CSAE Working Paper WPS/2016-31

Beyond the Stars: a New Method for Assessing the Economic Importance of Variables in Regressions

Olivier Sterck*

November 2016

Abstract

Economists lack a systematic method to assess the economic importance of effects in regressions. In this article, I use experimental evidence to show that for a large majority of economists, the economic importance of an explanatory variable refers to its contribution to deviations in the level of the dependent variable. Existing statistics, such as standardized beta coefficients and the partial or semi-partial r^2 and r, are only imperfect measures of the economic importance of explanatory variables: these statistics do not match with economists' common understanding of economic importance and are difficult to interpret. I therefore develop a new method, which consists in rescaling standardized beta coefficients such as to obtain the percentage contribution of each explanatory variable to deviations in the dependent variable. As an illustration, the method is applied to the study of the causes of long-run economic development.

*Correspondence: University of Oxford, Centre for the Study of African Economies, Manor Road, OX1 3UQ, United Kingdom, Phone: +32 473 37 50 00, Email: olivier.sterck@economics.ox.ac.uk. I thank Stefan Dercon, Arnaud Dufays, James Fenske, Natalie Quinn, Simon Quinn, Max Roser, Maxime Taquet, Stefan Thewissen and participants at the CSAE research workshop for useful comments and discussions. I am also grateful to all respondents to the online survey.

1 Introduction

A controversy erupted two decades ago when McCloskey and Ziliak (1996) highlighted that 70% of articles published in the American Economic Review in the 80s did not distinguish economic from statistical significance. One decade later, the same authors observed little change in this trend: 82% of papers published in the American Economic Review in the 90s *"mistook a merely statistically significant finding for an economically significant finding"* (Ziliak and McCloskey 2004).[1] The practice in economics has considerably changed since then. It is now frequent to hear in economic seminars or read in academic papers that an effect is "economically significant" or "economically important". The number of peer-reviewed articles discussing the "economic significance" or "economic importance" of their findings has considerably increased in the last decade (figure 1).

This trend is of course more than welcome for the credibility of the field. In this paper, I however argue that economists do not have a clear understanding of what economic importance means, and lack a systematic method corresponding to this understanding. The contribution of this research is fourfold. First, I clarify the common understanding of economic importance using survey data collected from economists who participated to an online experiment. I distinguish the concepts of absolute economic importance of a variable, which refers to its contribution to deviations in the level of the dependent variable, and relative economic importance, which is the ratio of the absolute economic importance of two explanatory variables in a regression.

Second, I show that the existing tools to assess the economic importance of each variable in linear regression model are highly imperfect. In particular, I argue that partial r^2, semi-partial r^2 and similar methods are inconvenient statistics, both because they relate to the variance of the dependent variable instead of the level of the dependent variable, and because they underestimate the real contributions of variables which are not orthogonal. Standardizing variables is a better tool to assess the relative economic importance of variables. Still, this method is imperfect to assess their absolute economic importance because the sum of coefficients of standardized beta coefficients does not add-up to a constant. In particular, the sum of standardized beta coefficients can be much higher than 1 in models with numerous explanatory variables and high R^2. This creates a risk of overestimating the economic importance of explanatory variables when interpreting results from standardized regressions. Experimental evidence confirms that economists are likely to misinterpret the importance of standardized beta coefficients.

[1] See e.g. Hoover and Siegler (2008) and Engsted (2009) for contrasting views. This paper does not aim to enter the debate on *significance* and the use of *p-values*, but rather to propose a new method for assessing the economic importance of variables of interest. This paper does not stand in favor or against the previously cited papers.

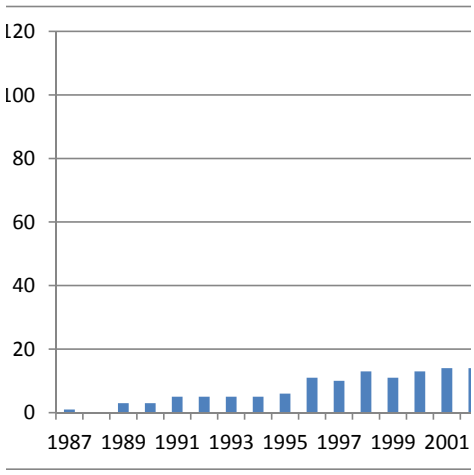


Figure 1: Results of a search in the Econlit bibliographic database using Proquest, with the query "economically significant" OR "economically important" (only peer-reviewed journals were considered).

Third, I develop a new method to measure the absolute and relative economic importance of variables in linear regression models. The method simply proposes to rescale the coefficients of a standardized regression by a certain coefficient such that the contributions of each variable and of the residuals add up to one. As outcome, the method provides the proportion of deviations in the dependent variable which is explained by each right-hand side variable and by the residuals. These statistics are expressed in percentage of the dependent variable, making them particularly straightforward to interpret.

Finally, the method is applied to the study of long-run causes of economic development. Results show that economic development is explained by a multitude of factors, from geographical circumstances (average temperature, distance to the nearest ice-free, the percentage of fertile soil, and terrain ruggedness), and historical conditions (slave trade intensity and state history), to the composition of countries' population (ethnic inequality, the percentage of the population of European descent, and genetic distance). None of these factors explain more than 11 percent of deviations in the log of GDP per capita.

The paper proceeds as follows. Section 2 clarifies the concepts of absolute and relative economic importance for economists. Section 3 reviews the existing methods to evaluate the contribution of explanatory variables in linear regression models. Section 4 presents the new method. The properties and limits of the new methods are discussed in Section 5. An application of the new method to the study of the causes of long-run growth is proposed in Section 6. Section 7 concludes the study.

2 Experiment 1: what does "economic importance" mean?

Researchers may pursue two objectives when assessing the economic importance of explanatory variables. First, they may seek to compare the relative effect size of different explanatory variables. For example, a researcher may wonder if it is wealth or distance to a healthcare center which is the best predictor of health outcomes and compare their impact. Second, they may be interested to know the absolute economic importance of an explanatory variable. Using the same example, a researcher may be interested to know the proportion of (changes in) the dependent variable - here health outcomes - which is explained by variables of interest, here wealth and the distance to a health care center.

The recent literature abounds with statements about the "economic significance" or "economic importance" variables of interest. Such statements are usually preceded or followed by an analysis of the predicted impact of an increase in the variable of interest on the dependent variable. The increase in the variable of interest is often expressed in standard deviation terms, and the impact on the dependent variable is either expressed in level, in percentage of the mean or in standard deviation terms. In the long-run growth literature for example (the focus of the application in Section 6), Spolaore and Wacziarg (2009) report standardized beta coefficients as *"a measure of the magnitude of the coefficients"*, noting that *"one standard deviation in FST genetic distance between plurality groups accounts for 16.79% of a standard deviation of income differences."* Similarly, Michalopoulos (2012) reports standardized coefficients to *"facilitate comparison of the quantitative effect across different specifications and across regressors."* Noting that *"a one-standard-deviation increase in the variation of elevation and a similar increase in the variation of land quality augments linguistic diversity by 0.31 and 0.34 standard deviations"*, the author concludes that *"these economically important findings reveal the geographic origins of contemporary ethnolinguistic diversity."* Similarly, in a section entitled *"Economic Magnitude of the Effects"*, Nunn and Puga (2012) report standardized coefficients to prove that *"the differential effect of ruggedness is statistically significant and economically meaningful."* In turn, Alesina et al. (2015) discuss the importance of their findings by noting that the *"standardized beta coefficient of the ethnic inequality index is around 0.20–0.30, quite similar to that of the works on the role of institutions on development (e.g., Acemoglu, Johnson, and Robinson 2001)."* The economic literature is replete with similar statements.

These quotes refer to the absolute economic importance of a variable, which seems to be understood as the size of its *ceteris paribus* impact on the dependent variable. Contrary to the theoretical literature on variables importance, which mostly focuses on

variance decomposition[2], the economic importance seems to refer to the proportion of deviations in the dependent variable which is explained by the explanatory variable, and not to the share of the explained variance.

To confirm this observation, I ran an online experiment among 207 economists.[3] The online questionnaire included the following problem:

"Three normal variables x_1, x_2 and x_3 were randomly generated following a normal distribution N(0,1). These variables are uncorrelated. They are combined as follows to generate the variable y: $y = c_0 + c_1x_1 + c_2x_2 + c_3x_3$. We assume that y, x_1, x_2 and x_3 are variables that are of high interest for economists and policy-makers (e.g. welfare, wages, health, age, etc.). A linear regression ($R^2 = 1$) of y on x_1, x_2 and x_3 can be used to obtain the coefficients c_1, c_2 and c_3. In percentage, how would you qualify the economic importance of the effect of x_1, x_2 and x_3 on y?"

I distinguish three possible answers to the question, which correspond to three understandings of the concept of economic importance. The economic importance of x_i may refer to (i) its contribution to the variance of the dependent variable $c_i^2/({c_1}^2+{c_2}^2+{c_3}^2)$, (ii) its effect compared to the mean of the dependent variable c_i/c_0, and (iii) its contribution to deviations in the dependent variable $|c_i|/(|c_1|+|c_2|+|c_3|)$.

Two vectors of coefficients c_i were randomly assigned to participants.[4] Half of the sample was randomly assigned to the vector $(0, 1, 5, 2)$, which allows relatively easy computation for the interpretations (i) and (iii), but may be confusing for interpretation (ii). The other half was randomly assigned to the vector $(6, 1, 5, 2)$, which allows straightforward computation for the three interpretations.

The results of this experiment are presented in figure 3.[5] When facing the first vector, a large majority of respondents (66%) of offered responses which are consistent with the interpretation (iii), confirming the hypothesis that most economists interpret the economic importance of a variable as its contribution to deviations in the dependent variable. Other respondents either did not know what to answer (27%), or had in mind the interpretation (i), on variables' contribution to the variance of the dependent variable (7%).

[2]See e.g. Bi (2012) or Grömping (2015) for reviews.

[3]The sample consists in 207 economists (1) who are current or former colleagues from the University of Oxford, the University of Namur or the Catholic University of Louvain, or (2) who participated at the 2015 or 2016 CSAE conferences at the University of Oxford. Respondents where contacted by email.

[4]I chose the coefficients such that the three interpretations require a similar level of effort for someone well-trained in econometrics. Furthermore, to facilitate the task, respondents were advised to enter a fraction (for example, 1/2 if they thought the answer was 50 percent).

[5]14% of responses could not be classified and were excluded. Most of these responses argue that the economic importance of the variable x_i is 33% , or is equal to its regression coefficient c_i.

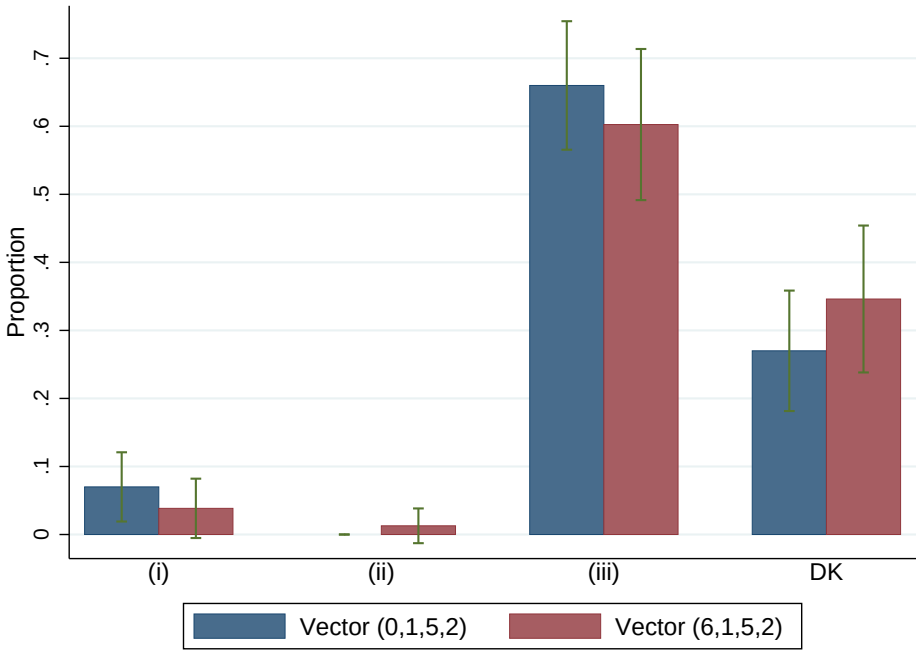


Figure 2: Results of experiment 1 ((i) to (iii) = response in line with interpretations (i) to (iii), DK = do not know).

With the second vector $(6, 1, 5, 2)$, 60% of respondents offered responses corresponding interpretation (iii).[6] Only 4% of responses corresponded to the interpretation (i), and 1% of responses corresponded to the interpretation (ii). As much as 35% of respondents replied that they could not reply to the question.

The online survey also included a question about the relative economic importance of explanatory variables. For this question, respondents are almost unanimous, and answers do not vary significantly across the two vectors of coefficients. When asked to assess the relative economic importance of x_1 versus x_2, 69% of respondents answered 5/1, which corresponds to the ratio c_1/c_2. This answer is consistent with interpretations (ii) and (iii). Only 5% of respondents answered 25/1. This latter answer compares variables' contributions to the variance of the dependent variable, and therefore pertains to the interpretation (i). These results demonstrate that the relative economic importance refers to the ratio of contributions to deviations in the dependent variable, and not to the ratio of contributions to its variance.

In conclusion, economists understand the absolute economic importance of a variable

[6]Perhaps surprisingly, a few respondents seem to have been confused when facing this vector. Eleven respondents replied that the contributions of variables are 1/14, 5/14 and 2/14 respectively; this corresponds to $|c_i| / (|c_0| + |c_1| + |c_2| + |c_3|)$. This response is close to interpretation (iii), and was classified as such; however, this understanding of economic importance is problematic because it is highly sensitive to a rescaling of the dependent variable (for degree Celsius to Kelvin for example).

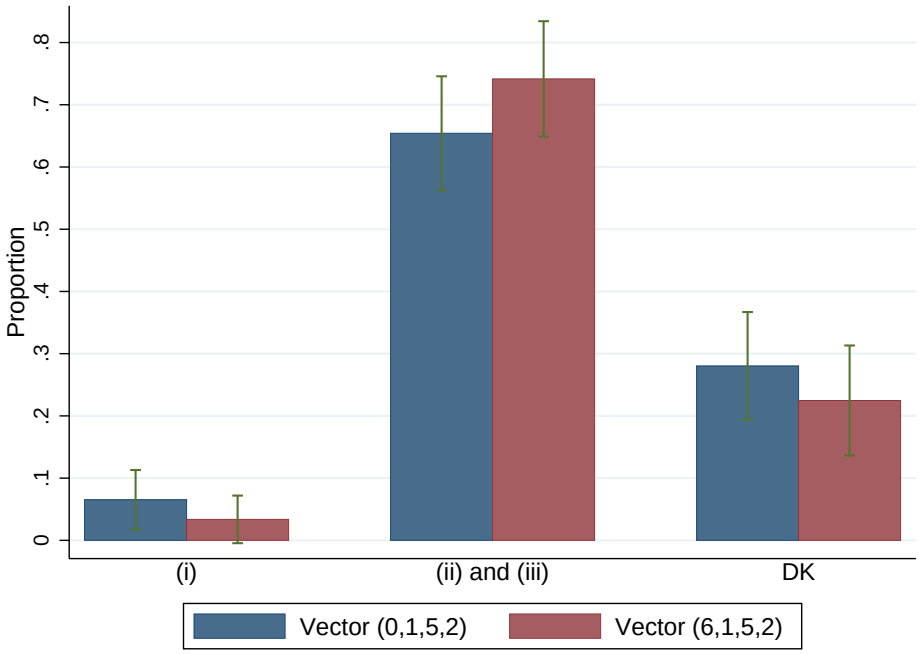


Figure 3: Results of experiment 1 ((i) to (iii) = response in line with interpretations (i) to (iii), DK = do not know).

as its percentage contribution to deviations in the dependent variable. This interpretation makes sense when it comes to policy implications: policy-makers are usually interested to influence the level of the dependent variable (and not its variance). The relative economic importance of two variables refers to the ratio of these contributions expressed in a same unit of measurement. The next section therefore examines if existing methods are consistent with this understanding.

3 A critique of existing methods

This section reviews the properties the two main methods used by economists to evaluate economic importance of explanatory variables in linear regression model: (1) the regressions with standardized variables, and (2) the partial and semi-partial r^2 and r.[7]

[7]Holgersson et al. (2014) recently rediscovered Achen (1982), and proposed using mean value decomposition to assess economic significance of variables. Assuming a model $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \epsilon$, the economic importance of variable x_i is defined by $\beta_i \frac{\bar{x_i}}{\bar{y}}$. This method is not appropriate because (1) it leads to negative values and, most importantly, because (2) the economic importance of variables can dramatically change when variables are rescaled, which is highly problematic. For example, it would be counter-intuitive if a research on the impact of weather on economic activity concludes that the economic importance of temperature is different when temperature is expressed in degrees Celsius compared to Kelvin or Fahrenheit. Similarly, the economic importance of a variable centered at zero is null according to their method, even if its coefficient is economically large and statistically significant. These flaws invite caution when using this method and interpreting its results.

3.1 Standardized regressions

Standardizing variables is the method which is the most widely used to interpret the economic importance of explanatory variables. I use a simple example to explore the properties of this method.[8] Suppose that the variable y is the sum of $n+1$ independent variables $x_1, ..., x_n$ and ϵ that are normally distributed:

$$y = \sum_{i=1}^{n} \beta_i x_i + \epsilon, \text{where } x_i \sim N(0, \sigma_i^2) \text{ and } \epsilon \sim N(0, e^2) \tag{1}$$

A simple regression of y on $x_1, ..., x_n$ gives unbiased estimates of regression coefficients β_i. As such, these estimates do not give information on the economic importance of $x_1, ..., x_n$.

The standardized transformations of $x_1, ..., x_n$ and ϵ are denoted with an asterisk. It is straightforward that $x_i^* = x_i/\sigma_i \sim N(0,1)$ and $\epsilon^* = \epsilon/e \sim N(0,1)$. The variable y is distributed as a normal variable $y \sim N(0, e^2 + \sum_{i=1}^{n}(\beta_i\sigma_i)^2)$, implying that $y^* = y/\sqrt{e^2 + \sum_{i=1}^{n}(\beta_i\sigma_i)^2} \sim N(0,1)$. Denoting $c_i = \beta_i\sigma_i$, and dividing each side of equation (1) by $\sqrt{e^2 + \sum_{i=1}^{n} c_i^2}$, we obtain:

$$y^* = \sum_{i=1}^{n} \frac{c_i}{\sqrt{e^2 + \sum_{i=1}^{n} c_i^2}} x_i^* + \frac{1}{\sqrt{e^2 + \sum_{i=1}^{n} c_i^2}} \epsilon \tag{2}$$

The coefficients of a regression of y^* on $x_1^*, ..., x_n^*$ therefore gives an unbiased estimate of coefficients b_i^*:

$$b_i^* = \frac{c_i}{\sqrt{e^2 + \sum_{i=1}^{n} c_i^2}}. \tag{3}$$

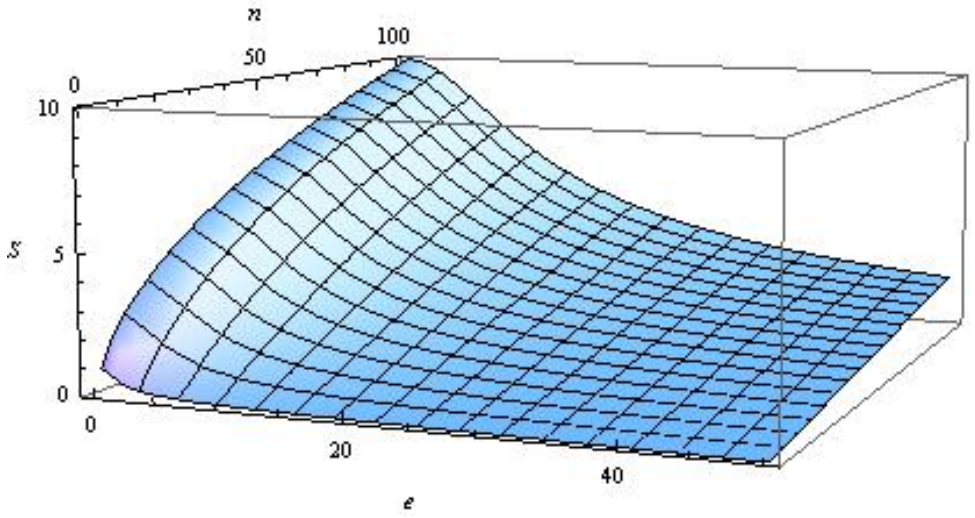
The coefficients of such *standardized regression* are commonly referred to as *standardized beta coefficients*.

Are the coefficients b_i^* useful to assess the economic importance of the explanatory variables x_i? Yes and no. On the one hand, the ratios of standardized beta coefficients allow comparing the relative importance of explanatory variables. It is indeed straightforward that $b_i^*/b_j^* = c_i/c_j = (\beta_i\sigma_i)/(\beta_j\sigma_j)$.

On the other hand, b_i^* is not a measure of the absolute economic importance of x_i. It is indeed not a direct measure of the proportion of deviations in the dependent variable which is explained by x_i. Contrary to the analysis of Section 2, the denominator of equation (3) is the sum of the square of the coefficients c_i and e^2 instead of the simple

[8] See e.g. King (1986) and Luskin (1991) for other discussions on the merits and problems of standardization.

Figure 4: Sum S of coefficients of standardized variables as a function of the number of variables n and contribution e of the residuals ($c_i = 1 \forall i$).



sum of the coefficients c_i and e. What is also worrying is that the sum of the coefficients b_i^* can be much larger than 1.

$$S = \sum_{i=1}^{n} b_i^* = \frac{\sum_{i=1}^{n} c_i}{\sqrt{e^2 + \sum_{i=1}^{n} c_i^2}} \tag{4}$$

It follows from equation (4) that the sum S can be very high if the error term is relatively small and if the number of explanatory variables n is large. This relationship is illustrated in figure 4 for the case $c_i = 1$. This implies that the coefficients of a standardized regression cannot be interpreted as the percentage contribution of explanatory variables to deviations in the dependent variable. Researchers missing this property face the risk of overestimating the economic importance of their variables of interest. This risk is particularly important when the R^2 and the number of explanatory variables are large.

It is straightforward to extend this analysis to the study of non-independent and non-normal distributions. Keeping the same notation[9], and denoting ρ_{ij} the correlation between x_i^* and x_j^*, a standardized regression of y^* on the variables x_i^* gives unbiased estimates of coefficients b_i^*:

$$b_i^* = \frac{c_i}{\sqrt{e^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} c_i c_j}}. \tag{5}$$

[9]The DGP is $y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$ where σ_i^2 is the variance of x_i, and $c_i = \beta_i \sigma_i$. The error term ϵ is assumed to be independent.

As before, the sum of the coefficients b_i^* can be large, especially if R^2 and the number of explanatory variables are large, and if the correlation between explanatory variables is low or negative.

3.2 Experiments 2 and 3: how do economists interpret standardized regressions?

The online experiment included a module examining how economists interpret the results of a regression.[10] Respondents were randomly assigned to two groups: half of them only received the results of a standard OLS regression (column 1 of table 1), while the other half also received the results of a standardized regression (column 2 of table 1). For the sake of comparison, column 3 reports estimates of the absolute economic importance of variables as calculated using the new method expounded in Section 4. This last column was not shown to respondents.

The question was asked as follows, with the information in square brackets only provided to the second group:

"Please find below the results of a simple OLS regression (column 1) [as well as the results of the same regression where all variables were standardized (column 2). Standardization consists in subtracting the mean of the variable and dividing the result by the standard deviation of the variable. This process was done for all variables (y, x_1, x_2, x_3, ...)]. It is assumed that all variables (y, x_1, x_2, ...) are of high interest for economists and policy-makers (e.g. welfare, wages, health, age, etc.). The table only shows results for 5 variables, but 45 supplementary control variables are included in the regression (and also standardized). Please note that the regression is well-specified, and without any statistical bias in the estimated coefficients and standard errors. Coefficients can be interpreted as causal. Variables are normally distributed with different variances and means. Standard errors are reported in parentheses. In percentage, how would you qualify the economic importance of the effect of x_1 on y? And of x_2 on y?"

Results of this experiment are presented in figure 5.[11] There is both good and bad news. On the one hand, a large majority of respondents from the first group (77%) correctly noted that they could not assess the absolute economic importance of variables when only OLS results were available. On the other hand, as much as 40% of respondents from the second group wrongly interpreted the results of the standardized regression in

[10] The data generating process of this regression is: $y = 0.1x_1 + 0.3x_2 + 0.2x_3 + 2x_4 - 1x_5 + 1.5x_6 - 0.2x_7 + 0.1x_8 - 0.5x_9 + 0.2x_{10} + \sum_{i=11}^{50} x_i + \epsilon$, where the x_i and ϵ are normally distributed with $\sigma_1 = 140$, $\sigma_2 = 3.3$, $\sigma_3 = 20$, $\sigma_4 = 5.45$, $\sigma_5 = -0.1$, $\sigma_6 = 2.6$, $\sigma_7 = -10$, $\sigma_8 = 1$, $\sigma_9 = -5.8$, $\sigma_{10} = 5$, $\sigma_i = 1 \forall i \in 11...50$, $\sigma_e = 20$.

[11] 12% of responses could not be classified, and were excluded.

	(1) Basic OLS	(2) Standardized OLS	(3) Contribution
x_1	0.099***	0.488***	14.3 %
	(0.003)	(0.014)	
x_2	0.324***	0.039***	1.1 %
	(0.118)	(0.014)	
x_3	0.213***	0.153***	4.4 %
	(0.020)	(0.014)	
x_4	2.056***	0.391***	11.9 %
	(0.074)	(0.014)	
x_5	-0.156	-0.001	0.0 %
	(3.928)	(0.014)	
Controls (45 variables)			48.0%
Residuals			20.6%
Sum contributions			100%
Controls (45 variables)	Yes	Yes	Yes
Observations	2583	2583	
R^2	0.504	0.504	

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1: Results of OLS regressions proposed to the participants of the experiment

percentage terms, and 17% interpreted the results of the OLS regression in percentage terms. When asked about the relative economic importance of variables (results not shown), 63% of respondents from the first group (only OLS) reported they could not answer the question, while 37% wrongly interpret the ratio of OLS coefficients. A majority (55%) of respondents from the second group correctly inferred the relative economic importance of variables from the ratio of standardized coefficients.

The online questionnaire also asked respondents *"What is the sum of regression coefficients in a regression where all variables are standardized (y, x_1, x_2, x_3, ...)?"* Results of this experiment are also mixed (figure 6). Only 2% of respondents could reply correctly to the question. A short majority (54%) recognized that they could not answer the question. Worryingly, 34% of respondents answered that the sum is equal to 1, while the rest replied that the sum is equal to 0 or to the r^2. Overall, the results of these two experiments confirm the concern that was raised above: economists are often tempted to interpret the coefficients of standardized regression in percentage terms.

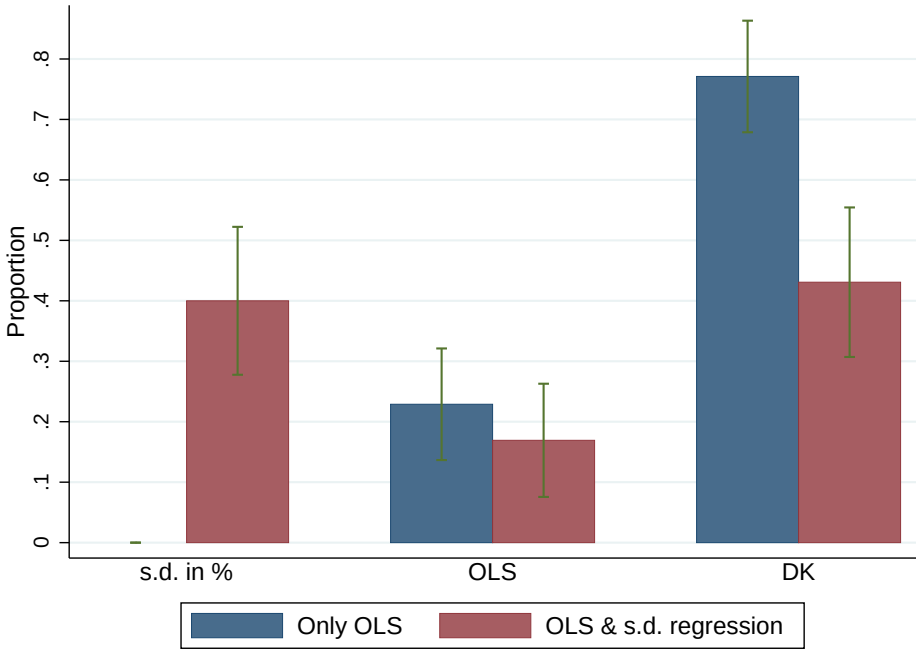


Figure 5: Results of experiment 2 (s.d. in percentage: coefficients of standardized regression are interpreted as percentage, OLS: coefficients of OLS regression are interpreted as percentage, DK = do not know).

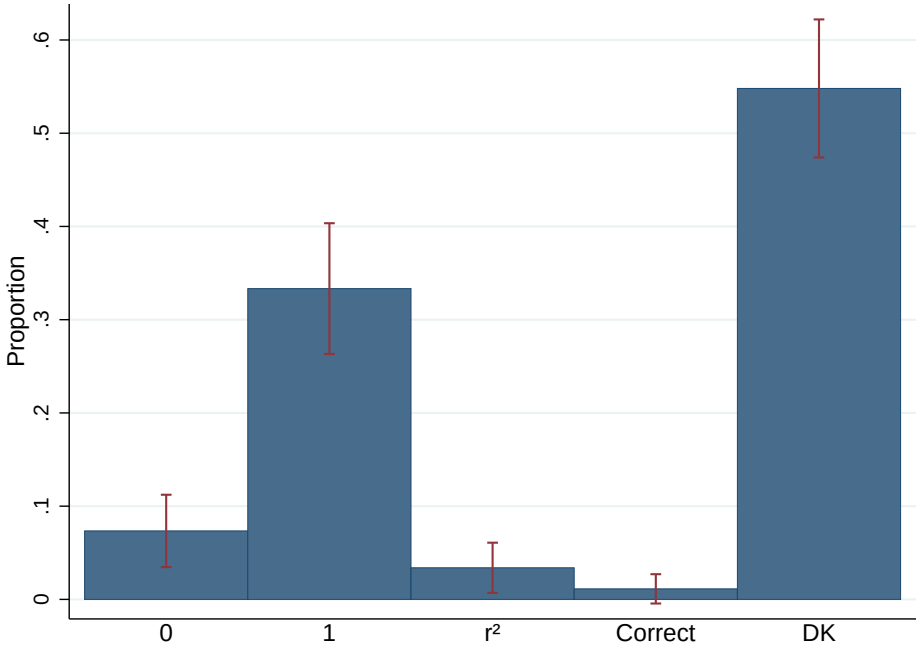


Figure 6: Results of experiment 3 (DK = do not know).

3.3 Partial r^2 and semi-partial r^2

Several methods aiming at decomposing the R^2 or the variance of the dependent variable have recently been developed by scholars from different fields (see e.g. Grissom and Kim (2012), Bi (2012) or Grömping (2015) for reviews). While these statistics identify which variable best contributes to the fit of the regression, they are less useful to assess the economic importance of variables of interest. I focus here on partial r^2 and semi-partial r^2, two statistics which are sometimes used by economists to assess the strength of a relationship (see e.g. Ashraf and Galor (2013) or Nakamura and Steinsson (2014)).

The partial and semi-partial r^2 measure the share of the variation in the dependent variable which is explained exclusively by the variation in the explanatory variable of interest.

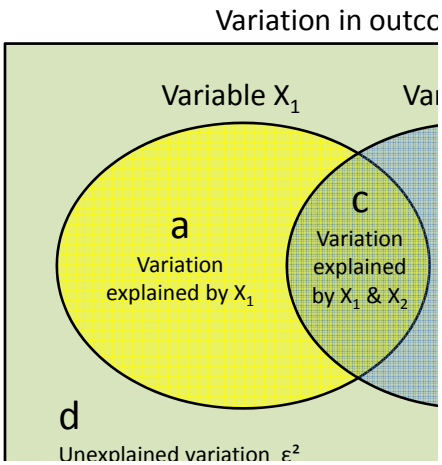
The partial r^2 compares the variation of interest to the sum of the unexplained variation and the variation of interest. In line with the notation of figure 7, the partial r^2 associated with the variable x_1 is the ratio $a/(a+d)$. It measures the mutual relationship between y and x_1 when other observable variables (x_2 here) are held constant with respect to the two variables involved (y and x_1 here).[12] In other words, it measures proportion of unexplained variation of y that becomes explained with the addition of x_1 to the model. The sum of partial r^2 is not an easily interpretable number. Even when the R^2 is large, the sum of partial r^2 can equal to 0 if variable are multicollinear. On the contrary, this sum can be much larger than 1 if explanatory variables are independent and if the R^2 is large.

In contrast, the semi-partial r^2 associated with a variable compares the variation explained exclusively by this variable to the total variation in the dependent variable. In line with the example in figure 7, the semi-partial r^2 associated with x_1 is the ratio $a/(a+b+c+d)$. It therefore captures the proportion of the variation in the dependent variable exclusively explained by x_1. The sum of semi-partial r^2 is equal to the R^2 if explanatory variables are perfectly orthogonal. This latter assumption is never satisfied in empirical applications, implying that the sum of semi-partial r^2 is always lower than the R^2 in practice.

The partial r^2 and semi-partial r^2 relate to the variance of the dependent variable, while Section 2 showed that economists are usually interested in deviations in the level of the dependent variable. The partial and semi-partial r offer a partial solution to this issue. However, as for standardized regressions, the sum of partial r, or semi-partial r, does not add-up to an easily interpretable number, implying that these statistics cannot

[12] In our example, the partial r^2 is given by the r^2 of a regression of the residuals of y with respect to x_2 on the residuals of x_1 with respect to x_2.

Figure 7: Decomposition of partial r^2 and semi-partial r^2



be used to estimate the absolute economic importance of each explanatory variable. Furthermore, the partial and semi-partial r ignores the variation correlated with different explanatory variables when these are not orthogonal. This property increases the risk of underestimating of the economic effect of variables that are not independent. These statistics should therefore not be used to assess the economic importance of explanatory variables.

4 A new method

The last section showed that standardized regressions are useful to compare the relative economic importance of explanatory variables, but are inappropriate to evaluate their absolute economic importance. In this section, I develop a simple and powerful method satisfying both objectives. Intuitively, the new method proposes to rescale standardized beta coefficients such as to obtain the percentage contribution of each explanatory variable to deviations in the dependent variable. I build on the simple example introduced in Section 3.1 and generalize the method afterward.

Equation (2) can be rewritten as follows:

$$y^* = \sum_{i=1}^{n} \underbrace{\frac{c_i}{\sqrt{e^2 + \sum_{i=1}^{n} c_i^2}}}_{b_i^*} x_i^* + \underbrace{\frac{e}{\sqrt{e^2 + \sum_{i=1}^{n} c_i^2}}}_{b_e^*} \epsilon^* \tag{6}$$

The absolute economic importance of x_i is measured by the following statistic:

$$\alpha_i \equiv \frac{|b_i^*|}{(|b_e^*| + \sum_{i=1}^n |b_i^*|)} = \frac{\left| \frac{c_i}{\sqrt{e^2 + \sum_{i=1}^n c_i^2}} \right|}{(\left| \frac{e}{\sqrt{e^2 + \sum_{j=1}^n c_j^2}} \right| + \sum_{j=1}^n \left| \frac{c_j}{\sqrt{e^2 + \sum_{j=1}^n c_j^2}} \right|)} = \frac{|c_i|}{|e| + \sum_{j=1}^n |c_j|} \tag{7}$$

The coefficient α_i measures the percentage contribution of the variable x_i to deviations in the dependent variable y.

Again, it is straightforward to extend this analysis to the study of non-independent and non-normal distributions. Keeping the same notation,[13] and noting that the coefficients b_i^* are given by equation (5) in the case of correlated variables, we have:

$$\alpha_i \equiv \frac{|b_i^*|}{(|b_e^*| + \sum_{i=1}^n |b_i^*|)} = \frac{\left| \frac{c_i}{\sqrt{e^2 + \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} c_i c_j}} \right|}{(\left| \frac{e}{\sqrt{e^2 + \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} c_i c_j}} \right| + \sum_{j=1}^n \left| \frac{c_j}{\sqrt{e^2 + \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} c_i c_j}} \right|)} = \frac{|c_i|}{|e| + \sum_{j=1}^n |c_j|} \tag{8}$$

The coefficient α_i and α_e can be obtained through following procedure:
- *Step 1:* Regress of $x_1, ..., x_n$ on y and predict the residuals of the regression ϵ.
- *Step 2:* Standardize the variables y, $x_1, ..., x_n$, and ϵ to obtain y^*, the x_i^* and ϵ^*.
- *Step 3:* Regress y^* on the x_i^* and ϵ^*. The coefficients of this regression are equal to the b_i^* and b_e^*.
- *Step 4:* Sum the absolute value of coefficients b_i^* and b_e^* to obtain: $|b_e^*| + \sum_{i=1}^n |b_i^*|$. The coefficients α_i are then given by: $\alpha_i = \frac{|b_i^*|}{(|b_e^*| + \sum_{i=1}^n |b_i^*|)}$.
- *Step 5:* The proportion of the dependent variable which is unexplained by $x_1, ..., x_n$ is given by: $\alpha_e = \frac{|b_e^*|}{(|b_e^*| + \sum_{i=1}^n |b_i^*|)}$.

The variance-covariance matrix associated with the coefficients α_i is obtained by:
- *Step 6:* multiplying x_i^* by $|b_e^*| + \sum_{i=1}^n |b_i^*|$ and doing a regression of y^* on these newly-generated variables.

It is important to note that the six steps procedure remains applicable when variables are not normally distributed and not orthogonal.

[13] The DGP is $y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \epsilon$ where σ_i^2 is the variance of x_i, $c_i = \beta_i \sigma_i$, and ρ_{ij} the correlation between x_i^* and x_j^*. The error term ϵ is assumed to be independent.

5 Properties

The new method allows comparing the relative importance of explanatory variables, and provides, in percentage terms, the proportion of deviations in the dependent variable explained by each explanatory variable. This new method is powerful, because simple to implement and easy to interpret. In addition, this method satisfies many desirable properties defined by Grömping (2015), as shown in table 2.

A range of properties are satisfied by all methods reviewed in this paper. The standardized beta coefficients (b_i^*), the partial and semi-partial r^2 and r, and the α_i of the new method are not affected by the positions of the regressors (anonymity), and not changed by linear transformations of individual variables. In large sample, these methods are not affected by the addition of a pure noise variable. Except for standardized beta coefficients[14], the outcomes of these methods are equal to zero if regression coefficients are equal to zero in large samples (exclusion), and the outcomes are strictly positive if regression coefficients are nonzero (inclusion).

The specificity of the new method and of standardized beta coefficients is that these methods relate to deviations of the dependent variables. However, the new method proposes to decompose these deviations into percentage contributions. The semi-partial r^2 also satisfies a decomposition property, but only when explanatory variables are orthogonal: it is a decomposition of the R^2. The new method and standardized beta coefficients are not affected by collinearity in the explanatory variables.

	(1)	(2)	(3)	(4)	(5)	(6)
	b_i^*	Partial		Semi-partial		α_i
		r^2	r	r^2	r	
Relates to variance		v	v			
Relates to deviations	v		v		v	v
Anonymity	v	v	v	v	v	v
Invariant to linear transformations	v	v	v	v	v	v
Not affected by a pure noise in large samples	v	v	v	v	v	v
Non-negativity		v	v	v	v	v
Exclusion in large samples (=0 if $\beta = 0$)	v	v	v	v	v	v
Inclusion (>0 if $\beta > 0$)		v	v	v	v	v
Not affected by collinearity	v					v
Decomposition				R^2		dev.

Table 2: Properties of the different methods

[14]The absolute value of standardized beta coefficients would satisfy these properties.

The new method has some limitations that should be highlighted. First, the method only applies to linear regression models. Future research should extend the scope of the method to more complex functional forms, and in particular, to quadratic terms which are widely used in econometrics.

Second, in small sample, the inclusion of variables that are irrelevant (not part of the true model) creates a downward bias in the estimated contributions of relevant variables. This problem is expected to be marginal in large samples, but relatively important in small samples. There is a trade-off between the inclusion of potentially irrelevant controls and this downward bias. In small sample, significance thresholds can be used as criteria for excluding irrelevant variables and mitigate this source of bias.

Finally, the new method implicitly assumes that the error term consists in one single variable which is unobserved. In practice, however, the error term is expected to be composed of a multitude of unobserved variables which are more or less important to predict the dependent variable. The number and the variance of these unobserved variables cannot be measured. The new method will overestimate the contribution of variables when the error term is composed of multiple unobserved variables.

Denoting e_i with $i = 1, ..., m$ the standard deviation of the unobserved variables composing the error term, we have that the contribution of each variable is given by $\alpha_i^{bias} \equiv \frac{|c_i|}{\sqrt{\sum_{i=1}^{m} {e_i}^2}+\sum_{j=1}^{n}|c_j|}$ instead of $\alpha_i^* \equiv \frac{|c_i|}{\sum_{i=1}^{m}|e_i|+\sum_{j=1}^{n}|c_j|}$. Because $\sum_{i=1}^{m}|e_i| > \sqrt{\sum_{i=1}^{m} {e_i}^2}$, the contribution of variables is overestimated when the error term is composed of multiple unobserved variables.

Different assumptions can be used to approximate the number and the standard deviation of variables composing the error term. Assuming that the variables composing the error term are independent and have the same variance, we have that $e = \sqrt{m{e_i}^2} \Rightarrow m = (e/e_i)^2$. The standard deviation of unobserved variables, e_i can be approximated by the mean of the $|c_i|$ ($e_i \approx \sum_j c_j/n = \bar{c}$). With this latter assumption, we posit that the economic importance of each unobserved variable is equal to the average economic importance of observables. When applying this adjustment, I recommend to exclude irrelevant (insignificant) variables, as these, if numerous, may sharply reduce the estimation of $\bar{b_i^*}$. Building on these assumptions, we construct an adjusted measure of economic importance that accounts for the fact that the error term is expected to be composed of numerous unobserved variables:

$$\alpha_i^{adj.} \approx \frac{|c_i|}{m\bar{c} + \sum_{j=1}^{n}|c_j|} = \frac{|c_i|}{(e^2/\bar{c}) + \sum_{j=1}^{n}|c_j|} \tag{9}$$

6 Applications

6.1 Application 1: a simple data generating process

A simple example illustrates the method and its alternatives. I generated the normal variables $x_1 \sim N(0,1)$, $x_2 \sim N(0,16)$, and $\epsilon \sim N(0,25)$. The correlation between x_1 and x_2 was arbitrarily set at 0.5, while ϵ is independent from the two other variables. These variables are combined to generate $y = x_1 + x_2 + \epsilon \sim N(0,44)$. The number of observations was set at 10,000 to obtain precise estimates. In line with Section 2, the absolute economic importance of x_1 and x_2 is equal to 10% and 40% respectively, while the error term captures 50% of deviations in y.

Column (1) of table 3 presents the results of a regression of y on x_1 and x_2. As expected, coefficients associated with x_1 and x_2 are close to 1, which does not tell much about the absolute and relative importance of these two variables. Column (2) shows the results of a regression with standardized variables. The coefficient of x_1^* is 0.154, which is approximately $1/\sqrt{44}$. Partial and semi-partial r^2 and r are displayed in columns (3) to (6). The outcome of the new method is presented in column (7). In line with the data generating process, the economic importance of x_1, x_2 and ϵ are estimated as: $\hat{\alpha_1} \approx 10.4\%$, $\hat{\alpha_2} \approx 40.1\%$ and $\hat{\alpha_e} \approx 49.5\%$. In this example, the sum $|b_1^*| + |b_2^*| + |b_e^*|$ is equal to 1.47, implying that the standardized beta coefficients overestimate the percentage contribution of x_1 and x_2 by 47%.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	y	y	Partial		Semi-partial		New method
		Standardized	r	r^2	r	r^2	
x_1	1.047***	0.154***	0.179	0.032	0.133	0.018	0.104***
	(0.057)	(0.008)					(0.006)
x_2	1.007***	0.592***	0.574	0.330	0.513	0.263	0.401***
	(0.014)	(0.008)					(0.006)
Residuals							0.495
Constant	0.024	-0.000					
	(0.050)	(0.007)					
N	10,000	10,000					
r^2	0.465	0.465					

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Linear regression model with non-orthogonal variables

6.2 Application 2: the data generating process of Experiment 2

The data generating process of underlying regressions in table 1 is: $y = 0.1x_1 + 0.3x_2 + 0.2x_3 + 2x_4 - 1x_5 + 1.5x_6 - 0.2x_7 + 0.1x_8 - 0.5x_9 + 0.2x_{10} + \sum_{i=11}^{50} x_i + \epsilon$. The variables x_i and ϵ are independent and normally distributed with $\sigma_1 = 140$, $\sigma_2 = 3.3$, $\sigma_3 = 20$, $\sigma_4 = 5.45$, $\sigma_5 = -0.1$, $\sigma_6 = 2.6$, $\sigma_7 = -10$, $\sigma_8 = 1$, $\sigma_9 = -5.8$, $\sigma_{10} = 5$, $\sigma_i = 1 \forall i \in 11...50$, $\sigma_e = 20$.

Keeping the same notation as before ($c_i = \beta_i \sigma_i$), we have $c_1 = 14$, $c_2 = 1$, $c_3 = 4$,..., with $|e| + \sum_{i=1}^{n} |c_i| = 100$. Formula 8 implies that $\alpha_1 = 14\%$, $\alpha_2 = 1\%$, $\alpha_3 = 4\%$,... The last column of table 1 shows estimates of the α_i with a sample of 2,583 observation randomly generated following this data generating process. These estimates are very close to their expected values. In this example, the sum $|b_e^*| + \sum_{i=1}^{n} |b_i^*|$ is equal to 3.53, implying that the regression with standardized variables overestimate the percentage contribution of variables by about 253%. From our specific sample and associated estimates, we indeed have that the coefficients of the standardized regression (column 2) are 241% higher than the estimated contribution in percentage (column 3). This example highlights the risk of overestimating the absolute economic importance of variables using a standardized regression, especially when the estimated model includes numerous variables with high predictive power.

6.3 Application 3: the roots of economic development

The new method is applied to study the "deep roots of economic development" (Spolaore and Wacziarg 2013). The empirical literature on the determinants of growth has recently moved from studying the proximate determinants of growth - such as capital accumulation and technology - to analyzing ever deeper, more fundamental factors of development. In this burgeoning literature, a central point of debate relates to the importance of geographical, historical and population composition variables.

The study of long-run causes of development is particularly relevant to illustrate the new method. Articles from this literature usually proceed as follows: (1) the authors describe a variable that they have newly constructed, (2) they then show that the new variable is a *statistically significant* and robust predictor of development, even when other variables discussed in the literature are controlled for, and finally (3) they show that the new variable is *economically important*, referring to standardized beta coefficients, partial-r^2 or similar methods to justify their argument. The presence of many competing predictors of long-run development makes the measurement of their relative and absolute economic importance very relevant.

I use OLS regressions to analyze the determinants of the logarithm of GDP per capita in 2000. Following the literature, I distinguish geographic factors, historic factors and the composition of populations (data sources are detailed in Appendix A).

The list of geographic variables includes the log of absolute latitude (Sala-i Martin 1997) and the average annual temperatures in Kelvin degrees (Dell et al. 2012), as well as a measure of terrain ruggedness, the distance to the nearest ice-free coast, a variable measuring carats of gem-quality diamonds extracted per square kilometer between 1958 and 2000, the percentage of each country with fertile soil, and the percentage of tropical land (Nunn and Puga 2012).[15]

Historical variables comprise a measure of slave trade intensity (Nunn and Puga 2012), a dummy for British legal origin (La Porta et al. 2008), an index of state history adjusted for population ancestry (Putterman and Weil 2010), and the number of years since a country transitioned from hunting and gathering to agriculture adjusted for population ancestry (Putterman and Weil 2010).

Population variables include measures of ethnolinguistic fragmentation (Alesina et al. 2003) and polarization (Reynal-Querol and Montalvo 2005), measures of ethnic inequality and spatial inequality (Alesina et al. 2015), the percentage of the population of European descent (Putterman and Weil 2010), and the genetic distance to the USA (Spolaore and Wacziarg 2009).[16]

Results are presented in table 4. Only geographical variables are included in the first block of four regressions (columns 1 to 4). Variables capturing the composition of the population are added in the second block (columns 5 to 8). Historical variables - for which data is missing for a substantial amount of countries - are included in the last two blocks (columns 9 to 16). The first column of each block shows results from a simple OLS regression. The second column presents the results from a standardized regression. Results from the new method are shown in the third column, according to equation (8). In the last column of each block, results from the new method are adjusted to account for the fact that the error term is likely composed of multiple variables (equation (9)).

A general overview of table 4 show that the predictive power of the estimated model is large ($R^2 = 0.72$ with all variables of interest). The coefficients of the standardized regression are much larger than the contributions of variables as calculated by the new method. From column (13) to column (14), all coefficients are divided by 2.8. Again,

[15]The measure of settlers mortality proposed by Acemoglu et al. (2001) is omitted due to limited data availability.

[16]I use the measure of genetic distance to the USA from Spolaore and Wacziarg (2009) rather than the measure of genetic diversity of Ashraf and Galor (2013) and its square to avoid including a quadratic term in the regression.

this highlights that the risk of overestimating the importance of variable is large when interpreting standardized beta coefficients. The estimated contribution of variables is slightly lower with the adjusted method.

Geographical variables explain a substantial percentage of deviations in the logarithm of GDP per capita. The estimated contribution of average temperature, distance to the nearest ice-free, the percentage of fertile soil, and terrain ruggedness are 11%, 8%, 7% and 7% respectively (based on column 15). Absolute latitude, the percentage of tropical land, and the measure of diamonds extraction are insignificant at conventional level in columns (9) and (13).

Among population composition variables, only the indicator of ethnic inequality and the percentage of the population of European descent are statistically significant across all specifications. Their estimated contributions to deviations in the dependent variable are 6% and 11% respectively. Genetic distance to the USA is statistically significant and its estimated contribution is equal to 8% when the indicator of state history adjusted for population ancestry is excluded from the regression. The same results are obtained with an indicator of genetic distance to the UK. This suggests that genetic distance could impact contemporary development indirectly, through its effect on state history. Regressions of state history on genetic distance to the USA or the UK, controlling for geographical variables, confirm that genetic distance is negatively and significantly correlated with state history. Genetic proximity could have impacted economic development through the diffusion of state-level institutions, but not so much through the diffusion of technological innovations (Spolaore and Wacziarg 2009). The coefficient associated with ethnolinguistic fragmentation is negative and significant in column (5) but not statistically significant in columns (9) and (13). This could be due to omitted variable bias when the intensity of slave trade is not controlled for. In line with findings from Whatley and Gillezeau (2011), slave trade is indeed likely to have favored fragmentation. Another explanation is that fragmentation in the past could have favored the slave trade, thereby indirectly impacting contemporary development.

In line with expectations, the coefficient associated of slave trade intensity is negative and statistically significant, while the coefficient of state history is positive and statistically significant. The estimated contributions of these factors to deviations in contemporary development are 5% and 8% respectively. Legal origins and years since the adoption of agriculture do not seem to explain economic development once other factors are controlled for.

I conclude that contemporary development of countries is explained by geographical and historical factors as well as by the composition of their population. Several factors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	GDP per capita in 2000 (log)															
	OLS	sd. OLS	% contrib.	% adj. contrib.	OLS	sd. OLS	% contrib.	% adj. contrib.	OLS	sd. OLS	% contrib.	% adj. contrib.	OLS	sd. OLS	% contrib.	% adj. contrib.
Absolute Latitude (log)	0.25*	0.17*	7.03	5.29	-0.23*	-0.17*	5.69	4.65	-0.19	-0.14	4.80	0.00	-0.15	-0.11	3.79	0.00
	(0.13)	(0.09)			(0.12)	(0.09)			(0.12)	(0.09)			(0.14)	(0.10)		
Avg. annual temp (K)	-0.09***	-0.53***	21.15	15.91	-0.05***	-0.31***	10.59	8.65	-0.04***	-0.25***	8.79	7.55	-0.05***	-0.31***	10.99	9.28
	(0.01)	(0.08)			(0.02)	(0.09)			(0.02)	(0.10)			(0.01)	(0.09)		
Ruggedness	-0.23***	-0.23***	9.35	7.04	-0.16***	-0.16***	5.41	4.42	-0.21***	-0.21***	7.29	6.27	-0.18***	-0.18***	6.53	5.51
	(0.06)	(0.07)			(0.06)	(0.06)			(0.06)	(0.06)			(0.06)	(0.06)		
Distance to Coast	-1.86***	-0.53***	21.10	15.87	-0.94***	-0.27***	9.11	7.44	-0.95***	-0.27***	9.40	8.08	-0.75***	-0.21***	7.57	6.39
	(0.20)	(0.06)			(0.21)	(0.06)			(0.20)	(0.06)			(0.21)	(0.06)		
Diamonds (carats)	-0.00	-0.05	2.19	0.00	0.00	0.04	1.52	0.00	0.00	0.02	0.85	0.00	0.00	0.05	1.86	0.00
	(0.00)	(0.09)			(0.00)	(0.07)			(0.00)	(0.06)			(0.00)	(0.05)		
% of fertile soil	-0.01***	-0.22***	8.81	6.63	-0.02***	-0.31***	10.49	8.57	-0.01***	-0.25***	8.64	7.43	-0.01**	-0.19**	6.93	5.85
	(0.00)	(0.07)			(0.00)	(0.07)			(0.00)	(0.07)			(0.00)	(0.08)		
% of tropical land	0.00	0.02	0.75	0.00	-0.00	-0.03	1.09	0.00	-0.00	-0.08	2.68	0.00	0.00	0.06	2.24	0.00
	(0.00)	(0.10)			(0.00)	(0.10)			(0.00)	(0.10)			(0.00)	(0.10)		
Ethnolinguistic fragmentation					-0.87**	-0.16**	5.63	4.60	-0.35	-0.07	2.31	0.00	-0.69	-0.13	4.70	3.96
					(0.42)	(0.08)			(0.44)	(0.08)			(0.47)	(0.09)		
Ethnolinguistic polarization					1.59	0.08	2.63	0.00	1.09	0.05	1.85	0.00	1.07	0.05	1.84	0.00
					(1.24)	(0.06)			(1.19)	(0.06)			(1.20)	(0.06)		
Ethnic Inequality					-1.46***	-0.28***	9.52	7.78	-1.16***	-0.22***	7.84	6.74	-0.87**	-0.17**	5.91	4.99
					(0.43)	(0.08)			(0.40)	(0.08)			(0.38)	(0.07)		
Spatial Inequality					0.07	0.01	0.50	0.00	0.09	0.02	0.65	0.00	0.03	0.01	0.24	0.00
					(0.39)	(0.08)			(0.34)	(0.07)			(0.34)	(0.07)		
% European descent					0.01**	0.20**	6.78	5.54	0.01**	0.21**	7.23	6.22	0.01***	0.30***	10.71	9.05
					(0.00)	(0.09)			(0.00)	(0.09)			(0.00)	(0.09)		
Genetic distance to the USA					-0.00***	-0.33***	11.40	9.31	-0.00***	-0.26***	9.01	7.75	-0.00	-0.08	2.83	0.00
					(0.00)	(0.07)			(0.00)	(0.07)			(0.00)	(0.11)		
Slave export intensity (log)									-0.12***	-0.21***	7.22	6.21	-0.08*	-0.13*	4.54	3.83
									(0.03)	(0.05)			(0.04)	(0.07)		
British legal origin dummy									0.18	0.06	2.16	0.00	0.18	0.06	2.20	0.00
									(0.16)	(0.06)			(0.17)	(0.06)		
State history (adj.)													1.46***	0.23***	8.10	6.84
													(0.53)	(0.08)		
Transition to agric. (adj.)													-0.00	-0.01	0.44	0.00
													(0.00)	(0.08)		
Residuals			29.63	49.26			19.64	39.05			19.27	43.75			18.58	44.29
Observations	173	173			154	154			154	154			135	135		
R^2	0.456	0.456			0.679	0.679			0.708	0.708			0.722	0.722		

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Long-run growth regressions

are statistically significant, but the contribution of each of these factors to deviations in economic development never exceeds 11%.

7 Conclusion

Assessing and discussing the economic importance of variables in regressions is becoming widespread in empirical economics. This trend is more than welcome. In this paper, I have however shown that the tools available to assess the economic importance of variables do not match with economists' understanding of the concept. Experimental data shows that economists understand the economic importance of variables as their contribution in percentage to deviations in the dependent variable. Neither the coefficients of standardized regressions nor partial and semi-partial r^2 or r correspond to this definition. This paper therefore proposed a new method to assess the economic importance of variables. The new method was applied to the study of the causes of long-run economic development. Results show that the economic development of countries is explained by a multitude of geographical and historical factors, as well as by the composition of their population.

References

Acemoglu, D., S. Johnson, and J. A. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review 91*(5), 1369–1401.

Achen, C. H. (1982). *Interpreting and using regression*, Volume 29. London, UK: Sage.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic growth 8*(2), 155–194.

Alesina, A., S. Michalopoulos, and E. Papaioannou (2015). Ethnic inequality. *Journal of Political Economy 123*(3), 547–724.

Ashraf, Q. and O. Galor (2013). The "out of africa" hypothesis, human genetic diversity, and comparative economic development. *The American Economic Review 103*(1), 1–46.

Bi, J. (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies 27*(2), 87–101.

Dell, M., B. F. Jones, and B. A. Olken (2012). Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics 4*(3), 66–95.

Engsted, T. (2009). Statistical vs. economic significance in economics and econometrics: Further comments on mccloskey and ziliak. *Journal of Economic Methodology 16*(4), 393–408.

Grissom, R. J. and J. J. Kim (2012). *Effect sizes for research: Univariate and multivariate applications*. Routledge.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics 7*(2), 137–152.

Holgersson, H., T. Norman, and S. Tavassoli (2014). In the quest for economic significance: assessing variable importance through mean value decomposition. *Applied Economics Letters 21*(8), 545–549.

Hoover, K. D. and M. V. Siegler (2008). Sound and fury: Mccloskey and significance testing in economics. *Journal of Economic Methodology 15*(1), 1–37.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 666–687.

La Porta, R., F. Lopez-de Silanes, and A. Shleifer (2008). The economic consequences of legal origins. *Journal of Economic Literature 46*(2), 285–332.

Luskin, R. C. (1991). Abusus non tollit usum: standardized coefficients, correlations, and r 2s. *American Journal of Political Science*, 1032–1046.

McCloskey, D. N. and S. T. Ziliak (1996). The standard error of regressions. *Journal of Economic Literature 34*(1), 97–114.

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity. *The American Economic Review 102*(4), 1508–1539.

Nakamura, E. and J. Steinsson (2014). Fiscal stimulus in a monetary union: Evidence from us regions. *The American Economic Review 104*(3), 753–792.

Nunn, N. and D. Puga (2012). Ruggedness: The blessing of bad geography in africa. *Review of Economics and Statistics 94*(1), 20–36.

Putterman, L. and D. N. Weil (2010). Post-1500 population flows and the long run determinants of economic growth and inequality. *The Quarterly Journal of Economics 125*(4), 1627.

Reynal-Querol, M. and J. G. Montalvo (2005). Ethnic polarization, potential conflict and civil war. *American Economic Review 95*(3), 796–816.

Sala-i Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 178–183.

Spolaore, E. and R. Wacziarg (2009). The diffusion of development. *The Quarterly Journal of Economics 124*(2), 469–529.

Spolaore, E. and R. Wacziarg (2013). How deep are the roots of economic development? *Journal of Economic Literature 51*(2), 325–369.

Whatley, W. and R. Gillezeau (2011). The impact of the transatlantic slave trade on ethnic stratification in africa. *The American Economic Review 101*(3), 571–576.

Ziliak, S. T. and D. N. McCloskey (2004). Size matters: the standard error of regressions in the american economic review. *The Journal of Socio-Economics 33*(5), 527–546.

A Data sources

The following variables are taken from the dataset of Alesina et al. (2015): the log of real GDP per person in 2000, the absolute latitude of countries, the average annual temperatures in Kelvin degrees, the dummy for British legal origin, the measures of ethnolinguistic fragmentation and polarization, and the measures of ethnic inequality and spatial inequality.

The following variables are taken from the dataset of Nunn and Puga (2012): the measure of terrain ruggedness, the distance to the nearest ice-free coast, the indicator of diamonds extraction, the percentage of each country with fertile soil, the percentage of tropical land, the measure of slave export intensity, and the percentage of the population of European descent.

The following variables are taken from the dataset of Spolaore and Wacziarg (2013): the index of state history adjusted for population ancestry, the number of years since a country transitioned from hunting and gathering to agriculture adjusted for population ancestry, and measures of genetic distance to the USA and the UK.