The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 1, pp. 92–106

Regression anatomy, revealed

Valerio Filoso
Department of Economics
University of Naples "Federico II"
Naples, Italy
filoso@unina.it

Abstract. The regression anatomy theorem (Angrist and Pischke, 2009, *Mostly Harmless Econometrics: An Empiricist's Companion* [Princeton University Press]) is an alternative formulation of the Frisch–Waugh–Lovell theorem (Frisch and Waugh, 1933, *Econometrica* 1: 387–401; Lovell, 1963, *Journal of the American Statistical Association* 58: 993–1010), a key finding in the algebra of ordinary least-squares multiple regression models. In this article, I present a command, `reganat`, to implement graphically the method of regression anatomy. This addition complements the built-in Stata command `avplot` in the validation of linear models, producing bidimensional scatterplots and regression lines obtained by controlling for the other covariates, along with several fine-tuning options. Moreover, I provide 1) a fully worked-out proof of the regression anatomy theorem and 2) an explanation of how the regression anatomy and the Frisch–Waugh–Lovell theorems relate to partial and semipartial correlations, whose coefficients are informative when evaluating relevant variables in a linear regression model.

Keywords: st0285, reganat, regression anatomy, Frisch–Waugh–Lovell theorem, linear models, partial correlation, semipartial correlation

1 Inside the black box

In the case of a linear bivariate model of the type

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

the ordinary least-squares (OLS) estimator for β has the known simple expression

$$\beta = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} = \frac{\mathrm{Cov}(y_i, x_i)}{\mathrm{Var}(x_i)}$$

In this framework, a bidimensional scatterplot can be a useful graphical device during model building to detect, for instance, the presence of nonlinearities or anomalous data.

When the model includes more than a single independent variable, there is no straightforward equivalent for the estimation of β, and the same bivariate scatterplot between the dependent variable and the independent variable of interest becomes potentially misleading because, in the general case, the independent variables are not orthogonal between them. Consequently, most econometric textbooks limit themselves to providing the formula for the β vector of the type

$$\beta = (X'X)^{-1} X'y$$

and drop altogether any graphical depiction of the relation of interest. Although compact and easy to remember, this formulation is a sort of black box because it hardly reveals anything about what really happens during the estimation of a multivariate OLS model. Furthermore, the link between the β and the moments of the data distribution disappears, buried in the intricacies of matrix algebra.

Luckily, an enlightening interpretation of the β's in the multivariate case exists and has relevant interpreting power. It was originally formulated more than 70 years ago by Frisch and Waugh (1933), revived by Lovell (1963), and implemented in applied econometrics by Angrist and Pischke (2009) under the catchy phrase "regression anatomy". According to this result, given a model with K independent variables, the coefficient β for the kth variable can be written as

$$\beta_k = \frac{\mathrm{Cov}\left(y_i, \widetilde{x}_i^k\right)}{\mathrm{Var}\left(\widetilde{x}_i^k\right)}$$

where $\widetilde{x}_i^k$ is the residual obtained by regressing x_i^k on all remaining $K-1$ independent variables.

The result is striking because it establishes the possibility of breaking a multivariate model with K independent variables into K simpler bivariate models and also sheds light on the machinery of multivariate OLS. This property of OLS does not depend on the underlying data-generating process or on its causal interpretation: it is a purely numerical property of the estimator that holds because of the algebra behind it.

For example, the regression anatomy theorem makes transparent the case of the so-called problem of multicollinearity. In a multivariate model with two variables that are highly linearly related, the theorem implies that for a variable to have a statistically significant β, it must retain sufficient explicative power after the other independent variables have been partialled out. Obviously, this is not likely to happen in a highly multicollinear model because the most variability is between the regressors and not between the residual variable $\widetilde{x}_i^k$ and the dependent variable y.

While this theorem is widely known as a standard result of the matrix algebra of the OLS model, its practical relevance in the modeling process has been overlooked, say Davidson and MacKinnon (1993), most probably because the original articles had a limited scope; it nonetheless illuminated a very general property of the OLS estimator. Hopefully, the introduction of a Stata command that implements it will help to spread its use in econometric practice.

2 The Frisch–Waugh–Lovell theorem

The regression anatomy theorem is an application of the Frisch–Waugh–Lovell (FWL) theorem about the relationship between the OLS estimator and any vertical partitioning of the data matrix $\mathbf{X}$. Originally, Frisch and Waugh (1933) tackled a confusing issue in time-series econometrics. Because many temporal series exhibit a common temporal trend, it was typical during the early days of econometrics to detrend these variables

before entering them in a regression model. The rationale behind this two-stage methodology was to purify the variables from spurious temporal correlation and use only the residual variance in the regression model of interest.

In practice, when an analyst was faced with fitting a model of the type

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i \tag{1}$$

with each variable possibly depending linearly on time, the analyst first estimated a set of K auxiliary regressions of the type

$$x_{ki} = c_k + c_{1k} t + e_{ki}$$

and an analogous regression for the dependent variable,

$$y_i = c_{0y} + c_{1y} t + e_{yi}$$

The analyst then used the residuals from these models to build an analogue to (1):

$$\widetilde{y}_i = \beta'_0 + \beta'_1 \widetilde{x}_{1i} + \cdots + \beta'_k \widetilde{x}_{ki} + \cdots + \beta'_K \widetilde{x}_{Ki} + e'_i$$

Alternatively, other analysts directly entered the time variable in (1) and fit the full model:

$$y_i = \beta^*_0 + \beta^*_1 x_{1i} + \cdots + \beta^*_k x_{ki} + \cdots + \beta^*_K x_{Ki} + dt + e^*_i$$

These two schools of econometric practice debated over the merits and the shortcomings of the respective methods until Frisch and Waugh quite surprisingly demonstrated that the two estimation methods are numerically equivalent; that is, they provide exactly the same results

$$\beta'_k = \beta^*_k$$

and

$$e'_i = e^*_i$$

In broader terms, the theorem applies to any regression model with two or more independent variables that can be partitioned into two groups:

$$\mathbf{y} = \mathbf{X}'_1 \boldsymbol{\beta}_1 + \mathbf{X}'_2 \boldsymbol{\beta}_2 + r \tag{2}$$

Consider the general OLS model $\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + e$, with $\mathbf{X}_{N,K}$. Next partition the $\mathbf{X}$ matrix in the following way: let $\mathbf{X}_1$ be an $N \times K_1$ matrix and let $\mathbf{X}_2$ be an $N \times K_2$ matrix, with $K = K_1 + K_2$. It follows that $\mathbf{X} = (\mathbf{X}_1 \mathbf{X}_2)$. Let us now consider the model

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \boldsymbol{\beta}_2 + e \tag{3}$$

where $\mathbf{M}_1$ is the matrix projecting off the subspace spanned by the columns of $\mathbf{X}_1$. In this formulation, $\mathbf{y}$ and the K_2 columns of $\mathbf{X}_2$ are regressed on $\mathbf{X}_1$; then the vector of residuals $\mathbf{M}_1\mathbf{y}$ is regressed on the matrix of residuals $\mathbf{M}_1\mathbf{X}_2$. The FWL theorem states that the $\boldsymbol{\beta}$'s calculated for (3) are identical to those calculated for (2). A complete proof can be found in advanced econometric textbooks such as those by Davidson and MacKinnon (1993, 19–24) and Ruud (2000, 54–60).

3 The regression anatomy theorem

A straightforward implication of the FWL theorem states that the β_k coefficient also can be estimated without partialling the remaining variables out of the dependent variable y_i. This is exactly the regression anatomy (RA) theorem that Angrist and Pischke (2009) have advanced as a fundamental tool in applied econometrics. In this section, for the sake of simplicity and relevance to my Stata command `reganat`, I provide a proof restricted to the case in which $X_{N,K}$, $K_1 = 1$, and $K_2 = K - 1$, building on the indications provided in Angrist and Pischke (2009).

Theorem 3.1 (Regression anatomy)**.** *Given the regression model*

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i \tag{4}$$

and an auxiliary regression in which the variable x_{ki} is regressed on all the remaining independent variables,

$$x_{ki} = \gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i \tag{5}$$

with $\widetilde{x}_{ki} = x_{ki} - \widehat{x}_{ki}$ being the residual for the auxiliary regression, the parameter β_k can be written as

$$\beta_k = \frac{\mathrm{Cov}(y_i, \widetilde{x}_{ki})}{\mathrm{Var}(\widetilde{x}_{ki})} \tag{6}$$

Proof. To prove the theorem, plug (4) and the residual $\widetilde{x}_{ki}$ from (5) into the covariance $\mathrm{Cov}(y_i, \widetilde{x}_{ki})$ from (6) and obtain

$$\begin{aligned}\beta_k =& \frac{\mathrm{Cov}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, \widetilde{x}_{ki})}{\mathrm{Var}(\widetilde{x}_{ki})} \\ =& \frac{\mathrm{Cov}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, f_i)}{\mathrm{Var}(f_i)}\end{aligned}$$

1. Because by construction $E(f_i) = 0$, it follows that the term $\beta_0 E(f_i) = 0$.

2. Because f_i is a linear combination of all the independent variables with the exception of x_{ki}, it must be that

$$\beta_1 E(f_i x_{1i}) = \cdots = \beta_{k-1} E(f_i x_{k-1i}) = \beta_{k+1} E(f_i x_{k+1i}) = \cdots = \beta_K E(f_i x_{Ki}) = 0$$

3. Consider now the term $E(e_i f_i)$. This can be written as

$$\begin{aligned}E(e_i f_i) &= E(e_i f_i) \\ &= E(e_i \widetilde{x}_{ki}) \\ &= E\left\{e_i \left(x_{ki} - \widehat{x}_{ki}\right)\right\} \\ &= E(e_i x_{ki}) - E(e_i \widehat{x}_{ki})\end{aligned}$$

Because e_i is uncorrelated with any independent variable, it is also uncorrelated with x_{ki}; accordingly, we have $E(e_i x_{ki}) = 0$. With regard to the second term of the subtraction, substituting the predicted value from (5), we get

$$E\left\{e_i\left(\gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki}\right)\right\}$$

Once again, because e_i is uncorrelated with any independent variable, the expected value of the terms is equal to 0. Thus it follows that $E(e_i f_i) = 0$.

4. The only remaining term is $E\left(\beta_k x_{ki} \widetilde{x}_{ki}\right)$. The term x_{ki} can be substituted by using a rewriting of (5) such that

$$x_{ki} = E\left(x_{ki} | X_{-k}\right) + \widetilde{x}_{ki}$$

This gives

$$\begin{aligned} E\left(\beta_k x_{ki} \widetilde{x}_{ki}\right) &= \beta_k E\left[\widetilde{x}_{ki}\left\{E\left(x_{ki}|X_{-k}\right) + \widetilde{x}_{ki}\right\}\right] \\ &= \beta_k \left(E\left\{\widetilde{x}_{ki}^2\right\} + E\left[\left\{E\left(x_{ki}|X_{-k}\right)\widetilde{x}_{ki}\right\}\right]\right) \\ &= \beta_k \mathrm{Var}(\widetilde{x}_{ki}) \end{aligned}$$

which follows directly from the orthogonality between $E\left(x_{ki}|X_{-k}\right)$ and $\widetilde{x}_{ki}$.

5. From previous derivations, we finally get

$$\mathrm{Cov}(y_i, \widetilde{x}_{ki}) = \beta_k \mathrm{Var}(\widetilde{x}_{ki})$$

which completes the proof.

□

4 A comparison between reganat and avplot

Let us sum up our results so far: the value of the coefficient β_k can be obtained by the FWL theorem and the RA theorem. While the FWL theorem states that

$$\beta_k = \frac{\mathrm{Cov}(\widetilde{y}_i, \widetilde{x}_i^k)}{\mathrm{Var}(\widetilde{x}_i^k)}$$

the RA theorem states that

$$\beta_k = \frac{\mathrm{Cov}(y_i, \widetilde{x}_i^k)}{\mathrm{Var}(\widetilde{x}_i^k)}$$

There are good reasons to use both formulations when building a multivariate model: both have advantages and shortcomings.

1. **Variance of residuals**
   The OLS residuals obtained by the FWL theorem and the RA theorem are generally different. In particular, those obtained via the FWL theorem coincide with those obtained for the multivariate full OLS model and are valid for inferences about β_k, while the residuals obtained via the RA theorem tend to be inflated because

$$\mathrm{Var}\,(y_i) \geq \mathrm{Var}\,(\widetilde{y}_i)$$

   This holds true because the variance of y can be written, in the simple case of a univariate model $y_i = \alpha + \beta x_i + \epsilon_i$, as

$$\sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_\epsilon^2$$

   where $\beta^2 \sigma_x^2$ is the variance of $\widetilde{y}$.

2. **Partial and semipartial correlations**
   In a regression model with just one independent variable, the OLS estimator can be written as

$$\beta = \frac{\mathrm{Cov}(y_i, x_i)}{\mathrm{Var}(x_i)} = \rho_{yx} \frac{\sigma_y}{\sigma_x}$$

   where ρ_{yx} is the correlation coefficient between x and y. The same relation applied to a multivariate model provides two alternative expressions when using either the FWL method or the RA method. In the case of the FWL method, we have

$$\beta_k = \frac{\mathrm{Cov}(\widetilde{y}_i, \widetilde{x}_i^k)}{\mathrm{Var}(\widetilde{x}_i^k)} = \rho_{\widetilde{y}\widetilde{x}} \frac{\sigma_{\widetilde{y}}}{\sigma_{\widetilde{x}}}$$

   while in the case of the RA theorem, we have

$$\beta_k = \frac{\mathrm{Cov}(y_i, \widetilde{x}_i^k)}{\mathrm{Var}(\widetilde{x}_i^k)} = \rho_{y\widetilde{x}} \frac{\sigma_y}{\sigma_{\widetilde{x}}}$$

   The term $\rho_{\widetilde{y}\widetilde{x}}$ is the partial correlation coefficient, while $\rho_{y\widetilde{x}}$ is the semipartial correlation coefficient. Because the FWL and the RA methods provide the same estimate for β_k, we can write the relation between the two types of correlation coefficients as

$$\rho_{y\widetilde{x}} = \frac{\sigma_{\widetilde{y}}}{\sigma_y} \rho_{\widetilde{y}\widetilde{x}}$$

   from which is evident that $\rho_{y\widetilde{x}} \leq \rho_{\widetilde{y}\widetilde{x}}$ because the variance of y is larger than the variance of $\widetilde{y}$.

   The advantage of using the semipartial coefficient over the partial coefficient is that the former is expressed in terms of σ_y units, whereas the latter's metrics depend on the independent variable under study. Thus using the semipartial coefficient allows for a comparison of the relative strength of different independent variables.

3. **Semipartial correlations and $\mathbf{R}^2$**
   In a multivariate OLS model, each independent variable's variance can be split into three components:

   a. Variance not associated with y
   b. Variance associated with y and shared with other regressors
   c. Variance associated with y and not shared with other regressors

   When you construct an OLS model, the inclusion of a new regressor is valuable when the additional explaining power contained in it is not already fully captured by the other K regressors. Accordingly, the new variable must mainly provide the kind of variance denoted with (c).

   A measure of the value of this informative variance for a new regressor is its semipartial correlation coefficient: this fact can be used to decompose the variance in a multivariate model. Under normal conditions, the sum of the squared semipartials can be subtracted from the overall R^2 for the complete OLS regression to get the value of common variance shared by the independent variables with y.

   The squared semipartial coefficient can also be expressed as the gain to the R^2 due to the inclusion of the kth variable, weighted by the portion of unexplained variance. In formula, this is

$$\rho^2_{y\widetilde{x}_k} = \frac{R^2_{\text{with}} - R^2_{\text{without}}}{(1 - R^2_{\text{with}})\,(N - K - 1)}$$

   Finally, a correspondence between the correlation coefficient and the R^2's from either the FWL regression or the RA regression can be established. In the case of the univariate model $y_i = \alpha + \beta x_i + \epsilon_i$, the coefficient of determination R^2 is defined as $\beta^2\sigma_x^2/\sigma_y^2$ and is equal to ρ^2_{yx}, that is, the squared simple correlation coefficient between y and x. In the same fashion, the R^2 from the FWL regression is equal to the squared partial correlation coefficient, while the R^2 from the RA regression is equal to the squared semipartial correlation coefficient.

I must note that Stata includes the official command `avplot`, which graphs the variable $\widetilde{x}_{ki}$ against $\widetilde{y}_{ki}$ (the residual of a regression of y on all variables except the kth). Though germane in scope and complementary in many walks of statistical life, `reganat` is more congruent than `avplot` with the quantitative interpretation of a multivariate linear model: the former permits an appreciation of the original metrics of y_i, while the latter focuses on $\widetilde{y}_{ki}$, whose metrics are less appealing to the general reader.

In the causal interpretation of the regression model (Angrist and Pischke 2009, chap. 1), the coefficient β is the size of the effect of a causing variable on a dependent variable, free of other competing factors. The same logic relies on the concept of ceteris paribus, that is, the evaluation of a cause with all other factors being equal. While the variable $\widetilde{x}_{ki}$ is the statistical counterpart of the causing variable, the variable $\widetilde{y}_{ki}$ is less informative than the original y_i because it is constrained to have a zero mean.

In applied statistical practice—for example, in econometrics (Feyrer, Sacerdote, and Stern 2008)—it is customary to present, early in an article, a bidimensional scatterplot of a dependent variable against an explanator of interest, even though the plot is potentially misleading because the variance shared by other potential confounders is not taken into account. Usually, in later pages, the main explanator is plugged into a set of other explanators to fit a regression model, but any scatterplot of the main relation of interest is seldom presented. This is unfortunate because the valuable graphical information derived from the FWL theorem gets lost. Nonetheless, to be worth the effort, the postestimation graph must resemble the original relation of interest. This is exactly the context in which `reganat` can enrich the visual apparatus available to the applied statistician while saving the original metrics of the variables involved as much as possible.

5 The command reganat

The estimation command `reganat` is written for Stata 10.1. It has not been tested on previous versions of the program.

5.1 Syntax

The command has the following syntax:

`reganat` *depvar varlist* [*if*] [*in*] [, `dis(`*varlist*`)` `label(`*varname*`)` `biscat` `biline` `reg` `nolegend` `nocovlist` `fwl` `semip` `scheme(`*graphical_scheme*`)`]

Just like any other standard OLS model, a single dependent variable and an array of independent variables are required.

By default, when the user specifies K covariates, the command builds a multigraph made of K bidimensional subgraphs. In each of them, the x axis displays the value of each independent variable free of any correlation with the other variables, while the y axis displays the value of the dependent variable. Within each subgraph, the command displays the scatterplot and the corresponding regression line.

5.2 Options

`dis(`*varlist*`)` restricts the output to the variables in *varlist* and excludes the rest. Only the specified *varlist* will be graphed; nonetheless, the other regressors will be used in the background calculations.

`label(`*varname*`)` uses *varname* to label the observations in the scatterplot.

`biscat` adds to each subgraph the scatterplot between the dependent variable and the original regressor under study. The observations are displayed using a small triangle. Because $E(\widetilde{x}_{ki}) = 0$ by construction and because $E(x_{ki})$ is in general different from 0,

the plotting of x_{ki} and $\widetilde{x}_{ki}$ along the same axis requires the variable $E(x_{ki})$ to be shifted by subtracting its mean.

`biline` adds to each subgraph a regression line calculated over the univariate model in which the dependent variable is regressed only on the regressor under study. To distinguish the two regression lines that appear on the same graph, `biline` uses a dashed pattern for the one for the univariate model.

`reg` displays the output of the regression command for the complete model.

`nolegend` prevents the legend from being displayed.

`nocovlist` prevents the list of covariates from being displayed.

`fwl` uses the FWL formulation in place of RA.

`semip` adds a table with a decomposition of the model's variance.

`scheme(`*graphical_scheme*`)` specifies the graphical scheme to be applied to the composite graph. The default is `scheme(sj)`.

6 An example

Consider the following illustrative example of `reganat`, without any pretense of establishing a genuine causality model. Suppose that we are interested in the estimation of a simple hedonic model for the price of cars dependent on their technical characteristics. In particular, we want to estimate the effect, if any, of a car's length on its price.

First, we load the classic `auto.dta` and regress `price` on `length`, obtaining

```
. sysuse auto
(1978 Automobile Data)

. regress price length

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =   16.50
       Model |  118425867     1  118425867           Prob > F      =  0.0001
    Residual |  516639529    72  7175549.01           R-squared     =  0.1865
-------------+------------------------------           Adj R-squared =  0.1752
       Total |  635065396    73  8699525.97           Root MSE      =  2678.7

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   57.20224   14.08047     4.06   0.000     29.13332    85.27115
       _cons |  -4584.899   2664.437    -1.72   0.090    -9896.357     726.559
------------------------------------------------------------------------------
```

The estimated β is positive. Then because other technical characteristics could influence the selling price, we include `mpg` (mileage) and `weight` as additional controls to get

```
. regress price length mpg weight

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   12.98
       Model |  226957412     3  75652470.6           Prob > F      =  0.0000
    Residual |  408107984    70  5830114.06           R-squared     =  0.3574
-------------+------------------------------           Adj R-squared =  0.3298
       Total |  635065396    73  8699525.97           Root MSE      =  2414.6

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |  -104.8682   39.72154    -2.64   0.010    -184.0903   -25.64607
         mpg |  -86.78928   83.94335    -1.03   0.305     -254.209    80.63046
      weight |   4.364798   1.167455     3.74   0.000     2.036383    6.693213
       _cons |   14542.43   5890.632     2.47   0.016      2793.94    26290.93
------------------------------------------------------------------------------
```

With this new estimation, the sign of `length` has become negative. The RA theorem states that this last estimate of β for `length` also could be obtained in two stages, which is exactly the method deployed by the command.

In the first stage, we regress `length` on `mpg` and `weight`:

```
. regress length mpg weight

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  2,    71) =  312.22
       Model |  32497.5726     2  16248.7863           Prob > F      =  0.0000
    Residual |  3695.08956    71  52.0435149           R-squared     =  0.8979
-------------+------------------------------           Adj R-squared =  0.8950
       Total |  36192.6622    73  495.789893           Root MSE      =  7.2141

------------------------------------------------------------------------------
      length |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -.3554659   .2472287    -1.44   0.155    -.8484259     .137494
      weight |    .024967   .0018404    13.57   0.000     .0212973    .0286366
       _cons |   120.1162    10.3219    11.64   0.000     99.53492    140.6975
------------------------------------------------------------------------------
```

Here it becomes clear that `length` and `weight` are remarkably correlated. In the second stage, we get the residual value of `length` conditional on `mpg` and `weight` by using the model just estimated, and then we regress `price` on this residual `reslengthr`.

```
. predict reslengthr, residuals
. regress price reslengthr
      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =    4.92
       Model |  40636131.6     1  40636131.6           Prob > F      =  0.0297
    Residual |   594429265    72  8255962.01           R-squared     =  0.0640
-------------+------------------------------           Adj R-squared =  0.0510
       Total |   635065396    73  8699525.97           Root MSE      =  2873.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  reslengthr |  -104.8682   47.26845    -2.22   0.030    -199.0961   -10.64024
       _cons |   6165.257   334.0165    18.46   0.000     5499.407    6831.107
------------------------------------------------------------------------------
```

The value of the β from this bivariate regression coincides with that obtained from the multivariate model, although the standard errors are not equal because of different degrees of freedom used in the calculation.

The command `reganat` uses the decomposability of the RA theorem to plot the relation between `price` and `length` on a bidimensional Cartesian graph, even though the model we are actually using is multivariate. Actually, the command plots `price` and `reslengthr` by using the following command, which produces the graph in figure 1.

```
. reganat price length mpg weight, dis(length)

Regression Anatomy
------------------------------------------------------------------------------
Dependent variable ...... : price
Independent variables ... : length mpg weight
Plotting ................ : length
```

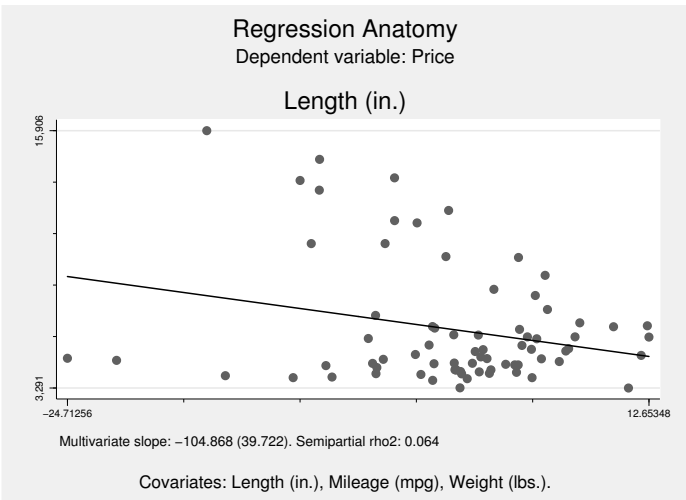

Figure 1. Regression anatomy

The graph displays the variable `length` after partialling out the influence of `mpg` and `weight`. Remarkably, this variable now also assumes negative values, which did not happen in the original data. This happens because residuals have zero expected value by construction; accordingly, the original data have been scaled to have zero mean displayed on the x axis together with residuals.

It is instructive to compare graphically the bivariate model and the multivariate model with the options `biscat` and `biline`. This command produces the graph of figure 2.

```
. reganat price length mpg weight, dis(length) biscat biline

Regression Anatomy
------------------------------------------------------------------------------

Dependent variable ...... : price
Independent variables ... : length mpg weight
Plotting ............... : length
```

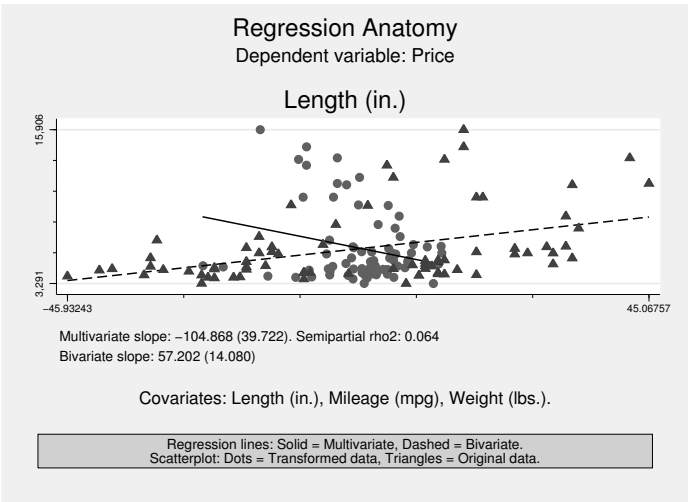

Figure 2. Regression anatomy: Original and transformed data

The graph also displays, for both models, the numerical value of β and its standard error at 95% in parentheses. Furthermore, on the same line, the command displays the squared semipartial correlation coefficient. The calculation is obtained using Stata's built-in `pcorr` command.

The other variables of the model also can be plotted on the graph to check whether the inclusion of additional controls does influence their effect on the dependent variable. This produces the composite graph of figure 3.

```
. reganat price length mpg weight, dis(length weight) biscat biline

Regression Anatomy
------------------------------------------------------------------------------

Dependent variable ...... : price
Independent variables ... : length mpg weight
Plotting ................ : length weight
```

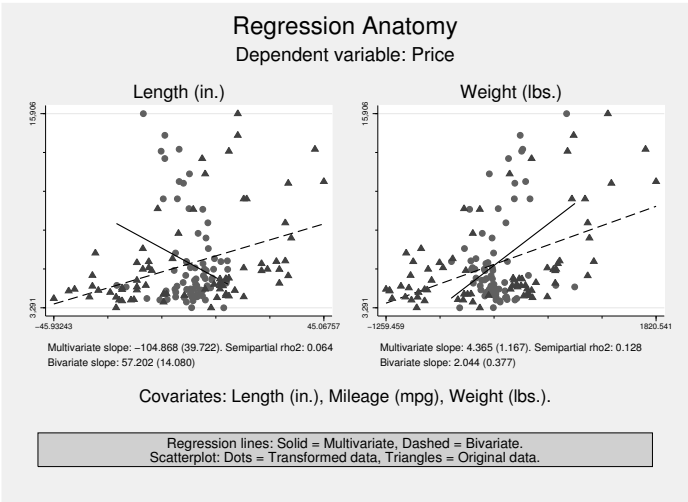

Figure 3. Regression anatomy: Composite graph

The inclusion of additional controls also affects the β for `weight`; in the bivariate model, its value is less than half as much as in the multivariate model (as is clear from the observation of the different slopes in the right panel).

The command is also useful to decompose the model's variance to get an idea of both the idiosyncratic and the joint contributions of the independent variables. Using the option `semip`, we get an additional table with partial correlations, semipartial correlations, squared partial correlations, squared semipartial correlations, relevant significance values, and some summary statistics. The results are obtained using Stata's built-in `pcorr` command.

```
. reganat price length mpg weight, dis(length) semip

Regression Anatomy
------------------------------------------------------------------------------
Dependent variable ...... : price
Independent variables ... : length mpg weight
Plotting ............... : length
(obs=74)
Partial and semipartial correlations of price with
                 Partial   Semipartial      Partial   Semipartial   Significance
   Variable        Corr.         Corr.      Corr.^2       Corr.^2          Value
------------------------------------------------------------------------------
     length      -0.3009       -0.2530       0.0906        0.0640         0.0102
        mpg      -0.1226       -0.0991       0.0150        0.0098         0.3047
     weight       0.4080        0.3582       0.1664        0.1283         0.0004

Model´s variance decomposition                              Value          Perc.
------------------------------------------------------------------------------
Variance explained by the X´s individually                 0.2021         0.5656
Variance common to X´s                                     0.1553         0.4344
------------------------------------------------------------------------------
Variance explained by the model (R-squared)                0.3574
```

The final table decomposes the model's variance: the vector of the three variables `length`, `mpg`, and `weight` explains 35.74% of `price`. This explained variance can be broken into the idiosyncratic contribution of each variable ($6.4\% + 0.98\% + 12.83\% = 20.21\%$) and the common variance (15.53%). In conclusion, around 57% of the model's explained variance can be attributed to the specific contribution of the independent variables, while these same variables share around 43% of `price`'s explained variance.

7 Acknowledgments

The author gratefully acknowledges Joshua Angrist for the invaluable support and encouragement provided during the development of the `reganat` command and for suggesting the title of the article. An anonymous referee deserves thanks for providing several hints, which significantly expanded the scope of this work. The editor's competence and availability have proven crucial throughout the process of submission and revision. Thanks also to Tullio Jappelli, Riccardo Marselli, and Erasmo Papagni for useful suggestions. Any remaining errors are solely the author's responsibility.

8 References

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Davidson, R., and J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. 2nd ed. New York: Oxford University Press.

Feyrer, J., B. Sacerdote, and A. D. Stern. 2008. Will the stork return to Europe and Japan? Understanding fertility within developed nations. *Journal of Economic Perspectives* 22(3): 3–22.

Frisch, R., and F. V. Waugh. 1933. Partial time regressions as compared with individual trends. *Econometrica* 1: 387–401.

Lovell, M. C. 1963. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association* 58: 993–1010.

Ruud, P. A. 2000. *An Introduction to Classical Econometric Theory*. Oxford: Oxford University Press.

About the author

Valerio Filoso is an assistant professor in public finance in the Department of Economics at the University of Naples "Federico II", where he teaches economic analysis of law. His research interests include family and labor economics, the effects of taxation on entrepreneurship, political economy, and monetary governance. He has been a visiting professor in macroeconomics at San Diego State University and a visiting researcher in the Eitan Berglas School of Economics at Tel Aviv University.