicollinearity using exact and stochastic restrictions: The case of an aggregate agricultural production function in Thailand. *American Journal of Agricultural Economics*, 62(2):199–210, 1980.

V. Moutinho, M. Madaleno, and M. Robaina. The economic and environmental efficiency assessment in EU cross-country: Evidence from DEA and quantile regression approach. *Ecological Indicators*, 78:85–97, 2017.

M.P. Murray. *Econometrics: A modern introduction*. Pearson Higher Education, 2005.

R.H. Myers. *Classical and modern regression with applications*. Duxbury, Thomson Learning, Belmont, CA, second edition, 1990.

D. Neeleman. *Multicollinearity in linear economic models*. Tilburg Institute of Economics, Netherlands, 1973.

J. Neter, W. Wasserman, and Kutner M.H. *Applied Linear Regression Models*. Irwin, Homewood, IL (USA), third edition, 1989.

A. Novales. *Econometría*. McGraw-Hill, Madrid, 1988.

A. Novales. *Econometría*. McGraw-Hill, Madrid, second edition, 1993.

A. Novales. Análisis de Regresión. `https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf`, 2010. Visited 7 April 2015.

A. Novales, R. Salmerón, C. B. García, J. García, and M. M. López-Martín. Tratamiento de la multicolinealidad aproximada mediante variables ortogonales. In *Anales de Economía Aplicada. XXIX Congreso Internacional de Economía Aplicada*, pages 1212–1227, 2015.

R.M. O'Brien. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41:673–690, 2007.

J. O'Hagan and B. McCabe. Tests for the severity of multicolinearity in regression analysis: A comment. *The Review of Economics and Statistics*, pages 368–370, 1975.

M.P. Pablo-Romero and J. De Jesús. Economic growth and energy consumption: The Energy-Environmental Kuznets Curve for Latin America and the Caribbean. *Renewable and Sustainable Energy Reviews*, 60:1343–1350, 2016.

Y. Pan and R.T. Jackson. Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males. *Epidemiology & Infection*, 136(3):421–431, 2008.

H.T. Pao and C.M. Tsai. CO_2 emissions, energy consumption and economic growth in BRIC countries. *Energy Policy*, 38(12):7850–7860, 2010.

G.R. Pasha and M.A.A. Shah. Application of ridge regression to multicollinear data. *Journal of research (Science)*, 15(1):97–106, 2004.

R.K. Paul. Multicollinearity: Causes, effects and remedies. M. Sc. (Agricultural Statistics), Roll No. 4405. IASRI, New Delhi, 2006.

K. Pearson. Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, 6(2):566, 1901.

M.H. Pesaran. *Time series and panel data econometrics*. Oxford University Press, 2015.

A.A. Rafindadi. Revisting the concept of environmental Kuznets curve in period of energy disaster and deteriorating income: Empirical evidence from Japan. *Energy Policy*, 94:274–284, 2016.

Y. Rasool, S.A.H. Zaidi, and M.W. Zafar. Determinants of carbon emissions in Pakistan's transport sector. *Environmental Science and Pollution Research*, 26:22907–22921, 2019.

M. Robaina-Alves, V. Moutinho, and P. Macedo. A new frontier approach to model the eco-efficiency in European countries. *Journal of Cleaner Production*, 103:562–573, 2015.

J.T. Roberts and P.E. Grimes. Carbon Intensity and Economic Development 1962-1991: A Brief Exploration of the Environmental Kuznets Curve. *World Development*, 25(2):191–198, 1997.

J. Roca and E. Padilla. Emisiones atmosféricas y crecimiento económico en España: la curva de Kuznets ambiental y el Protocolo de Kyoto. *Economía Industrial*, 351(1):73–86, 2003.

R.C. Rockwell. Assessment of multicollinearity: The Haitovsky test of the determinant. *Sociological Methods & Research*, 3(3):308–320, 1975.

A. Rodríguez, R. Salmerón, and C. B. García. The coefficient of determination in the ridge regression. *Communications in Statistics-Simulation and Computation*, 2019.

M. Roy, S. Basu, and P. Pal. Examining the driving forces in moving toward a low carbon society: an extended STIRPAT analysis for a fast growing vast economy. *Clean Technologies and Environmental Policy*, 19(9):2265–2276, 2017.

R. Salmerón and E. Rodríguez. Métodos cuantitativos para un modelo de regresión lineal con multicolinealidad. Aplicación a rendimientos de letras del tesoro. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, 24:169–189, 2017.

R. Salmerón, J. García, C. B. García, and C. García. Treatment of collinearity through orthogonal regression: an economic application. *Boletín de Estadística e Investigación Operativa*, 32(3):184–202, 2016. URL: `https://bit.ly/2M7KxMO`.

R. Salmerón, J. García, C. B. García, and M. M. López-Martín. The raise estimators. estimation, inference and properties. *Communications in Statistics-Theory and Methods*, 46(13):6446–6462, 2017a.

R. Salmerón, J. García, C. B. García, and M. M. López-Martín. A note about the corrected VIF. *Statistical Papers*, 58(3):929–945, 2017b.

R. Salmerón, C. B. García, and J. García. Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation*, 88(12):2365–2384, 2018.

R. Salmerón, J. García, C. B. García, and M.M. López-Martín. Transformation of variables and the condition number in ridge estimation. *Computational Statistics*, 33(3):1497–1524, 2018.

R. Salmerón, A. Rodríguez, and C. B. García. Diagnosis and quantification of the non-essential collinearity. *Computational Statistics*, 2019.

P.A. Samuelson, T.C. Koopmans, and J.R.N. Stone. Report of the evaluative committe for Econometrica. *Econetrica*, 22:242–256, 1954.

H. Scheel. Undesirable outputs in efficiency valuations. *European Journal of Operational Research*, 132(2):400–410, 2001.

M.A. Schroeder. Diagnosing and dealing with multicollinearity. *Western Journal of Nursing Research*, 12(2):175–187, 1990.

P.C. Schulze. I = IPBAT. *Ecological Economics*, 40(2):149–150, 2002.

L.M. Seiford and J. Zhu. Modeling undesirable factors in efficiency evaluation. *European Journal of Operational Research*, 142(1):16–20, 2002.

L.S. Shapley. A value for n-person games. *Contributions to the Theory of Games (AM-28)*, 2:307, 2016.

A. Sherbinin, D. Carr, S. Cassels, and L. Jiang. Population and environment. *Annual Review of Environment and Resources*, 32:345–373, 2007.

C. Shuai, X. Chen, Y. Wu, Y. Tan, Y. Zhang, and L. Shen. Identifying the key impact factors of carbon emission in China: Results from a largely expanded pool of potential impact factors. *Journal of Cleaner Production*, 175:612–623, 2018.

S.D. Silvey. Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(3):539–552, 1969.

G. Smith and F. Campbell. A Critique of Some Ridge Regression Methods. *Journal of the American Statistical Association*, 75(369):74–81, 1980.

R.D. Snee and D.W. Marquardt. Comment: Collinearity diagnostics depend on the domain of prediction, the model, and the data. *The American Statistician*, 38(2):83–87, 1984.

A. Spanos and A. McGuirk. The problem of near-multicollinearity revisited: erratic vs systematic volatility. *Journal of Econometrics*, 108(2):365–393, 2002.

G.W. Stewart. Collinearity and least squares regression. *Statistical Science*, 2 (1):68–100, 1987.

J.M. Stock and M.M. Watson. *Introducción a la econometría*. Pearson, Madrid, third edition, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

G. Tintner. Methodology of mathematical economics and econometrics. The University of Chicago press, 1968.

M. Torras and J.K. Boyce. Income, inequality, and pollution: a reassessment of the environmental Kuznets Curve. *Ecological Economics*, 25(2):147–160, 1998.

G.A. Uddin, K. Alam, and J. Gow. Estimating the major contributors to environmental impacts in Australia. *International Journal of Ecological Economics and Statistics*, 37(1):1–14, 2016.

E. Uriel. *Econometría: el modelo lineal*. Alfa Centauro, Madrid, 1997.

A.H. Vencheh, R.K. Matin, and M.T. Kajani. Undesirable factors in efficiency measurement. *Applied Mathematics and Computation*, 163(2):547–552, 2005.

E. Vigneau, M.F. Devaux, E.M. Qannari, and P. Robert. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(3):239–249, 1997.

P.E. Waggoner and J.H. Ausubel. A framework for sustainability science: a renovated IPAT identity. *Proceedings of the National Academy of Sciences*, 99(12):7860–7865, 2002.

J. Walton and C. Ragin. Global and national sources of political protest: Third world responses to the debt crisis. *American Sociological Review*, 55:876–890, 1990.

T. Wei. What STIRPAT tells about effects of population and affluence on the environment? *Ecological Economics*, 72:70–74, 2011.

L. Wen and H. Shao. Analysis of influencing factors of the carbon dioxide emissions in China's commercial department based on the STIRPAT model and ridge regression. *Environmental Science and Pollution Research*, 26(26): 27138–27147, 2019.

C.R. Wichers. The detection of multicollinearity: A comment. *The Review of Economics and Statistics*, 57(3):366–368, 1975.

H. Wold. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis* (ed. P.R. Krishnaiaah), 391-420. New York: Academic Press, 1966.

J.M. Wooldridge. *Introducción a la econometría. Un enfoque moderno.* Thomson Paraninfo, Madrid, second edition, 2008.

B. Woolf. Computation and interpretation of multiple regressions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 100–119, 1951.

L.H. Wurm and S.A. Fisicaro. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, 72:37–48, 2014.

Q. Xie and J. Liu. Combined nonlinear effects of economic growth and urbanization on CO2 emissions in China: Evidence from a panel data partially linear additive model. *Energy*, 186:115868, 2019.

S.C. Xu, Y.W. Li, Y.M. Miao, C. Gao, Z.X. He, W.X. Shen, R.Y. Long, H. Chen, B. Zhao, and S.X. Wang. Regional differences in nonlinear impacts of economic growth, export and FDI on air pollutants in China based on provincial panel data. *Journal of Cleaner Production*, 228:455–466, 2019.

R. Yang and W. Chen. Spatial Correlation, Influencing Factors and Environmental Supervision on Mechanism Construction of Atmospheric Pollution: An Empirical Study on SO2 Emissions in China. *Sustainability*, 11(6):1742, 2019.

R. York. Residualization is not the answer: Rethinking how to address multicollinearity. *Social science research*, 41(6):1379–1386, 2012.

R. York, E.A. Rosa, and T. Dietz. STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts. *Ecological economics*, 46(3):351–365, 2003.

X. Yuan, L. Ye, L. Bao, Z. Ge, and Z. Song. Nonlinear feature extraction for soft sensor modeling based on weighted probabilistic PCA. *Chemometrics and Intelligent Laboratory Systems*, 147:167–175, 2015.

S. Zhang and T. Zhao. Identifying major influencing factors of CO2 emissions in China: Regional disparities analysis based on STIRPAT model from 1996 to 2015. *Atmospheric Environment*, 207:136–147, 2019.

Y. Zhang, Q. Zhang, and B. Pan. Impact of affluence and fossil energy on China carbon emissions using STIRPAT model. *Environmental Science and Pollution Research*, 26:1–11, 2019.

P. Zhou and B.W. Ang. Linear programming models for measuring economy-wide energy efficiency performance. *Energy Policy*, 36(8):2911–2916, 2008.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.