Bulletin de Méthodologie Sociologique
2018, Vol. 140 90–109

DOI: 10.1177/0759106318795218
journals.sagepub.com/home/bms
SAGE

Some Contributions from Geometry to Linear Models' Construction in Social Sciences

Frédéric Lebaron
Ecole Normale Supérieure Paris – Saclay, France

Andrés F. Castro T.
University of Pennsylvania, USA

Résumé

Sommes-nous assez prudents quand nous utilisons les modèles de régression linéaire ? Depuis les années 1970, ces derniers se sont imposés comme la méthode la plus commune dans les sciences sociales quantitatives. Leur champ d'application et leurs limites n'ont été que trop rarement abordés. Dans ce texte, nous questionnons une étape majeure du protocole de modélisation : la sélection des variables explicatives. Nous montrons tout d'abord la nécessité d'une comparaison systématique des modèles de régression bivariée et multivariée, dans la mesure où les variables indépendantes généralement sélectionnées ne sont pas toujours orthogonales entre elles, et peuvent aboutir à des coefficients incertains. Nous mobilisons ensuite la méthode de la représentation géométrique des modèles linéaires afin de visualiser les origines et les causes de l'instabilité des modélisations linéaires classiques, puis proposons l'usage de la *residual regression* comme alternative. Nous illustrons ces propos à partir des données collectées par Cukierman et al. (2002) concernant la régulation gouvernementale et l'inflation dans 26 pays. Nos conclusions soulignent l'importance de considérer les effets de structure et invitent à une posture parcimonieuse des modèles de régression, qui ne compteraient qu'un faible nombre de variables indépendantes.

Corresponding Author:
Andrés F. Castro T., Population Studies Center and Department of Sociology, University of Pennsylvania, Philadelphia, Pennsylvania, USA
Email: candres@sas.upenn.edu

Abstract
Are we cautious enough when using linear models? After the 1970s linear models became the most common method for quantitative social scientists. More discussion on their scope and limitations is needed. We focus on one stage of the modeling process, namely, variable selection. We show that a rigorous comparison between bivariate and multivariate regression models should be done in this stage as non-orthogonality among predictors can lead to ambiguous estimates. Further, we use geometrical representations of linear models for two purposes. First, to visualize sources of instability and the causes of ambiguous results. Second, to support residual regression as an alternate approach. We illustrate our ideas using data collected by Cukierman et al. (2002) on the relationship between government regulation and inflation in 26 countries. Our conclusions stress the need to assess structural effects and support parsimonious models with few predictors.

Mots clés
Analyse de données, modèles linéaires généralisés, géométrie, méthodologie quantitative en sciences sociales, effets de structure

Keywords
Data analysis, generalized linear models, geometry, quantitative methodology in social sciences, structural effects

Introduction

A substantial increase in the use of the linear modeling techniques in social science research has occurred since the late 1960 (Ollion, 2011; Cornwell, 2015: chap. 2). By linear models we refer to the statistical models in which the relationship between the expected value of a dependent variable $E[Y]$ and a set of covariates, comprised in the model matrix X, is specified through a link function $g(\cdot)$, and a set of parameters, typically presented as a vector denoted by the character β.

$$E[Y] = g(X\beta) \tag{1}$$

Under a classic statistical approach, the β-parameters are linear, fixed and unknown. Several estimation techniques have been developed and incorporated into statistical packages, for a wide range of probability distributions, in particular for the so-called family of exponential distributions (Nelder and Wedderburn, 1972; McCullagh, 1984; Dobson and Barnett, 2008). This situation has contributed to increase the number of studies relying on linear models, termed Generalized Linear Models (GLM) after Nelder & Wedderburn's (1972) seminal work. These models have served to produce and refine theories about social phenomena in multiple domains.

Considerable attention has been devoted to understanding the potential of this approach both technically and theoretically; less attention has been devoted, however, to its limitations. In technical terms, criteria for the selection of variables for the right-hand side of equation 1 and the interpretation of the estimates continue to be challenged in the literature (Rouanet et al., 2002; Deauvieau, 2010; Selz and Deauvieau, 2011; Bry et al., 2016). From a theoretical standpoint, concerns about the appropriateness of this approach to study

certain social phenomena have been raised, mainly opposing it to non-linear/exploratory statistical techniques (Darras, 1966; Abbot, 1988; Hirschman, 1994; Desrosières, 2001, 2008; Bourdieu, 2005)[1]. To give the reader a sense of the tone of this discussion – at least in Sociology – we selected a quote from Hirschman's article on theories of fertility change:

> The standard social science model is that society works pretty much like a regression equation: the task is to find the right set of predictors, solve the equation, and discover what factors are most important in predicting social outcomes. This framework does lead to empirical generalizations, but there seem to be endless qualifications about the measurement of variables, the meaning and interpretation of variables, the substitutability of one variable for another, and complex interactions with historical settings. If science is to discover parsimonious principles that explain complex patterns, we do not seem to be making progress (1994: 226).

This assessment does not differ from more recent debates about the nature of research and theory in the realm of family demographics, where regression-based approaches are also dominant (Johnson-Hanks et al., 2011: Introduction). On the other side of the discussion, linear models are often presented as the unique statistical tool to explain social phenomena, in particular among studies in quantitative sociology and demography; not to mention economics (Morgan, 1990; Lebaron, 2000). Other statistical methods are deemed as merely descriptive, i.e. as devoid of any explanatory power. Moderated views do not assume a hierarchy between description and explanation, and present the two approaches as potentially complementary (Bernard et al., 1989; Lebart et al., 1997; Lieberson and Horwich, 2008; Bry et al., 2016); this reconciliatory stands are rather scarce. Historically, the separation between multivariate (geometry-based) descriptive approaches and model-based ones seem to divide the Anglo-Saxon tradition of statistical analysis from the so-called French one. This situation is clearly portrayed in, but not limited to, discussions surrounding Pierre Bourdieu's work (Rouanet et al., 2000; Gollac, 2004; Savage et al., 2013; Lebaron and Le Roux, 2015).

Our aim is not to solve these discussions but to call for more caution when using linear models in the social sciences. We further suggest the need to relocate the opposition between explanation and description from the realm of statistical methods to the realm of theory. In other words, we contend that all statistical approaches can be explanatory (i.e. they can be used to develop theories about social behavior) when they are appropriately informed by theory. Hence, we call for more methodological openness when conducting quantitative research in the social sciences.[2]

More specifically, we critically assess two stages of the process of modeling, namely variable selection and the interpretation of the estimates. The assessment of these two stages is crucial given the rapid growth of large data sets on social issues and the predominance of linear modeling techniques in quantitative social sciences. Based on a systematic comparison of bivariate and multivariate models and their geometric representation, we claim that linear modeling, as any other statistical technique, does not provide by itself an explanation of social phenomena. Rather, statistical methods should be understood as tools to make data intelligible under certain theoretical frameworks (Bourdieu and Wacquant, 1992).

Building and interpreting linear models

Reliability and interpretability are desirable characteristics for the estimates of a linear model. As the estimates may change depending on the variables that are included in a model, variable selection is a crucial stage. Technical and theoretical criteria to include or exclude predictors do not necessarily align. For example, the inclusion of a variable, very important for the theory at hand, can easily worsen the goodness of fit of the model. Depending on the main goal of a model (summary vs prediction) one may favor technical or theoretical arguments. Going back to the basis of modeling and to their geometric interpretation can be informative for the purposes of variable selection.

For long time, statisticians have warned analysts by recalling them that *all models are wrong*, but some of them are useful for research (Box, 1976). Useful models are typically models in which the predictors are: (i) few in number, (ii) well-clarified and (iii) measured with small error (Box, 1976; Mosteller and Tukey, 1977; Dobson and Barnett, 2008). Commandment number one speaks directly to the variable selection stage. It is well known that when predictors are correlated, estimation techniques are not able to capture the so-called *true* relationship between the dependent and the explanatory variables. This situation is commonly known as *quasi-collinearity*.

In principle, a useful research-model should not include many predictors. However, including multiple variables on the right-hand side of equation 1 is desirable insofar as it allows one to adjust estimates by theoretically-relevant factors. For instance, in economic studies on wage differentials by gender, it is important to control for factors such as educational attainment and economic sector if one is interested in measuring the level of discrimination against women in the labor market (Bruno, 2010). In epidemiological studies looking at the contribution of smoking to mortality, researchers often want to control for educational attainment and race, because of the potential relationship between these two variables and the outcome of interest (Ho and Elo, 2013). Another classic example can be found in demographic studies on the role of education in fertility decline. In this case, educational attainment is an explanatory variable (not a control) as it is thought to influence women's fertility decisions. These studies have a strong case to control for: place of residence, migration status, occupation and wealth, given the large heterogeneity of fertility along these dimensions (Castro and Juárez, 1995).

Typically, research strategies start with a bivariate model where the outcome variable is predicted by the variable of interest (sex, smoking behavior, educational attainment). Further, control variables are added to the model. When both the direction and the significance of the estimates do not change after adding control variables, the interpretation supports the bi-variate association. For instance, if wage differences between men and women remain statistically significant after controlling for education and economic sector, the conclusion will go as follows: all things being equal, women have, on average, lower wages than men. On the contrary, changes in the magnitude of the estimates and their significance level after adding control variables are interpreted as if the control mediates the relationship between the variable of interest and the outcome. A classic example of this can be found in studies of educational attainment and migration. Vallet and Caille (1996) reported a negative association between educational attainment and migration. However, once the authors control for family background characteristics, this

negative associations reverses. An analogous idea is at the core of demographic techniques of standardization. It is well known that comparisons of crude mortality rates can be misleading when the two populations of interest differ in their age-structure. Thus, adjusting – controlling – for age-structure is required to appropriately compare mortality levels (Preston et al., 2001; Deauvieau, 2011).

Control variables play the role of adjusting factors, yet the joint inclusion of several controls – over controlling – makes the adjusting process unclear. Potential ambiguous results as well as artificial significant results can arise when models are estimated including too many control variables. The sources of these potential problems can be pedagogically described using geometric representations and further assessed through specific measures.

Geometric representations of linear models help us do two things: (1) to reinterpret the role of control variables in a linear model, and (2) to identify one potential solution when ambiguous results occur. By geometric representation we refer to the generation of a tri-dimensional space for the variables – one dependent and two independent ones – and the presentation of regression outputs as orthogonal (bivariate models) and oblique projections (multivariate models) of the dependent variable on the independent variables and on the plane spanned by them, respectively (Le Roux and Rouanet, 2004: chap. 1). Given that more than three dimensions are impossible for us to visualize simultaneously, for cases involving more than three variables we rely on the ratio between the size of the orthogonal and the oblique projection as an indicator to evaluate the influence of control variables. In other words, the tri-dimensional case is used for pedagogical purposes and its generalization is presented throughout numerical outputs. As for a potential solution, geometric representations also show that by subtracting residuals from original variables one can obtain predictors that are orthogonal to one another. These new predictors can be used to estimate new models, a technique known in the literature as residual regression. Residual regression constitutes the last step of our analysis.

We apply these procedures to a set of linear models based on Cukierman et al.'s (2002) data to assess two aspects: (1) how geometric representations (and ratios) can be informative for the selection of variables and for the construction of orthogonal predictors (residual regression), (2) the extent to which author's conclusions coincide with the conclusions we will draw from a more parsimonious model and from a residual regression approach. In this case, we find that geometric representations support the selection of a model with fewer predictors than a theoretical approach would suggest. However, we also find that, author's conclusions basically hold for both models. We argue that more attention should be devoted to selecting variables and interpreting coefficients since there is no guarantee for the two approaches to always produce the same results.

Data and methods

Data

The data was originally recorded by Cukierman et al. (2002) to test the extent to which the independence of central banks from the state could affect inflation within 26 former socialist countries. Up to three years were recorded for each country between 1989 and

Table 1. Descriptive Statistics, Standardized Names and Correlations Among Dependent and Explanatory Variables

Variable Sth. variable	**INF** y	**IND** x_1	**WAR** x_2	**GLI** x_3	**PLI** x_4	**MIN** x_5
Minimum	0.020	0.000	0.000	0.000	0.000	0.000
Mean	0.381	0.288	0.211	2.084	0.577	0.245
Std. deviation	0.209	0.268	0.411	1.784	0.317	0.288
Maximum	0.790	0.850	1.000	6.340	0.900	0.850
Correlations						
INF	1.000					
IND	−0.464	1.000				
WAR	0.425	−0.236	1.000			
GLI	−0.504	0.882	−0.199	1.000		
PLI	−0.234	0.779	−0.072	0.820	1.000	
MIN	−0.585	0.931	−0.220	0.889	0.684	1.000

1998. The final data set contains 57 observations (country-year) for which the following variables were recorded (refer to Table 1 for descriptive statistics).

- **Inflation** (INF): it is measured by the depreciation of currency rates. It was calculated using equation 2, where F represents the usual inflation indicator. By using a homographic function of the usual consumer price index, the indicator is standardized.

$$INF = F\,(1+F)^{-1} \tag{2}$$

- **Independence of central banking** (IND): an index that measures the level of independence of the central bank with respect to the government.
- **War/non-war conditions** (WAR): dummy variable coded as 1 for war and 0 for the absence of war.
- **Global liberalization** (GLI): an index that measures the degree of global liberalization.
- **Price liberalization** (PLI): an indicator taken from the previous index, it measures the degree of liberalization of prices.
- **Multiplicative independence** (MIN): a combination of the two previous variables defined as:

$$MIN = 0 \; if \; GLI \leq 2 \; or \; MIN = IND \; if \; GLI > 2 \tag{3}$$

Inflation is the dependent variable, while the remaining ones are separated into two groups. Variables that characterize the relative independence of central banking are treated as *variables of interest,* whereas the indicator for war is used as a *control*; that is, it is included to consider the potential effect of war on inflation.

Table 1 presents descriptive statistics and correlation coefficients for each of the six variables. Additionally, Table 1 shows the name for the standardized version of each variable that will be used for further comparisons.

There is a strong correlation among most of the variables. Eight out of the 15 correlation coefficients have an absolute value above 0.5. This is not surprising because the variables of interest are measuring the same concept (liberalization). Additionally, there are good theoretical reasons to expect a strong correlation between the variables of interest and inflation. Moreover, the PLI indicator is part of the GLI index, which implies correlation by construction. Similarly, as MIN is a positive function of GLI and IND, their correlation comes by construction. A weak correlation is recorded between PLI and WAR (−0.072) and between GLI and WAR (−0.199). We will refer to these results later.

Methods

We fit all possible models for each variable to check how estimates change across models. These models are specified by combining all predictors while keeping one of them at a time. For example, when we focus on the first variable (x_1) we estimate all models with one, two, three, four and five predictors keeping x_1 in all specifications. The first model uses only x_1 as predictor, the second one uses x_1 and x_2, the third one x_1 and x_3, etc. The number of possible models per variable corresponds to the number of combinations of size p taken a group of n variables, noted as C_n^p. Hence, for each variable (x_1, x_2, x_3, x_4 and x_5) there is one single possible model with one predictor, four models with two predictors ($C_4^1 = 4$), six models with three predictors ($C_4^2 = 6$), four models with four predictors ($C_4^3 = 4$) and one last model that include all covariates; for a total of 16 models per variable.[3] The five bivariate models constitute the baseline for our comparisons.

We then assess the variability of the coefficients associated to each variable with respect to the baseline model. Small changes are not problematic because the interpretation remains the same. Instead, substantial changes – i.e. changes in the direction of the association – can be problematic as they reduce the intelligibility of the coefficients. In other words, unstable estimates imply that one could draw contradictory conclusions depending on the set of variables included in the model. Based on these results we select a model that only includes variables with stable results across specifications – we term this model *parsimonious* model. Further, we provide the geometric representation of all the bi-variate, the parsimonious and the full model. Finally, we use residual regression as an alternative to avoid potential misinterpretation of linear models' estimates. This technique makes all predictors orthogonal to one another, making estimates robust to the inclusion of several predictors.

As a reminder, geometric representations serve two purposes. First, they provide evidence on the usefulness of interpreting regression outputs as orthogonal and oblique projections among vectors. Second, they help us to visualize the cases in which the inclusion of a control can lead to an increase, a decrease, or a reversal in the size of an estimate. Each of these cases is presented as an area within the correlation circle, and factors associated to the size of each area are discussed.

Table 2. Coefficients and R^2 for Bivariate Models

	x_1	x_2	x_3	x_4	x_5
Beta	−0.464	0.425	−0.504	−0.234	−0.585
R^2	(0.215)	(0.181)	(0.254)	(0.055)	(0.343)

Results

Bivariate, parsimonious and full models

Table 2 presents the bivariate associations between each of the five independent variables and inflation. These figures give an imperfect information but more stable/intelligible than the information coming from a full model as they do not depend on other variables (Rouanet et al., 2002). These numbers provide references for further comparisons.

As the variables are standardized, the bivariate associations are equal to the correlation coefficients presented in Table 1. For x_1 (degree of independence of the central banking) the correlation is -0.464, which implies that among the observed country-years, an increase in the independence of the central bank is associated with a decrease in inflation. The proportion of variance explained by this model corresponds to the square of the coefficient $(-0.464)^2 = 0.215$. All the other three variables of interest also exhibit negative associations with inflation. These results are consistent with the economic theory behind the model. Conversely, the variable related to war conditions has a positive association. This is not unexpected given the well-known inflationary effects of war.

Figure 1 summarizes the results from all the 80 models that were fitted (16 per variable). We are aware some models are duplicated; however, we chose to keep them all as the analysis is carried out by variable. The left panel displays the distribution of the 48 coefficients per variable (16 are non-redundant). For instance, the first boxplot (x_1) corresponds to the estimated coefficients for the variable x_1, i.e. β_1 from all models that include x_1. The gray line corresponds to the bivariate association (i.e. model including only x_1). On the right panel, each model is represented by several points depending on the number of covariates (e.g. four points for a model with four covariates), the R^2 and the estimated association between y and the covariates. For example, the top-most square marker (□) on the right panel correspond to one out of the four representations of the model $y \sim x_1 + x_2 + x_3 + x_5$. The x-coordinate of this point is the R^2 associated to this model (0.5), and the y-coordinate corresponds to the estimate of β_1, i.e. the conditional association of x_1 and y. The point is represented with a square as the variable of interest is x_1. This very same model is represented by three other points. These three points have the same x-coordinate ($R^2 = 0.5$) but different y-coordinate and marker, depending on the variable of interest. The y-coordinates for these three points correspond to the estimates of β_2 (0.336, ▲-triangle), β_3 (-0.107, ○-circle) and β_5 (-1.1, X-marker).[4]

A gray background was added to points representing models with two covariates. These models are particularly interesting for the following reason: if there are two important factors influencing inflation, namely war conditions and liberalization, then

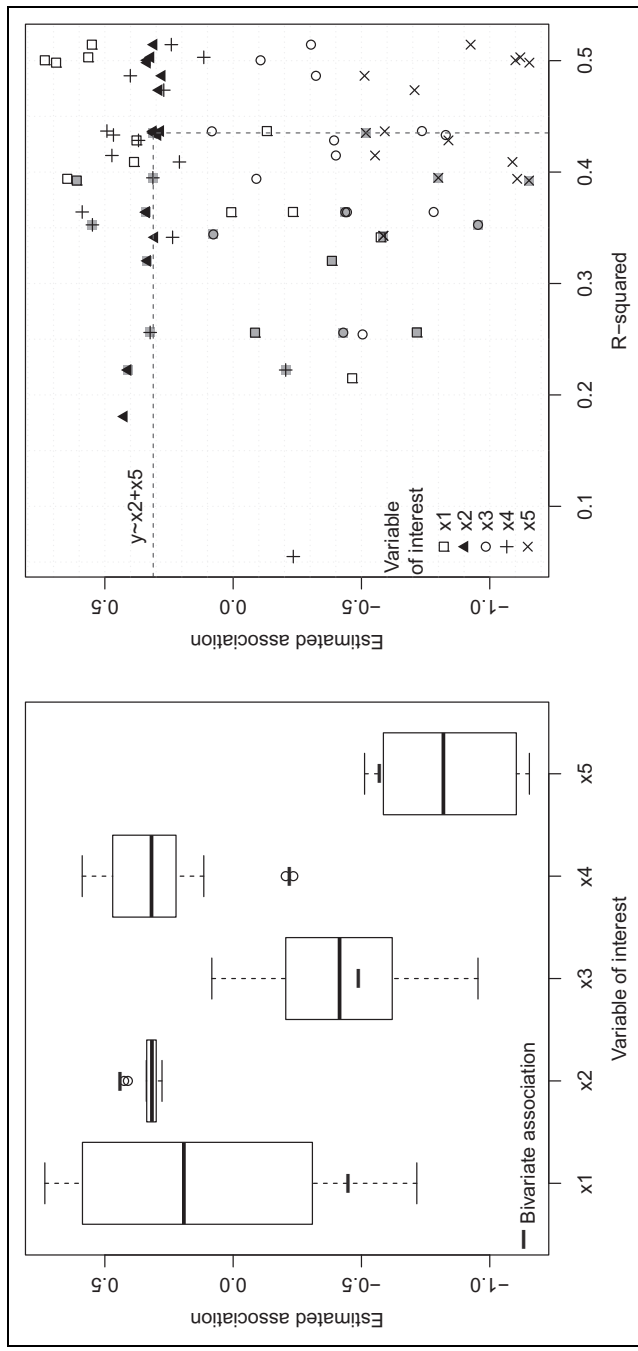

Figure 1. Summary results for models by variable of interest.

two good measures of these concepts should be sufficiently informative, i.e. they should explain a large proportion of the variance, while providing consistent estimates for the association between them and the outcome variable. Two segments were added to point-out the specification with two covariates that has the largest R^2, namely, $y \sim x_2 + x_5$.

Three conclusions can be derived from this figure.

1. The coefficients associated to variable x_2 are stable across models. This represents "the good case" and is due to the relatively low correlation between x_2 and the other variables (-0.072, -0.199, -0.220, -0.236 see Table 1).
2. For variables x_1, x_3, x_4 and x_5 coefficients are unstable. As compared to the bivariate association, the sign and magnitude of these coefficients vary from one model to another. This occurs to a lesser extent for x_5, for which the estimated coefficients do not change sign. This situation can be termed as "the bad case" to the extent that in some cases the relationship is positive and in some others the relationship is negative. Even though these changes may have a theoretical explanation (such as in the cases discussed in the previous section), in a context of several control variables (more than 3) an exploration of the sources of instability is necessary as it is hard to believe that a single theoretical explanation could accommodate multiple changes in several coefficients at a time.
3. Among the models with two variables (x_2 being one of them), the one that includes variables x_2 and x_5 has the largest R^2. Not surprisingly given the low correlation between x_2 and x_5 (see Table 1).

On the one hand, the model that includes x_2 and x_5 constitutes a *parsimonious* alternative to study the relationship between independence of central banking and inflation, controlling for the potential effect of war. On the other, the full model includes the five variables of interest, as the theory suggests that all these dimensions of liberalization can influence inflation. We termed this model *saturated*. For the sake of brevity, we label the saturated model as (A) and the parsimonious one as (B). Cukierman et al. relied on the full model and their main conclusion goes as follows:

> Once the process of liberalization has gone far enough legal independence turns out to be effective in slowing inflation down. [. . .] The cumulative index of liberalization developed by de Melo et al. (1996) exerts a significant negative influence on inflation, as is the case in their paper, mainly at low levels of cumulative liberalization (Cukierman et al., 2002: 19).

This conclusion is consistent with both models A and B, and with the bi-variate associations. Yet, the sizes of the estimates are different in each of them. These differences are particularly large for the variable of interest, namely, MIN (refer to Table 3). Moreover, the difference in the R^2 between model A and B is not that large relative to the difference in the number of predictors each model includes (five vs two).

The bottom panel in Table 3 displays as set of ratios termed structural effects. Structural effects correspond to the ratio between the bivariate association and the estimates of models A and B for each variable. For example, the ratio between the bivariate association of x_1 and y, and the estimate for x_1 in model A is computed as

Table 3. Comparison of Different Models

Model	IND x_1	WAR x_2	GLI x_3	PLI x_4	MIN x_5	R^2
Bivariate	−0.464	0.425	−0.504	−0.234	−0.585	
R^2	(0.215)	(0.181)	(0.254)	(0.055)	(0.343)	
(A)	0.551	0.309	−0.303	0.241	−0.926	(0.514)
(B)		0.311			−0.517	(0.435)
Structural effects						
(A) / Bivariate	−1.19	0.73	0.60	−1.03	1.58	
(B) / Bivariate		0.73			0.88	

$(0.551/-0.464) = -1.19$. This ratio reflects a reversal in the sign of the coefficient and an increase of 19% in its absolute value when all variables are included. In other words, compared to the bivariate association of x_1 and y, the conditional association of x_1 and y in model A is larger and it has an opposite sign. β_1 in model A is not intelligible since its size and direction become ambiguous. This case corresponds to an extreme scenario (sign reversal).

Structural effects in model A are large. Two coefficients (β_1, β_4) changed their sign, while the rest displayed changes in their magnitude. The smallest reduction occurred to β_2, which is 27% smaller in model A compared to the bivariate specification (Structural effect = 0.73). For the pair of variables with the highest correlation, ($\rho(x_1, x_5) = +0.931$) the conditional association is larger than the bivariate one. Conversely, structural effects in model B are substantially smaller compared to those in model A (more stable coefficients). The more information is added to the model, the higher the risk of affecting the intelligibility of the coefficients due to the potential redundancy of the variables. Geometric representations give us visual tools to explore the sources of this instability.

Geometric representations

A standardized variable can be represented by a vector of norm equal to one. Then, all the variables in a linear model with one dependent variable and two predictors can be represented within a sphere of ratio equal to one. The correlation coefficient $\rho(i,j)$ between two variables (x_i, x_j) corresponds to the squared cosine of the angle (θ) between any given pair of vectors (Le Roux and Rouanet, 2004).

If two variables are not correlated, their geometrical representation will correspond to two orthogonal vectors ($\theta = 90°$) because $\text{Cos}(90°) = 0$. Weak correlations correspond to angles close to the right angle. If the variables are positively correlated the angle between the vectors will lay between $-90°$ and $90°$ (excluding both extremes). If the correlation is negative, the angle between them will pertain to the both-sides-opened-interval (90°, 270°). In all cases θ is a distance index that satisfies the triangle inequality, and the relation $\rho = \cos(\theta)$ is an index of proximity defined in the interval [0,1] (Le Roux and Rouanet, 2004: chap. 1). Figure 2 displays the geometric representation of variables y, x_2 and x_5. Figure 2 also displays orthogonal and oblique projections of y on each

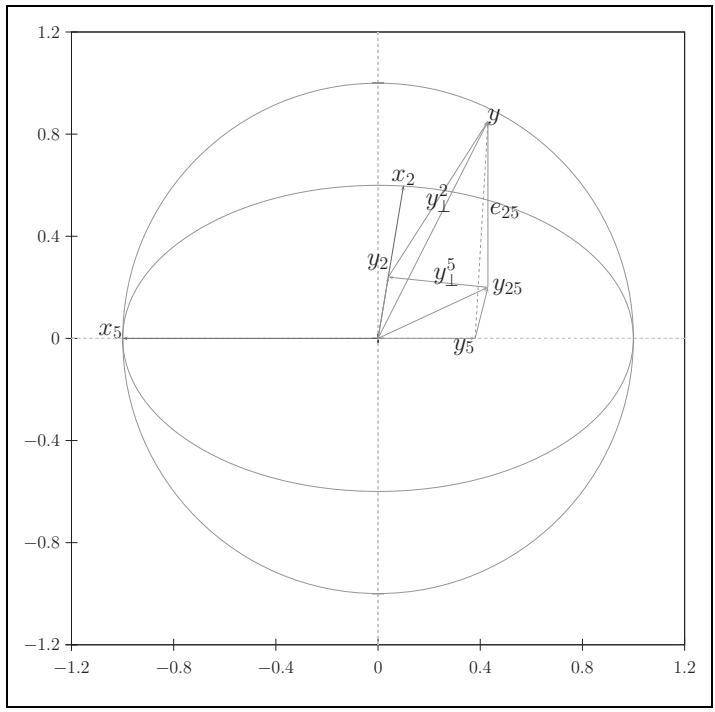

Figure 2. Correlation sphere with three reduced variables.

variable (x_2 and x_5) and on the plane they generate. These projections constitute the geometric representation of the regression outputs.

The regression of y on x_2 corresponds to the orthogonal projection of y on x_2 (labeled y_2). Note that the residual of this regression ($y_{\perp}^2 = y - y_2$) is perpendicular to x_2. The regression of y on x_2 and x_5 is the orthogonal projection of y on the plane defined by x_2 and x_5 (labeled y_{25}) and the residual ($e_{25} = y - y_{25}$) is an orthogonal vector with respect to the plane. This representation contains the basic ideas that help us better assess regression models.

All the aforementioned linear models can be defined in terms of orthogonal and oblique projections of the dependent variable (y) on the hyperplane formed by any combination of the independent variables (x_1, x_2, x_3, x_4 and x_5). The length of the orthogonal projections of y on each variable, noted as y_i (with $i = 1, 2, 3, 4$ and 5), corresponds to the bivariate association between variable x_i and y, the length of the oblique projection of y on x_i, noted as y^i, corresponds to the conditional one.

Figure 3 displays a cross-sectional view of Figure 2 to further explore model B. As established above (see Table 2), the conditional associations are smaller than the bivariate ones given the correlation among the two variables. Note that if x_2 and x_5 were independent (i.e. $\theta = 90°$) both the bivariate and the conditional associations will coincide. As it is not the case, vectors x_2 and x_5 delimit areas in which the bivariate and the conditional associations differ. The gray and the white semicircles within the

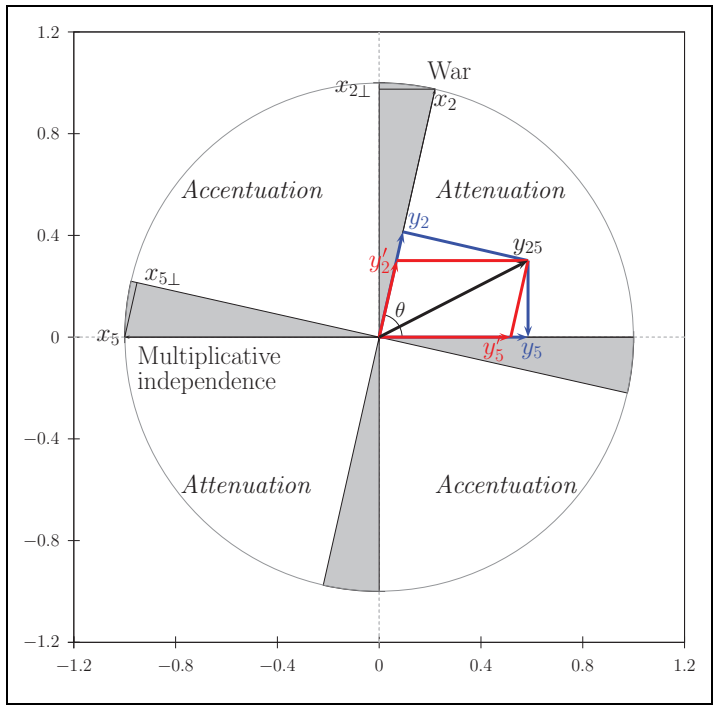

Figure 3. Correlation circle. Regression of y on (x_2, x_5).

circumference in Figure 3 correspond to areas of reversal and areas of non-ambiguous differences between bivariate and conditional associations, respectively. By reversal we mean that the conditional association has the opposite sign as compared to the bivariate one; by non-ambiguous change we mean that there is either *accentuation* or *attenuation* of the association in the same direction. The size of the gray area (reversal effect) depends on θ, i.e. on the correlation between the independent variables.

It follows that model B can be written as the sum of the two oblique projections $y_{25} = y'_2 + y'_5$. The R^2 of this regression (0.435) corresponds to the squared length of the vector y_{25}. A similar analysis can be done for model A. The components of model A are in a five-dimensional space which prevents us from making a plane representation of them. Equation 4 expresses model A as the sum of five oblique projections.

$$y_{12345} = y'_1 + y'_2 + y'_3 + y'_4 + y'_5 \tag{4}$$

For illustrative purposes and without any loss of generalization, equation 4 can be expressed as the sum of two oblique projections: $y_{12345} = y'_{1345} + y'_2$. The first term corresponds to the oblique projection of y on the hyperplane (x_1, x_3, x_4, x_5) parallel to x_2 and the last term to the oblique projection of y on x_2 parallel to the subspace (x_1, x_3, x_4, x_5). This plane also contains the regression of y on x_2 ($y_2 = 0.425\, x_2$). Figure 4 displays a graphic representation of these two oblique projections. As for model B, due to the correlation among the covariates, conditional and bivariate associations differ. Even

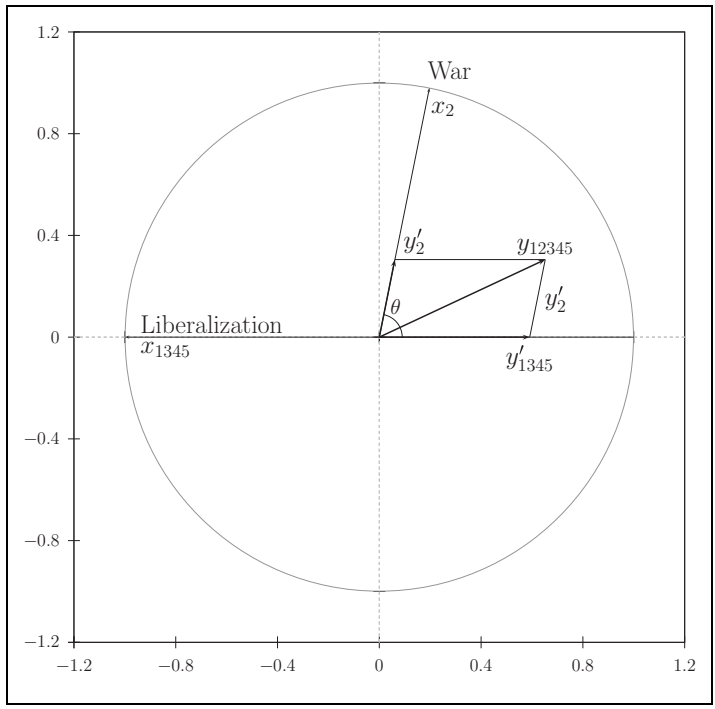

Figure 4. Regression of y on (x_2, x_{1345}).

though we restrict the analysis to two dimensions, our conclusions hold for any higher-order dimension as they are based on general mathematical properties.

In general, conditional and bivariate associations between a set of covariates and an outcome may differ. The potential difference among them is determined by the correlation of the variables included in the model (angles that influence the projections). Examples shown above led us to identify three cases: *accentuation* (which includes the *emergence* of the association), *attenuation* (which include the *dissipation* of the association) and *reversal* (changes in the sign of the association).

When these situations occur, we say that there is an effect due to the data's structure. We define the magnitude of this effect as the ratio between the conditional and the bivariate association (structural effect). If this ratio is larger than 1 it implies that the conditional association is bigger than the bivariate one (*accentuation*). If the ratio is below one and above zero there is *attenuation*, and if the ratio is negative there is *reversal*. The borderline cases among these three situations are stabilization (between attenuation and accentuation), disappearance (between attenuation and reversal) and emergence (between accentuation and reversal). For a rigorous study of such cases see the work by Rouanet et al. (2002), where a scheme of *rose des vents* is used to summarize the above-described situations.

Geometric representation suggests that residuals can be used to avoid the influence of the correlation among the variables of interest and the control. The residual of each regression is orthogonal with respect to the hyperplane generated by the independent

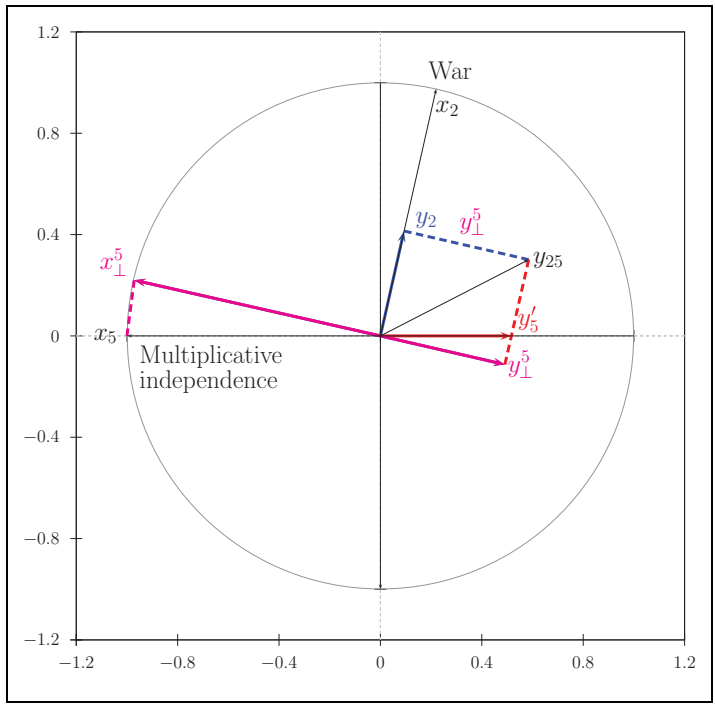

Figure 5. Residual of x_5 with respect to x_2.

variables. Using this fact, it is possible to use residuals as an orthogonal version of each variable of interest. This process is known as residual regression and will be described in the next section from a geometric point of view.

Residual regression

Each of the variables of interest (x_1, x_3, x_4 and x_5) was regressed on the control variable (x_2) to obtain a new version of each of them in relation to the control. This new version is orthogonal to the control variable x_2. We note these new variables as $x_\perp^j$ (j = 1, 3, 4, 5) to emphasize two things; first, that they correspond to variables x_j, and second, that they are orthogonal to x_2. As a result, we are eliminating the areas of ambiguous results (gray areas in Figure 3) by assuring orthogonality between the control and the variables of interest. For example, $x_\perp^5 = x_5 + 0.220x_2$, where 0.220 is estimated by regressing x_5 on x_2. Figure 5 displays the geometric representation of the calculation described above for x_5.

The dependent variable (y) was regressed on the orthogonal variables ($x_\perp^j$) for models A and B - we note them as model $A_\perp$ and model $B_\perp$. We denote the regression of y on $x_\perp^5$ as $y_\perp^5$. Note that y_{25} can be written as the sum of y_2 and $y_\perp^5$, that is $y_{25} = y_2 + y_\perp^5$ (we will use this fact later to compare the size of the effects). Equations 5 and 6 display the coefficients for the orthogonal variables.

$$y_{\perp}^{12345} = 0.551\, x_{\perp}^{1} - 0.303\, x_{\perp}^{3} + 0.241\, x_{\perp}^{4} - 0.926\, x_{\perp}^{5} \quad (5)$$

$$y_{\perp}^{5} = -0.517 x_{\perp}^{5} \quad (6)$$

Even though the coefficients in models A and B coincide with the coefficients in models $A_{\perp}$ and $B_{\perp}$ (refer to Table 2), this equivalence is not an identity. The residual variables have a smaller standard deviation given that they correspond to orthogonal projections. To make adequate comparisons in terms of the size of the associations between the control and the variables of interest, the latter must be re-parametrized based on the variables' standard deviations.

In model $B_{\perp}$ we have that: $\mathrm{Var}(x_{\perp}^{5}) = \sqrt{\left(1 - (0.220)^2\right)} = 0.975$. The re-parametrized version ($z_{\perp}^{5}$) is computed in equation 7:

$$z_{\perp}^{5} = \frac{x_{\perp}^{5}}{0.975} \quad (7)$$

Solving for $x_{\perp}^{5}$, we have $x_{\perp}^{5} = 0.975\, z_{\perp}^{5}$. Using this expression in equation 6 we have $y_{\perp}^{5} = -0.504\, z_{\perp}^{5}$. Since $y_{25} = y_2 + y_{\perp}^{5}$ we can rewrite model B as:

$$y_{25} = 0.425 x_2 - 0.504 z_{\perp}^{5} \quad (8)$$

Now the coefficients are comparable. The ratio $-0.504 / 0.425 = 1.19$ implies that the residual association between inflation and the multiplicative independence of the central bank ($x_{\perp}^{5}$) is bigger than the association between inflation and the control variable. This conclusion coincides with that of the authors, but there is no reason to expect this coincidence in all cases.

The same procedure can be applied to model A. Using the expressions for the orthogonal decomposition and standardizing the coefficients based on the variable's standard deviations, we present in equation 9 the final version of model A.

$$y_{12345} = 0.425\, x_2 - 0.577 z_{\perp}^{1345} \quad (9)$$

Where $z_{\perp}^{1345}$ represents a reduced variable of liberalization. The ratio between the sizes of the effects ($0.517 / 0.425 = 1.21$) is again larger than 1, which reinforces the above conclusion.

Conclusions and discussion

Stability of coefficients in bivariate and multivariate models is a crucial aspect to perform rigorous research in social sciences given the growing nature of large and diverse data sets. Evidence presented here suggests that both the number of predictors and the correlation among them have a direct impact on the stability of the estimates. The more variables are included, the higher the risk of having unstable estimates. The same can be said about the correlation among predictors: the higher the correlation among independent variables, the more erratic the coefficients would be. Parsimonious models shall be preferred as opposed to saturated ones. Even though we only presented one application, we believe our results can be generalized to contexts where analyses

are conducted on observational data to the extent that the variables of interest and the control variables may be correlated.

In that scenario, several and correlated predictors may lead to ambiguous results, i.e. results where the direction of the relationship between the predictors and the outcome can be reversed when comparing bivariate and conditional associations. It is unlikely that a single theory could account for all potential changes across a large set of predictors (say, more than three). The ratio between the conditional association and the bivariate one can be used to measure the stability of the coefficients – we termed this ratio *structural effect*. Although there is not a clear threshold for this ratio, either a reversal (change in the sign) or a substantial change in the estimates can be taken as signals of redundant information in the model. Starting from bivariate models and analyzing the structural effects is a good methodological practice in the process of variable selection. Moreover, reporting the bivariate association in a comparative fashion with respect to the conditional ones is key to assess results from multivariate regression analysis. There may also be potential paradoxical situations regarding the statistical significance of the coefficients, that is, substantial changes in the p-value. However, we leave that discussion for another work.

Geometric representation served the purpose of illustrating analytically the causes of these paradoxical results. Bivariate and multivariate regression can be represented as orthogonal and oblique projections. These representations help us in the variable selection process insofar as they display both the regression outputs and the multiple correlations among predictors simultaneously. We showed that areas of ambiguous results can be relevant if the correlation among the independent variables is large. Furthermore, in a model with several predictors these areas become difficult to assess visually, and the evaluation of structural effects is crucial. In sum, reporting bivariate and conditional associations along with structural effects can help us to assess our conclusions when using multivariate linear modeling techniques.

These suggestions are relevant for any research that uses linear models on observational data, broadly defined as data sets in which the marginal distribution of the independent variables is not controlled beforehand, i.e. when the correlation among predictors is a feature of the data on its own.[5] Overlooking this aspect may lead to ambiguous results as the inclusion of an additional control (or its omission) can produce two things: (1) a substantial change in the magnitude of the associations, (2) a change in their sign. If both associations are reported, a direct comparison can be made between them. Moreover, by comparing *structural effects* one can assess the extent to which the association between the variables of interest and the outcome is driven by the structure of the data, rather than by a true link between the outcome and the predictors. We believe these recommendations speak in particular to quantitative social scientists since most of the data we use are survey data, where several control variables are available and pertinent. A similar reflection ought to be done about *event-history* and *multilevel models* (Courgeau and Lelièvre, 1997; Courgeau, 2003). These two types of models constitute improvements to ordinary linear models as defined here as they account for features such as right and left censoring and interdependence among levels; yet they also share important characteristics with the former (Dobson and Barnett, 2008).

Residual regression is an alternative to avoid areas of ambiguous results. Geometric representations show that results from this approach need to be re-parametrized due to

the reduction of the variance in the orthogonal version of the predictors. Taking this into account yields either more convincing evidence to the question at hand, as in the case of our example, or can point out the necessity of more parsimonious models. This result suggests that caution ought to be exercised when using multivariate regression models.

All in all, we feel confident claiming that linear models should be used carefully when applied to research within the social sciences. Despite the attractiveness of including several controls, this practice should be, in principle, avoided or at least be accompanied by a systematic analysis of the *structural effects*. From this perspective, linear modeling appears as a tool to explore data structure – i.e. multivariate associations among outcomes and predictors – rather than a mechanism that explains social phenomena by itself. Our analysis confirms the two-way relationship between theory and methodology – which are often presented as separate matters. No theory can exist without rigorous data analysis that supports it, and all empirical analysis need to be theoretically informed. Consequently, explanation in the social sciences is not a matter of the statistical tool at hand, but a matter of making the appropriate connections between theory and methods.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support provided for Andrés' doctoral studies by the Fulbright Commission and the Population Studies Center at the University of Pennsylvania is gratefully acknowledged.

Notes

1. This discussion overpasses the scope of the present article for which we have limited the number of references to the articles that we are more familiar with.
2. We are particularly indebted to Henry Rouanet who significantly contributed to this article until his passing, and whose work has been a major source of inspiration to the development of these ideas both technically and theoretically. His books, articles and web publications continue to be an important source for academic reflection in this area. See: http://www.math-info.univ-paris5.fr/~lerb/rouanet/index.html.
3. The entire list of models is available upon request as part of the supplemental materials of this work.
4. Although none of the estimates should be larger than one in absolute value, six models yield results that violate this rule (x-markers in the bottom-right area of the plot). These six models include variables x_1 and x_5. The strong correlation among these two variables (0.931) affects the level of precision of the estimates which are also affected by numerical approximations carried on by the statistical software.
5. We are aware of techniques such as Randomized Control Trials, Regression Discontinuity, Propensity Score Matching and Instrumental Variable, which by construction rule out the issue of correlation among predictors; yet there are several areas of research where these techniques cannot be implemented.

References

Abbot A (1988) Transcending General Linear Reality. *Sociological Theory* 6(2): 169-186.

Bernard J, Le Roux B and Rouanet H (1989) L'analyse des données multidimensionnelles par le langage d'interrogation des données (lid): au-delà de l'analyse des correspondances. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 23: 3-46.

Bourdieu P (2005) *The Social Structures of the Economy* (first edition). Cambridge: Polity Press.

Bourdieu P and Wacquant L (1992) *An Invitation to Reflexive Sociology* (second edition). Cambridge: Polity Press.

Box GEP (1976) Science and Statistics. *Journal of the American Statistical Association* 71(356): 791-799.

Bruno AS (2010) Analyser les écarts de salaires à l'aide des modèles de régression. Vertus et limites d'une méthode. Le cas des migrants de Tunisie en région parisienne après 1956. *Histoire & Mesure* XXV(2): 121-156.

Bry X, Robette N and Roueff O (2016) A Dialogue of the Deaf in the Statistical Theater ? Adressing Structural Effects within a Geometric Data Analysis Framework. *Quality and Quantity* 50(3): 1009-1020. doi: 10.1007/s11135-015-0187-z.

Castro Martín MT and Juárez F (1995) The Impact of Women's Education on Fertility in Latin America: Searching for Explanations. *International Family Planning Perspectives* 21(2): 52-57.

Cornwell B (2015) *Social Sequence Analysis: Methods and Applications*. doi: 10.1017/CBO9781316212530.

Courgeau D (2003) (ed.) *Methodology and Epistemology of Multilevel Analysis. Approaches from Different Social Sciences*. Dordrecht: Springer.

Courgeau D and Lelièvre E (1997) Changing Paradigm in Demography. *Population* 9: 1-10.

Cukierman A, Miller GP and Neyapti B (2002) Central Bank Reform, Liberalization and Inflation in Transition Economies – An International Perspective. *Journal of Monetary Economics* 49(2): 237-264.

Darras (1966) *Le partage des bénéfices. Expansion et inégalité en France*. Paris: Editions de Minuit.

De Melo M, Denizer C and Gelb A (1996) From Plan to Market: Patterns of Transition. *Policy Research Working Papers* 1564: World Bank.

Deauvieau J (2010) Comment traduire sous forme de probabilités les résultats d'une modélisation logit? *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 105(1): 5-23.

Deauvieau J (2011) Est-il possible et souhaitable de traduire sous forme de probabilités un coefficient logit? Réponse aux remarques formulées par Marion Selz à propos de mon article paru dans le *BMS* en 2010. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 112(1): 32-42.

Desrosières A (2001) Entre réalisme métrologique et conventions d'équivalence: les ambiguités de la sociologie quantitative. *Genèses* 2(43): 112-127.

Desrosières A (2008) Analyse des données et sciences humaines: comment cartographier le monde social? *Journal Electronique d'Histoire des Probabilités et de la Statistique* 4(2): 21.

Dobson AJ and Barnett AG (2008) *An Introduction to Generalized Linear Models* (first edition) New York: Chapman and Hall.

Gollac M (2004) La rigueur et la rigolade. À propos de l'usage des méthodes quantitatives par Pierre Bourdieu. *Courrier des statistiques* 1(112): 29-36.

Hirschman C (1994) Why Fertility Changes. *Annual Review of Sociology* 20: 203-233.

Ho JY and Elo IT (2013) The Contribution of Smoking to Black-White Differences in US Mortality. *Demography* 50(2): 545-568.

Johnson-Hanks J, Bachrach CA, Morgan SP and Kohler HP (2011) *Understanding Family Change and Variation: Toward a Theory of Conjunctural Action*. Dordrecht: Springer.

Lebaron F (2000) *La croyance économique – les économistes entre science et politique* (first edition). Paris: Seuil.

Lebaron F and Le Roux B (2015) *La méthodologie de Pierre Bourdieu en action: espace culturel, espace social et analyse des données*. Paris: Dunod.

Lebart L Morineau A and Piron M (1997) *Statistique exploratoire multidimensionnelle* (second edition). Paris: Dunod.

Le Roux B and Rouanet H (2004) *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Dordrecht-Boston-London: Kluwer Academic Publisher.

Lieberson S and Horwich J (2008) Implicaction Analysis: A Pragmatic Proposal for Linking Theory and Data in the Social Sciences. *Sociological Methodology* 38(1): 1-100.

McCullagh P (1984) Generalized Linear Models. *European Journal of Operational Research* 16(3): 285-292.

Morgan M (1990) *The History of Econometric Ideas* (first edition). London: Cambridge University Press.

Mosteller F and Tukey J (1977) *Data Analysis and Regression: A Second Course in Statistics*. Reading: Addison-Wesley.

Nelder JA and Wedderburn W (1972) Generalized Linear Models. *Journal of the Royal Statistical Society* 135(3): 370-384.

Ollion E (2011) De la sociologie en Amérique. Éléments pour une sociologie de la sociologie étasunienne contemporaine. *Sociologie* 2(3): 277-294.

Preston S, Guillot M and Heuveline P (2001) *Demography: Measuring and Modeling Population Processes*. Malden, MA: Blackwell.

Rouanet H, Ackerman W, Lebaron F, Le Hay V and Le Roux B (2002) Régression et analyse géométrique des données: réflexions et suggestions. *Mathématiques et Sciences Humaines* 160: 13-45.

Rouanet H, Ackerman W and Le Roux B (2000) The Geometric Analysis of Questionnaires: The Lessons of Bourdieu's Distinction. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 65: 5-18.

Savage M, et al. (2013) A New Model of Social Class? Findings from the BBC's Great British Class Survey Experiment. *Sociology* 47(2): 219-250.

Selz M and Deauvieau J (2011) Pourquoi traduire sous forme de probabilités les résultats d'une modélisation logit?: réaction à J. Deauvieau (BMS 2010) Première réponse à Marion Selz. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 111: 76-79.

Vallet LA and Caille JP (1996) Les élèves étrangers ou issus de l'immigration dans l'école et le collège français. Une étude d'ensemble. *Les Dossiers d'Éducation et Formations* 67.