



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal (2017)
17, Number 2, pp. 422–441

Estimating responsiveness scores using `rscore`

Giovanni Cerulli
CNR-IRCrES
National Research Council of Italy
Institute for Research on Sustainable Economic Growth
Rome, Italy
giovanni.cerulli@ircres.cnr.it

Abstract. `rscore` computes unit-specific responsiveness scores using an iterated random-coefficient regression approach. The model fit by `rscore` considers a regression of a response variable y , that is, *outcome*, on a series of factors (or regressors) \mathbf{x} , that is, *varlist*, by assuming a different reaction (or “responsiveness”) of each unit to each factor contained in \mathbf{x} . `rscore` allows for i) ranking units according to the obtained level of the responsiveness score; ii) detecting more influential factors in driving unit performance; and iii) studying the distribution (heterogeneity) of factors’ responsiveness scores across units. Also, `rscore` offers useful graphical representation of results. We provide two illustrative applications of the model: the first is on a cross-section, and the second is on a longitudinal dataset.

Keywords: st0480, `rscore`, responsiveness scores, random-coefficient regression

1 Introduction

In biomedical and socioeconomic disciplines, it is commonly recognized that individual agents react heterogeneously to external stimuli. Such heterogeneity also characterizes aggregated units of analysis such as companies, regions, and entire countries.

Thus measuring unit-heterogeneous response to specific factors is relevant to providing a clearer understanding of the relationship between a stimulus (a drug administration, a policy program, etc.) and its effect on predefined target variables.

To this end, in this article, I present `rscore`, a user-written command for computing a unit-specific responsiveness score (RS) by means of an iterated random-coefficient regression (RCR) approach.

Simply put, the model fit by `rscore` considers a regression of a response variable y , that is, *outcome*, on a series of factors (or regressors) \mathbf{x} , that is, *varlist*, by assuming a different reaction (or “responsiveness”) of each unit to each factor contained in \mathbf{x} .

A simple example can better illustrate the usefulness of this command for applied research. Consider the popular Stata instructional dataset `auto.dta`, and suppose we are interested in regressing the variable `price` (“price of the car”) on `mpg` (“miles per gallon”). What we usually estimate is a common slope regression of this type,

$$\text{price}_i = \alpha + \beta \times \text{mpg}_i + e_i$$

where α and β are two parameters common to all units i , and e_i , an error term. However, suppose (when reasonable) each car has a different behavior in terms of the reaction of its price to its consumption. The coefficient β catches only the “average” effect of all cars, whereas knowing about the idiosyncratic behavior of each single car would be much more informative.

To this end, we may be interested in estimating a regression with a unit-specific slope, such as

$$\text{price}_i = \alpha + \beta_i \times \text{mpg}_i + e_i$$

where it is now clear that the slope refers to unit i . If, under reasonable assumptions, we could compute each β_i , we would obtain precious additional information about the relation between car price and consumption.

rscore aims at estimating each β_i that, in this context, we call the RS of **price** to **mpg**.¹ As will be clearer later on, it is a score, not an estimate, because we assume insufficient information in the data to perform proper inference on each β_i (without strong additional assumptions, as discussed in the next section). However, RSs may convey useful descriptive information for many phenomena in empirical research. More specifically, **rscore** may allow for i) ranking units according to the obtained level of the RS; ii) detecting more influential factors in driving unit performance; and iii) studying the distribution (heterogeneity) of factors’ RSs across units. Also, **rscore** offers useful graphical representation of results.

This article is organized as follows: section 2 provides a short statistical account of the RCR, which is useful to transition into section 2.5, where I describe the model fit by **rscore**. Section 4 illustrates the syntax of **rscore**. Section 5 and section 6 provide two illustrative applications, one on a cross-section and one on a longitudinal dataset. Finally, section 7 ends the article.

2 Statistical background

In recent years, random-coefficient models have been the objective of vibrant research (Lewbel and Pendakur Forthcoming; Hoderlein, Klemelä, and Mammen 2010; Beran, Feuerverger, and Hall 1996). This literature has tried to overcome some limits of the traditional regression model by incorporating either correlated or uncorrelated random coefficients, estimated parametrically or nonparametrically. In what follows, I provide a brief description of an (exogenous) linear random-coefficient model.

To set the stage, consider a standard regression model. In such a model, parameters (that is, “regression coefficients”) are singleton numbers. In a simple regression of the type

$$y_i = \alpha + \beta \times x_i + \epsilon_i \quad i = (1, \dots, N)$$

1. More precisely, **rscore** computes the expectation of β_i , conditional on the exogenous covariates (in this specific case, the covariate **mpg**). See section 2.5.

the regression coefficient β is a population parameter, which summarizes the effect of factor² x on outcome y , when it is assumed that each observation within such population shares the same β (as well as the same α).

A way to relax such an assumption is to assume each unit owns a specific regression coefficient in the population. The previous model thus becomes

$$y_i = \alpha_i + \beta_i \times x_i + \epsilon_i$$

where the generic unit i owns both a specific intercept (α_i) and a specific slope (β_i). Assume that x_i is exogenous, implying that $E(\epsilon_i|x_i) = 0$, and let's focus on β_i by letting $\alpha_i = \alpha$ for the sake of clarity. With no further information than that specified in the previous model, identifying and estimating an intercept and a slope for each individual within a sample of size N would be impossible.

However, to identify such parameters, one can impose assumptions over the probabilistic distribution of β_i by holding, for instance, that each β_i is a draw from a compounded random variable of the type

$$\beta_i = \beta + v_i$$

where v_i is a random variable with $E(v_i) = 0$ and β is a (common) parameter. This implies that $E(\beta_i) = \beta$. Under this assumption, we obtain—by substitution—a simplified version of the previous model:

$$y_i = \alpha + \beta \times x_i + v_i x_i + \epsilon_i$$

If we assume $E(v_i|x_i) = 0$, we get

$$E(y_i|x_i) = \alpha + \beta \times x_i$$

which is a standard regression model, where the common slope β is consistently estimated by ordinary least squares (OLS) or, in the case of a heteroskedastic error, generalized least squares.

When longitudinal data are available, one can also estimate each β_i by unit-specific OLS. For this purpose, one can refer to [Swamy \(1970\)](#) for estimating RCRs within longitudinal data, as well as the related implementation provided by [Poi \(2003\)](#).

With cross-section datasets, another possible route is assuming $E(v_i|x_i) \neq 0$, which implies that β_i is a function of the covariates (or a subset of those) included in the regression model. In this case,

$$\begin{aligned} E(y_i|x_i) &= \alpha + \beta x_i + E(v_i|x_i) \times x_i + \epsilon_i \\ &= \alpha(x_i) + \beta(x_i)x_i \end{aligned} \tag{1}$$

provided again that $E(\epsilon_i|x_i) = 0$. In the previous equation, $\alpha(x_i) = \alpha + \beta x_i$ and $\beta(x_i)$ is a generic function of x that can be modeled either parametrically or nonparametrically, depending on the way one models the conditional expectation $E(v_i|x_i)$.

2. In this article, we use the term “factor” to indicate a generic covariate, although “factor” generally refers to a variable taking a discrete number of values.

Following (1), one can compute the partial effect of x on y as

$$\frac{\partial}{\partial x_i} E(y_i|x_i) = \alpha'(x_i) + \beta'(x_i)x_i + \beta(x_i) \quad (2)$$

where it is clear that each unit i owns a different partial effect depending on x_i . The mean over x of (2) is known as average partial effect (APE), which is defined as

$$\text{APE} = E_x \left\{ \frac{\partial}{\partial x_i} E(y_i|x_i) \right\} = E_x \{ \alpha'(x_i) + \beta'(x_i)x_i + \beta(x_i) \} \quad (3)$$

To estimate APE, one can use the sample equivalent of (3), that is

$$\widehat{\text{APE}} = \frac{1}{N} \sum_{i=1}^N \left\{ \widehat{\alpha}'(x_i) + \widehat{\beta}'(x_i)x_i + \widehat{\beta}(x_i) \right\}$$

It is clear that the heterogeneous response of each unit to a given covariate (or factor) can be interesting to evaluate per se because it brings several pieces of information about how APE takes a specific value whenever one looks at APE as the global average effect of x on y .

In this article, the partial effect of a given factor x on a unit's outcome y is called the RS of a unit's y to a unit's x , all other things being equal. Once such RSs are estimated (using a proper estimation procedure, as I will show later on), one can use them for various purposes, such as i) ranking units according to the obtained level of the RSs; ii) detecting factors that are more influential in driving unit performance; and iii) studying the distribution (heterogeneity) of factors' RSs across units.³

`rscore` is a command for computing these unit-specific RSs. Thus it uses an iterated RCR model whose baseline regression is similar to (1). Before presenting the syntax of `rscore`, I will illustrate the RS model's structure and assumptions.

3 The model

RSs measure the change of a given outcome y when a given factor x_j (with $j = 1, \dots, Q$) changes, conditional on all other $(Q - 1)$ factors:

$$\mathbf{x}_{-j} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_Q)$$

Algebraically, it is the derivative of y on x_j , given \mathbf{x}_{-j} , when one allows for each observation to have its own RS. Importantly, we assume \mathbf{x}_{-j} is a vector of all exogenous variables. RSs are obtained using an iterated RCR model, the basic econometrics of which can be found in Wooldridge (2002, 638–642; 2010, 141–145).

3. An early empirical application of the RS approach proposed in this article can be found in Cerulli (2014).

Calculating an RS follows this simple protocol:

1. Define y , the outcome (or response) variable.
2. Define a set of Q factors thought of as affecting y , and indicate the generic factor with x_j .
3. Define an RCR model linking y to the various x_j , and extract a unit-specific responsiveness effect of y to all set of factors x_j , with $j = 1, \dots, Q$.
4. For the generic unit i and factor j , indicate such effects as b_{ij} , and collect all of them in a matrix \mathbf{B} .
5. Finally, aggregate by unit (row) or by factor (column) the $E(b_{ij}|\mathbf{x}_{i,-j})$, thus getting synthetic unit and factor responsiveness measures.

Analytically, an RS is the “partial effect” of a factor x in an RCR (Wooldridge 1997; 2003; 2004). Indeed, for each $j = 1, \dots, Q$, define an RCR model of this kind as

$$\begin{cases} y_i = a_{ij} + b_{ij}x_{ij} + e_i \\ a_{ij} = \gamma_0 + \mathbf{x}_{i,-j}\gamma + u_{ij} \\ b_{ij} = \delta_0 + \mathbf{x}_{i,-j}\delta + v_{ij} \end{cases}$$

where e_i , u_{ij} , and v_{ij} are freely correlated error terms with

$$E(e_i|\mathbf{x}_{i,-j}; x_{ij}) = E(u_{ij}|\mathbf{x}_{i,-j}; x_{ij}) = E(v_{ij}|\mathbf{x}_{i,-j}; x_{ij}) = 0$$

It is easy to see that the regression parameters, a_{ij} and b_{ij} , are both nonconstant because they depend on all the other inputs x except x_j (this is, in fact, the meaning of the vector $\mathbf{x}_{i,-j}$). Observe that δ_0 and γ_0 are, on the contrary, constant parameters. According to this model, we can define the regression line as

$$E(y_i|x_{ij}, \mathbf{x}_{i,-j}) = E(a_{ij}|\mathbf{x}_{i,-j}) + x_{ij} \times E(b_{ij}|\mathbf{x}_{i,-j})$$

Given this, we define the responsiveness effect of x_{ij} on y_i as the derivative of y_i with respect to x_{ij} ; that is,

$$\frac{\partial}{\partial x_{ij}} \{E(y_i|x_{ij}, \mathbf{x}_{i,-j})\} = E(b_{ij}|\mathbf{x}_{i,-j})$$

where $E(b_{ij}|\mathbf{x}_{i,-j})$ is the partial effect of x_{ij} on y_i . We can repeat the same procedure for each x_{ij} (with $j = 1, \dots, Q$)—so it is eventually possible to define, for each unit $i = 1, \dots, N$ and factor $j = 1, \dots, Q$, the $N \times Q$ matrix \mathbf{B} of the partial effects as follows:

$$\mathbf{B} = \begin{pmatrix} E(b_{11}|\mathbf{x}_{1,-j}) & \dots & E(b_{1Q}|\mathbf{x}_{1,-j}) \\ \vdots & E(b_{ij}|\mathbf{x}_{i,-j}) & \vdots \\ E(b_{N1}|\mathbf{x}_{N,-j}) & \dots & E(b_{NQ}|\mathbf{x}_{N,-j}) \end{pmatrix}$$

If all variables are standardized with zero means and unit variance, partial effects are beta coefficients and therefore independent of the unit of measurement. Thus they can be compared with each other and summed.⁴

Once the matrix \mathbf{B} is known, we can define, for each unit i , the total unit responsiveness (TUR) and the mean unit responsiveness (MUR) as

$$\text{TUR}_i = \sum_{j=1}^Q b_{ij}$$

and

$$\text{MUR}_i = \frac{1}{Q} \sum_{j=1}^Q b_{ij}$$

and define, for each factor j , the total (or mean) responsiveness of y to factor j 's unit change (TFR and MFR) as

$$\text{TFR}_j = \sum_{i=1}^N b_{ij}$$

and

$$\text{MFR}_j = \frac{1}{N} \sum_{i=1}^N b_{ij}$$

In a cross-section data setting, OLS provide a consistent estimation of each b_{ij} within this regression,⁵

$$\begin{aligned} y_i &= \gamma_0 + \mathbf{x}_{i,-j}\gamma + (\delta_0 + \bar{\mathbf{x}}_{-j}\delta)x_{ij} + x_{ij}(\mathbf{x}_{i,-j} - \bar{\mathbf{x}}_{-j})\delta + \eta_i \\ \eta_i &= u_{ij} + x_{ij}v_{ij} + e_i \end{aligned}$$

where $\bar{\mathbf{x}}_{-j}$ is the vector of the sample means of $\mathbf{x}_{i,-j}$. Once previous regression parameters are estimated, we can obtain an estimate of the partial effect of factor x_j on y for the generic unit i as

$$\hat{E}(b_{ij}|\mathbf{x}_{i,-j}) = \hat{\delta}_0 + \mathbf{x}_{i,-j}\hat{\delta}$$

By repeating this procedure for each unit i and factor j , we can finally obtain $\hat{\mathbf{B}}$, that is, the estimation of matrix \mathbf{B} .

When a longitudinal dataset is available, the estimation of \mathbf{B} can be obtained by using either random-effects or fixed-effects estimation of the following panel-data regression,

$$y_{it} = \gamma_0 + \mathbf{x}_{i,-j,t}\gamma + (\delta_0 + \bar{\mathbf{x}}_{-j,t}\delta)x_{ijt} + x_{ijt}(\mathbf{x}_{i,-j,t} - \bar{\mathbf{x}}_{-j,t})\delta + \alpha_i + \eta_{it} \quad (4)$$

4. Because beta coefficients are measured in standard deviations, they can be compared. The meaning of a beta coefficient is straightforward; suppose that in a regression of y on x , the beta is found to be equal to 0.3. This means that one standard-deviation increase in x leads to a 0.3 standard-deviation increase in the predicted y with all the other variables in the model held constant.

5. Indeed, OLS are consistent because for each $j = 1, \dots, Q$, we have $E(\eta_i|\mathbf{x}_{ij}) = E(u_{ij}|\mathbf{x}_{ij}) + x_{ij} \times E(v_{ij}|\mathbf{x}_{ij}) + E(e_i|\mathbf{x}_{ij}) = 0$. However, η_i is clearly heteroskedastic. Thus robust OLS provide more correct standard errors.

where the added parameter α_i represents a unit-specific effect accounting for unobserved heterogeneity. In particular, fixed-effects estimation, assuming free correlation between α_i and η_{it} , can mitigate a potential endogeneity bias due to misspecification of previous equation and measurement errors in the variables considered in the model (Wooldridge 2010, 281–315). As such, a panel dataset may allow for more reliable estimates of true RS than usual OLS.

Finally, as long as variables are standardized, (4) becomes

$$y_{it} = \gamma_0 + \mathbf{x}_{i,-j,t}\gamma + \delta_0 x_{ijt} + x_{ijt} \times \mathbf{x}_{i,-j,t}\delta + \alpha_i + \eta_{it}$$

which simplifies the formula.

4 The *rscore* command

4.1 Syntax

As seen above, *rscore* computes unit-specific RSs using an iterated RCR model. The model fit by *rscore* considers a regression of a response variable y , that is, *outcome*, on a series of factors \mathbf{x} , that is, *varlist*, by assuming a different reaction (or “responsiveness”) of each unit to each factor contained in \mathbf{x} . The basic syntax of *rscore* is

```
rscore outcome [varlist] [if] [in] [weight], model(modeltype) rs_name(stub)
    [factors(varlist_f) xlist(varlist_c) graph(#) radar(numlist)
    id_string(varname) vce(vcetype) save_graph1(filename)
    save_graph2(filename) ]
```

fweights, *iweights*, and *pweights* are allowed; see [U] 11.1.6 **weight**.

4.2 Options

model(*modeltype*) specifies the model to be fit, where *modeltype* must be one of the following models: **ols** (OLS), **fe** (panel fixed effect), or **re** (panel random effect). **model**() is required.

rs_name(*stub*) specifies the beginning part of the name of the RS variables generated by *rscore*. RS variables are thus named as *stub1*, *stub2*, *stub3*, ..., *stubQ*. **rs_name**() is required.

factors(*varlist_f*) specifies that factor variables have to be included among the regressors.

xlist(*varlist_c*) specifies that control variables (which are not factors) have to be included among the regressors.

`graph(#)` provides a combined graph of the densities of the RSs. The number `#` defines the width of the graph's x axis. The user can set a proper `#` for providing a good rendering of the graph.

`radar(numlist)` provides a radar plot of the RSs for the units specified in *numlist*. To run this option, the user must specify the `id_string()` option. This option uses the user-written `radar` command provided by Mander (2007).

`id_string(varname)` requests that a string variable as identifier of each observation be specified. This is required if the user wishes to provide a radar plot of the RSs.

`vce(vcetype)` allows the user to choose *vcetype* as either `robust` or `cluster clustvar`.

`save_graph1(filename)` saves the graph generated by the `graph()` option in the user-specified *filename*.

`save_graph2(filename)` saves the graph generated by the `radar()` option in the user-specified *filename*.

4.3 Stored results

Finally, for each factor regression, `rscore` returns goodness-of-fit statistics—that is, R -squared—stored in scalars `e(R1)`, `e(R2)`, `e(R3)`, ..., `e(RQ)`, as well as the average R -squared, which is the overall goodness of fit of the model, stored in the scalar `e(R)`.

5 Application 1

This section presents an illustrative application of `rscore` within a cross-section data structure. It is intended to allow users to become familiar with the use of this command.

To this end, we consider the usual `auto.dta`, with a full specification of the `rscore` syntax. In this application, we are interested in identifying the main factors driving the price of a car and identifying the distribution of such an effect over observations for each declared factor. We focus on price responsiveness to five covariates (that is, `mpg`, `trunk`, `weight`, `length`, and `displacement`) by controlling for two factor variables (that is, `foreign` and `rep78`) and two controls (that is, `gear_ratio` and `headroom`). The application of `rscore` results in code like this:

```
. sysuse auto
(1978 Automobile Data)

. rscore price mpg trunk weight length displacement, rs_name(RS)
> model(ols) factors(foreign rep78) xlist(gear_ratio headroom)
> graph(100) id_string(make) radar(4 9 13 40) save_graph1(mydistr)
> save_graph2(myradar)

*****
*** DESCRIPTIVE STATISTICS FOR SINGLE FACTOR RESPONSIVENESS SCORES ***
*****
```

Responsiveness scores for variable mpg_std				
	Percentiles	Smallest		
1%	-.9230999	-.9230999		
5%	-.3402585	-.7910841		
10%	-.2775861	-.7330251	Obs	69
25%	-.090533	-.3402585	Sum of Wgt.	69
50%	.0403205		Mean	.0112316
		Largest	Std. Dev.	.2559057
75%	.1841674	.3397723		
90%	.2829912	.3649702	Variance	.0654877
95%	.3397723	.4167671	Skewness	-1.354045
99%	.477404	.477404	Kurtosis	5.911784
Responsiveness scores for variable trunk_std				
	Percentiles	Smallest		
1%	-.5023578	-.5023578		
5%	-.2346051	-.3120776		
10%	-.1627576	-.2949739	Obs	69
25%	.0015847	-.2346051	Sum of Wgt.	69
50%	.1770424		Mean	.2170236
		Largest	Std. Dev.	.3706054
75%	.3745154	.718573		
90%	.6115505	1.25186	Variance	.1373483
95%	.718573	1.303261	Skewness	1.583251
99%	1.77037	1.77037	Kurtosis	7.317773
Responsiveness scores for variable weight_std				
	Percentiles	Smallest		
1%	.4967597	.4967597		
5%	.5914793	.5649452		
10%	.6201076	.5688947	Obs	69
25%	.7530873	.5914793	Sum of Wgt.	69
50%	.9084411		Mean	.9705862
		Largest	Std. Dev.	.3131494
75%	1.128973	1.620982		
90%	1.530207	1.666719	Variance	.0980625
95%	1.620982	1.668086	Skewness	.8406102
99%	1.763967	1.763967	Kurtosis	2.886559
Responsiveness scores for variable length_std				
	Percentiles	Smallest		
1%	-.8865399	-.8865399		
5%	-.8249316	-.8677544		
10%	-.6593894	-.8445593	Obs	69
25%	-.5387048	-.8249316	Sum of Wgt.	69
50%	-.4376124		Mean	-.4303998
		Largest	Std. Dev.	.1895008
75%	-.3094746	-.143986		
90%	-.2292118	-.0476573	Variance	.0359106
95%	-.143986	.0030893	Skewness	-.0796322
99%	.0920853	.0920853	Kurtosis	3.649125

```

                                Responsiveness scores for variable
                                displacement_std
-----
      Percentiles      Smallest
1%      -.8768004      -.8768004
5%      -.5376742      -.7265469
10%     -.4575502      -.6088864      Obs          69
25%     -.2181616      -.5376742      Sum of Wgt.    69
50%     -.0629257
                                Mean          .0022286
                                Std. Dev.     .4293259
                                Largest
75%     .1630541      .698191
90%     .5437063      1.304059      Variance      .1843207
95%     .698191      1.332656      Skewness      1.316481
99%     1.496196      1.496196      Kurtosis      5.795581
-----

*****
***** RSCORE GOODNESS-OF-FIT *****
*****

The R-squared for mpg_std is:
.63707556

-----

The R-squared for trunk_std is:
.65798794

-----

The R-squared for weight_std is:
.65783691

-----

The R-squared for length_std is:
.62031188

-----

The R-squared for displacement_std is:
.73278452

-----

The mean R-squared is:
.66119936
-----

```

We call `rscore` using an OLS estimation through the `model(ols)` option. This option is appropriate with a cross-section dataset such as `auto.dta`. The default output of `rscore` is a series of summary statistics tables for each factor RS considered and a table reporting the single-factor regression *R*-squared and the model mean (or overall) *R*-squared. In this specific case, we obtain five summary statistics and six *R*-squared. From summary statistics, we see that the variable `weight` sets out the highest average RS (with a value of 0.97). Notice that RSs are beta coefficients because `rscore` *z*-standardizes each variable in advance. Therefore, a mean RS for `weight` of 0.97 means that—on average—one standard-deviation change in cars' weight yields a one standard-deviation increase in cars' price. We do not have a significance test for this value because it is held only as a score. From the goodness-of-fit output, we see that the best fit is reached by variable `displacement` with an *R*-squared of about 0.73, which is substantial. The average *R*-squared of the model is 0.66, meaning that—on average—our factors' specification explains around 66% of the total variance of cars' prices. Because it is a rather high *R*-squared, we can accept our specification as a good one.

Because we set the `rs_name(RS)` option, the command generates five new variables called `RS1`, `RS2`, `RS3`, `RS4`, and `RS5`, which contain the RSs for the five factors. The variables' labels suggest which factor each generated variable refers to.

The `factors(foreign rep78)` and `xlist(gear_ratio headroom)` options set the factor variables and a set of additional (continuous) variables to control for. They are specified only as controls; that is, they provide `rscore` with a correct specification for price.

To generate graphical results, `rscore` offers two options that can also be combined: `graph(#)` and `radar(4 9 13 40)`. Note that the latter option needs to be specified jointly with the string identifier of the observations, which in this case is the `make` variable. Hence, the `radar(4 9 13 40)` option needs to be combined with the `id_string(make)` option.

If the `graph(#)` option is specified, Stata returns the graph in figure 1. This is the joint plot of the distribution of all factors' RSs. This graph allows one to visually detect two aspects:

- Factor importance: This is (positively) higher as soon as the factor's RS distribution lies on the right side of the figure.
- Factor response heterogeneity: This is higher whenever the factor's RS distribution presents long tails.

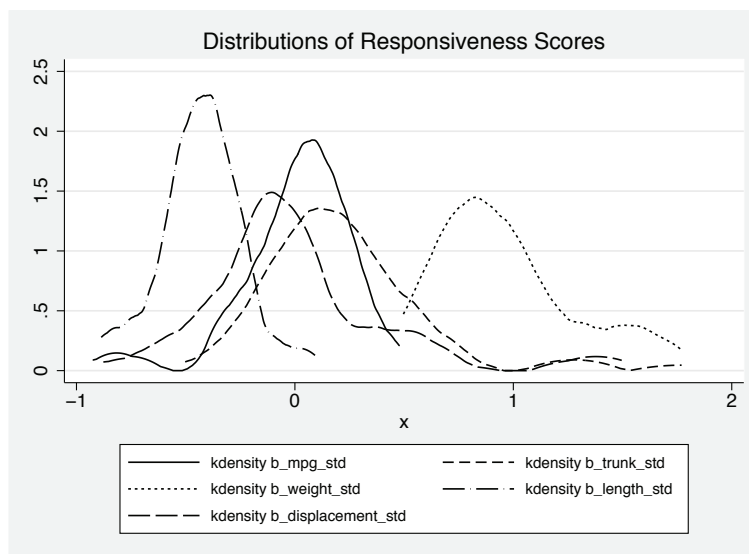


Figure 1. Joint plot of the distribution of factors' RSs

Looking at figure 1, we see that the factor whose distribution stands out in terms of a positive effect on price is a car's weight. This distribution is located at the extreme

right of the graph and presents a long right tail. However, the factors with larger RS dispersions are **trunk** and **displacement**, while **length** shows a very concentrated distribution, albeit with a negative impact on price on average. It means that cars' price response to their length are weak, negative, and very similar among all car models.

If the `radar(4 9 13 40)` and `id_string(make)` options are (jointly) specified, we obtain the radar graph of figure 2.

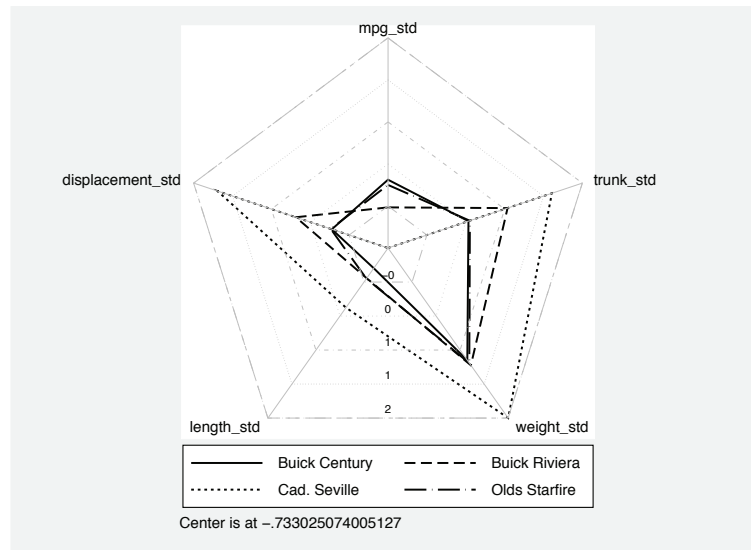


Figure 2. Units' RS radar graph

This graph is useful for comparing—factor by factor—different unit RSs. In this example, we aim to compare units 4, 9, 13, and 40 in our dataset, which correspond to specific models of cars. By looking at figure 2, we immediately see that the car model “Cadillac Seville” presents a higher RS for each factor, except for **mpg**. The car models “Buick Century” and “Old Starfire” show a similar RS on each factor, while “Buick Riviera” has a larger impact on price through **trunk** and **displacement**. Thus the radar graph is useful to have a quick snapshot of units' comparative results on single factors' RSs. This is a handy option, which emphasizes the advantage of using an RS approach over traditional regression methods.

6 Application 2

In this second application, I use `rscore` to identify the main drivers of countries' gross domestic product (GDP). I address three research questions:

- Factor importance rank: Among countries' GDP components, what are those whose growth change produces a larger or smaller response in terms of GDP growth?
- Factor response heterogeneity: Is a country's GDP growth response more or less heterogeneously or homogeneously distributed among its driving factors?
- Unit responsiveness rank: Which units have larger or smaller RSs for each given factor?

For a dataset, we consider the World Bank's "Economy & Growth" indicators database, which is made of 283 macroeconomic indicators for 250 countries collected from 1960–2014. We consider 13,695 observations, thus obtaining a huge longitudinal dataset.

As drivers, we consider the main components of GDP formation, plus the government surplus or deficit, that is,

- Cash surplus or deficit (percent of GDP)
- General government final consumption expenditure (annual percent growth)
- Household final consumption expenditure (annual percent growth)
- Gross fixed capital formation (annual percent growth)
- Exports of goods and services (annual percent growth)
- Imports of goods and services (annual percent growth)

Within this dataset, we consider the variable `ny_gdp_pcap_kd_zg` representing the "real GDP rate of growth". We plot the time pattern of this variable from 1990–2013 for the five largest European countries, namely, Great Britain (GBR), Germany (DEU), Italy (ITA), France (FRA), and Spain (ESP). Figure 3 shows the time pattern, and the Stata code to obtain this figure is

```

. clear all
. set maxvar 30000

. use "data_worldbank_economy&growth.dta", clear
. global time year>=1990

. twoway
> line Y year if countrycode == "GBR" & $time, sort ||
> line Y year if countrycode == "FRA" & $time, sort ||
> line Y year if countrycode == "ITA" & $time, sort ||
> line Y year if countrycode == "ESP" & $time, sort ||
> line Y year if countrycode == "DEU" & $time, sort
> xlabel(1990(2)2015, gmax angle(horizontal))
> legend(label(1 "GBR") label(2 "FRA") label(3 "ITA")
> label(4 "ESP") label(5 "DEU")) title("GDP annual growth")

```

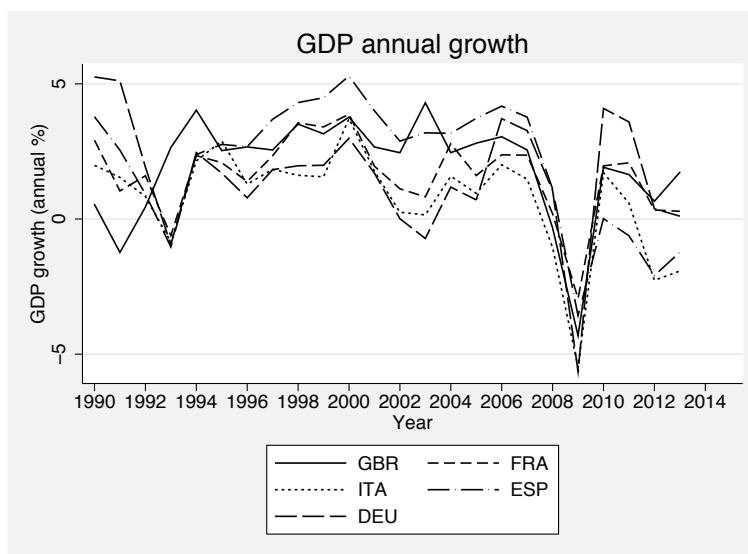


Figure 3. GDP annual growth rate: 1990–2012

We now apply `rscore` by running a fixed-effects model (because a panel-data structure is now available):

```

. * Estimate RS for the "GDP annual growth" (Y)
. global xvars B G C I E M
. * Model with fixed effects
. capture drop w
. generate w=10
. capture drop id_country
. encode countryname, generate(id_country)
. tsset id_country year
    panel variable: id_country (strongly balanced)
    time variable: year, 1960 to 2014
                  delta: 1 unit

```

```

. capture drop year2
. tostring year, gen(year2)
year2 generated as str4
. capture drop countryr
. generate countryr=countryname+year2
. rscore Y $xvars [pweight=w], model(fe) rs_name(_bx) graph(3)
> radar(4508 4233 5938 12978 11383 13033) id_string(countryr)
(output omitted)

```

For the sake of brevity, we omit the RS summary tables and consider only the graphical outcomes. Indeed, as in the previous application, *rscore* generates the following combined plot of the distributions of RS for each variable considered. This graph is illustrated in figure 4.

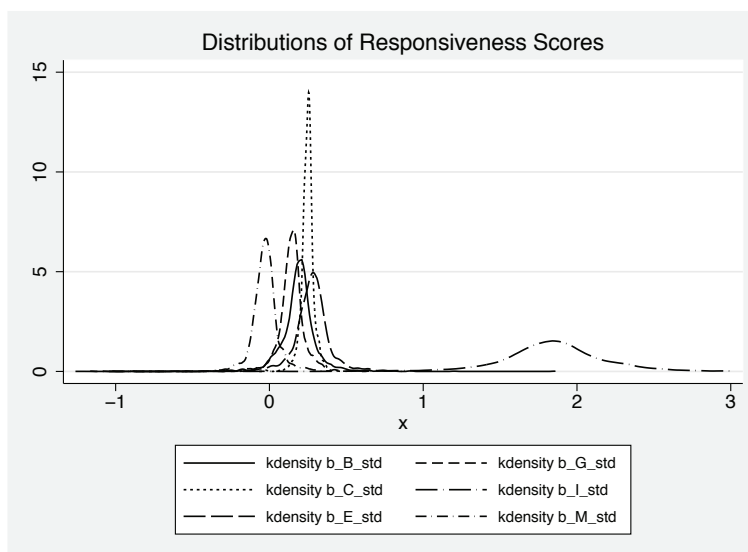


Figure 4. Distribution of RSs: 1990–2012

Quite surprisingly, figure 4 shows that the most relevant factor driving larger GDP rate of growth (by country and by time) is represented by the private fixed investment. Larger annual growth in this variable is associated with larger GDP growth, other things being equal. More precisely, one standard-deviation increase in private investment turns out to be associated—on average—with about two standard-deviations' positive increase in the GDP rate of growth.

However, private investment is also the factor showing the largest RS variance around its mean, which can be interpreted as a measure of risk whenever one wants to draw policy advice from this result. The second most relevant driver of GDP growth is export, which shows a much more concentrated RS distribution than investment. As expected, imports rank in the last position, with a negative average RS. Table 1 shows factor importance results for all the factors considered in this example.

Table 1. Factor importance results

Factor	Observations	Mean	Standard deviation	Min	Max
Deficit	1,877	0.1973794	0.3941552	−4.902884	15.50016
Public spending	1,877	0.1369482	0.2697644	−10.85175	0.6500276
Consumption	1,877	0.2479453	0.0429293	−0.1079709	0.4484287
Investment	1,877	1.847772	0.3886027	−1.261205	5.092096
Export	1,877	0.2691284	0.3522494	−14.00593	0.8352496
Import	1,877	−0.0319164	0.2415005	−9.790175	0.3935999

Figure 5 shows the time pattern of country GDP growth RS to the growth in fixed investments. As can be seen, from 2000–2010, Spain was the country with the highest GDP growth responsiveness to fixed investment, while Great Britain was the best performer during the '90s. The last year (that is, 2012) shows that Germany and Italy are particularly responsive.

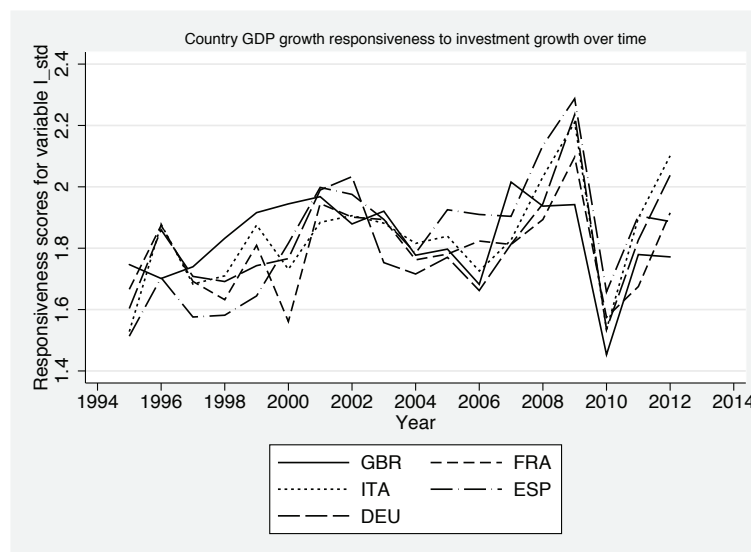


Figure 5. Time pattern of GDP growth RS to the growth in fixed investments: 1990–2012

An interesting piece of information we can obtain from matrix **B** is a ranking of observations according to their RS. For example, we could be interested in knowing the observation in our dataset with the highest investment RS. To obtain such information, we simply sort observations over variable `_bx4` and list them as follows:

```
. sort _bx4
. list countryname year _bx4 if _bx4>=3 & _bx4!=.
```

	countryname	year	_bx4
1865.	Mali	2005	3.036985
1866.	Belarus	1992	3.045852
1867.	Congo, Rep.	2008	3.075329
1868.	Nigeria	2012	3.094015
1869.	Macao SAR, China	2009	3.143898
1870.	Seychelles	2008	3.196258
1871.	Argentina	2002	3.227396
1872.	Indonesia	1999	3.309149
1873.	Trinidad and Tobago	2007	3.3421
1874.	Iran, Islamic Rep.	1994	3.403374
1875.	Bulgaria	1991	3.416531
1876.	Bulgaria	1990	4.174945
1877.	Nigeria	2004	5.092096

We see that Nigeria in 2004 exhibits the highest GDP growth response to investment growth with an RS equal to 5.09; this means that one standard-deviation change in the rate of growth of fixed investment is associated with about a five standard-deviation increase in Nigerian GDP growth, which is rather high if one considers that the average RS over countries and years in this case is around two.

Finally, as a global responsiveness index, we calculate the MUR and show the first and last 10 observations according to this index:

```
. generate MUR = (_bx1 + _bx2 + _bx3 + _bx4 + _bx5 + _bx6)/6
(11,818 missing values generated)
. sort MUR
. list countryname year MUR in 1/10
```

	countryname	year	MUR
1.	Sierra Leone	2000	-3.208911
2.	Nigeria	2004	-.0919238
3.	Sierra Leone	2010	-.0733859
4.	Rwanda	1991	-.0355732
5.	Congo, Dem. Rep.	2003	-.0349704
6.	Nigeria	2007	-.0310322
7.	Venezuela, RB	1997	.0403034
8.	Sierra Leone	2011	.0478267
9.	Azerbaijan	1996	.0880654
10.	Madagascar	2003	.0898289

```
. gsort - MUR
. list countryname year MUR in 1/10
```

	countryname	year	MUR
1.	Bulgaria	1990	.8761727
2.	Bulgaria	1991	.8273186
3.	Congo, Rep.	2008	.779387
4.	Macao SAR, China	2009	.7609017
5.	Mali	2006	.7523293
6.	Argentina	2002	.7505001
7.	Belarus	1992	.7342721
8.	Iran, Islamic Rep.	1994	.7301948
9.	Congo, Dem. Rep.	1994	.7280833
10.	Seychelles	2003	.7250628

Bulgaria in 1990 and 1991 displays the highest average global response to all the drivers considered in this application, while the worst-performing country was Sierra Leone in 2000.

Lastly, we consider the radar graph provided by **rscore** in this second application. The result for 2012 is illustrated in figure 6.

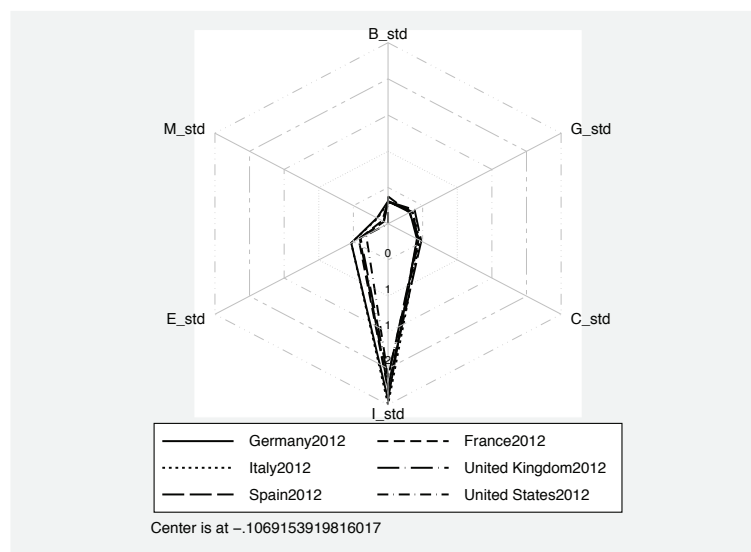


Figure 6. Radar graph for RS country comparison—year 2012

It is evident to see that no remarkable differences arise among European countries (and United States) in 2012. As expected, private fixed investment stands out as the factor with leading responsiveness.

7 Conclusion

The `rscore` command presented in this article can be a useful tool to detect both factor importance and factor heterogeneous response in a regression analysis measuring the impact of a factor x_j on an outcome y . `rscore` also allows the analyst to exploit fixed-effects estimation for a regression of y on x_j , thus mitigating potential factor endogeneity within each factor regression. This command allows one to suitably rank both factors and observations according to their RSSs, providing the analyst with more detailed idiosyncratic information of the response of an outcome y on factor x_j whenever the analysis of such a relation appears to be meaningful.

8 References

- Beran, R., A. Feuerverger, and P. Hall. 1996. On nonparametric estimation of intercept and slope distributions in random coefficient regression. *Annals of Statistics* 24: 2569–2592.
- Cerulli, G. 2014. The impact of technological capabilities on invention: An investigation based on country responsiveness scores. *World Development* 59: 147–165.
- Hoderlein, S., J. Klemelä, and E. Mammen. 2010. Analyzing the random coefficient model nonparametrically. *Econometric Theory* 26: 804–837.
- Lewbel, A., and K. Pendakur. Forthcoming. Unobserved preference heterogeneity in demand using generalized random coefficients. *Journal of Political Economy*.
- Mander, A. 2007. radar: Stata module to draw radar (spider) plots. Statistical Software Components S456829, Department of Economics, Boston College. <https://ideas.repec.org/c/boc/bocode/s456829.html>.
- Poi, B. P. 2003. From the help desk: Swamy's random-coefficients model. *Stata Journal* 3: 302–308.
- Swamy, P. A. V. B. 1970. Efficient inference in a random coefficient regression model. *Econometrica* 38: 311–323.
- Wooldridge, J. M. 1997. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters* 56: 129–133.
- . 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
- . 2003. Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters* 79: 185–191.
- . 2004. 03.2.1. Fixed effects estimation of the population-averaged slopes in a panel data random coefficient model—Solution. *Econometric Theory* 20: 428–429.

———. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

About the author

Giovanni Cerulli is a researcher at CNR-IRCrES, National Research Council of Italy, Institute for Research on Sustainable Economic Growth. He received a degree in statistics and a PhD in economic sciences from Sapienza University of Rome, and is editor-in-chief of the *International Journal of Computational Economics and Econometrics*. His main research interest is applied microeconometrics, with a focus on counterfactual treatment-effects models for program evaluation. He is the author of the book *Econometric Evaluation of Socio-Economic Programs: Theory and Applications* (Springer, 2015). He has published articles in high-quality, refereed economics journals.