The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–12

Regression Anatomy, Revealed

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

Abstract. This paper presents a Stata command, `reganat`, which implements graphically the method of regression anatomy as described by Angrist and Pischke (2009). This tool can help the analyst in the validation of linear models, since it produces a bi-dimensional scatterplot obtained under the control of other covariates.[1]

Keywords: `reganat`, Regression anatomy, Frisch-Waugh-Lovell theorem.

JEL Codes: C5, C51, C3, C52.

Disclaimer: The final version of this paper has been submitted to the Stata Journal.

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 OLS estimator for β has the known simple expression

$$\beta = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} = \frac{\text{cov}(y_i, x_i)}{\text{var}(x_i)}.$$

In the context of such a simple regression framework, a scatterplot can be a useful graphical device during the process of model building to detect, for instance, the presence of nonlinearities or anomalous data.

When the model includes more than a single independent variable, the simple formula for the estimation of β breaks up and a bivariate scatterplot between the dependent variable and a variable of interest is not informative anymore since the regressors need not to be orthogonal among them. Consequently, most econometrics textbooks limit themselves to providing the formula for the β vector of the type

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

Although compact and easy to remember, this formulation is a sort black box, since it hardly reveals anything about what really happens during the estimation of a multi-

1. The author gratefully acknowledges Joshua Angrist for the invaluable support provided during the development of the command. Thanks also to Tullio Jappelli, Riccardo Marselli and Erasmo Papagni for the useful suggestions. All the remaining errors are solely the author's responsibility.

variate OLS model. Furthermore, the link between the β and the moments of the data distribution disappear 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 seventy years ago by Frisch and Waugh (1933), revived by Lovell (1963), and recently brought to a new life 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 k-th variable can be written as

$$\beta_k = \frac{\text{cov}(y_i, \tilde{x}_i^k)}{\text{var}(\tilde{x}_i^k)}$$

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

The results is striking since it establishes the possibility of breaking a multivariate model with K independent variables into K bivariate models and also sheds light into 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 mechanical property of the estimator which 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 which are highly linearly related, our theorem states that for a variable to have a significant β it must retain explicative power after the other independent variables have been partialled out. Obviously, this is not likely to happen in a multicollinear model as the most part of variability is between the regressors and not between the residual variable $\tilde{x}_i^k$ 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, Davidson and MacKinnon (1993) say, most probably because the original articles had a limited scope, but they nonetheless tackled a very general problem. Hopefully, the introduction of a Stata command which implements it will help spreading its use in econometric practice.

2 The theorem

The regression anatomy is an application of the Frisch-Waugh-Lovell theorem about the relationship between the OLS estimator and any vertical partitioning of the data matrix X. The theorem applies to any regression model with two or more independent variables which can be partitioned in two groups

$$y = X_1'\beta_1 + X_2'\beta_2 + r. \tag{1}$$

Consider the general OLS model $y = X'\beta + e$, with $X_{N,K}$. Next, partition the X matrix in the following way: let X_1 be a $N \times K_1$ matrix and X_2 be a $N \times K_2$ matrix,

with $K = K_1 + K_2$. It follows that $X = [X_1 X_2]$. Let us now consider the model

$$M_1 y = M_1 X_2 \beta_2 + e \tag{2}$$

where M_1 is the matrix projecting off the subspace spanned by the columns of X_1. In this formulation, y and the K_2 columns of X_2 are regressed on X_1; then, the vector of residuals $M_1 y$ is regressed on the matrix of residuals $M_1 X_2$. The Frisch-Waugh-Lovell theorem states that the β's calculated for the model (2) are identical to those calculated for the model (1). A complete proof can be found in advanced econometrics textbooks like Davidson and MacKinnon (1993, p. 19–24) or Ruud (2000, p. 54–60), but for the sake of simplicity and relevance to our Stata command `reganat`, here we limit ourself to a simple expansion of the proof provided in Angrist and Pischke (2009) restricted to the case in which $X_{N,K}$, $K_1 = 1$ and $K_2 = K - 1$.

Theorem 1 (Regression anatomy) *Given the regression model*

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + e_i \tag{3}$$

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

$$x_{ki} = \gamma_0 + \gamma_1 x_{1i} + \ldots + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \ldots + \gamma_K x_{Ki} + f_i \tag{4}$$

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

$$\beta_k = \frac{\operatorname{cov}(y_i, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})} \tag{5}$$

Proof. To prove the theorem, plug (3) and the residual $\tilde{x}_{ki}$ from (4) into the covariance $\operatorname{cov}(y_i, \tilde{x}_{ki})$ from (5) and obtain

$$\begin{aligned} \beta_k =& \frac{\operatorname{cov}(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})} \\ =& \frac{\operatorname{cov}(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + e_i, f_i)}{\operatorname{var}(f_i)} \end{aligned} \tag{6}$$

1. Since by construction $E[f_i] = 0$, it follows that the term $\beta_0 E[f_i] = 0$.
2. Since f_i is a linear combination of all the independent variables with the exception of x_{ki}, it must be that

$$\beta_1 E[f_i x_{1i}] = \ldots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \ldots = \beta_K E[f_i x_{Ki}] = 0$$

3. Consider now the term $E[e_i f_i]$. This can be written as

$$\begin{aligned} E[e_i f_i] &= E[e_i f_i] \\ &= E[e_i \tilde{x}_{ki}] \\ &= E[e_i (x_{ki} - \hat{x}_{ki})] \\ &= E[e_i x_{ki}] - E[e_i \hat{x}_{ki}] \end{aligned} \tag{7}$$

Since 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 (4) we get

$$E[e_i(\gamma_0 + \gamma_1 x_{1i} + \ldots + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \ldots + \gamma_K x_{Ki})].$$

Once again, since e_i is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then, it follows that $E[e_i f_i] = 0$.

4. The only remaining term is $E\left[\beta_k x_{ki} \tilde{x}_{ki}\right]$. The term x_{ki} can be substituted using a rewriting of the model (4) such that

$$x_{ki} = E\left[x_{ki}|X_{-k}\right] + \tilde{x}_{ki}.$$

This gives

$$\begin{aligned} E\left[\beta_k x_{ki} \tilde{x}_{ki}\right] &= \beta_k E\left[\tilde{x}_{ki}\left(E\left[x_{ki}|X_{-k}\right] + \tilde{x}_{ki}\right)\right] \\ &= \beta_k \left\{E\left[\tilde{x}_{ki}^2\right] + E\left[\left(E\left[x_{ki}|X_{-k}\right]\tilde{x}_{ki}\right)\right]\right\} \\ &= \beta_k \text{var}(\tilde{x}_{ki}) \end{aligned} \tag{8}$$

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

From previous derivations we finally get

$$\text{cov}(y_i, \tilde{x}_{ki}) = \beta_k \text{var}(\tilde{x}_{ki})$$

which completes the proof. ■

3 The command `reganat`

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

The files `reganat.ado` and `reganat.sthlp` can be freely downloaded from the web address `http://wpage.unina.it/filoso/Stata/` and are also accessible through the SSC system.

3.1 Syntax

The command has the following syntax:

`reganat` *depvar varlist* [*if*] [*in*] [, `disp(vars)` `l(varname)` `biscat` `biline` `reg` `nolegend` `nocovlist` `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 user specifies K covariates, the commands builds a multi-graph made of K bi-dimensional subgraphs. In each of them, the x-axis displays the value of each independent variable *net 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.

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

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

The option `biscat` adds on each subgraph the scatterplot between the dependent variable and the original regressor under study. The observations are displayed using a small triangle. Since $E(\tilde{x}_{ki}) = 0$ by construction, while $E(x_{ki})$ is in general different from zero, the plotting of x_{ki} and $\tilde{x}_{ki}$ along the same axis requires the variable $E(x_{ki})$ to be shifted by subtracting its mean.

The option `biline` adds on 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 which appear on the same graph, the one for the univariate model uses a dashed pattern.

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

The option `nolegend` prevents the legend to be displayed.

The option `nocovlist` prevents the list of covariates to be displayed.

The option `scheme(graphical scheme)` can be used to specify the graphical scheme to be applied to the composite graph. By default, the command uses the `sj` scheme.

4 An example

Consider the following illustrative example of the command, without any pretense of genuine causality. Suppose that we are interested in the estimation of a simple hedonic model for the price of cars as depending 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` dataset and regress `price` on `length`, obtaining

(*Continued on next page*)

```
. sysuse auto, clear
(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, since other technical characteristics could influence the selling price, we include `mpg` (mileage) and `weight` as additional controls and we 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 regression anatomy theorem states that this last estimate of β for `length` could be also obtained in two stages and this is exactly the method deployed by the command.

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

(*Continued on next page*)

```
. 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
------------------------------------------------------------------------------
```

from which 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` using the model just estimated and then regress `price` on this residual `reslength`.

```
. predict reslengthr, r

. regress price reslength

      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 slightly differ.

The command `reganat` uses the decomposability of the regression anatomy theorem to plot the relation between `price` and `length` on a bi-dimensional cartesian graph, *even though the model we are actually using is multivariate*. Actually, the command plots `price` and `reslength` using the command

```
. reganat price length mpg weight, dis(length)
Dependent variable: price
Independent variables: length mpg weight
Plotting: length
```

which produces the graph of fig. (1). The graph displays the variable `length` after partialling out the influence of `mpg` and `weight`. Remarkably, this variable now assumes also negative values, which it 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 in order to be displayed on the x-axis together with residuals.

It is instructive to compare graphically the model obtained using the bivariate model and the multivariate model adding the options `biscat` and `biline`.

```
. reganat price length mpg weight, dis(length) biscat biline
Dependent variable: price
Independent variables: length mpg weight
Plotting: length
```

This command produces the graph of fig. (2). The graph also displays, for both models, the numerical value of β and its standard error at 95% in parentheses.

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

```
. reganat price length mpg weight, dis(length weight) biscat biline
Dependent variable: price
Independent variables: length mpg weight
Plotting: length weight
```

This produces the composite graph of fig. (3). 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 it is clear from the observation of the different slopes in the right panel.

The command is also useful when searching for outliers. Using the option `label` adds labels to the points in the scatterplot.

```
. reganat price length mpg weight, dis(length) label(make)
Dependent variable: price
Independent variables: length mpg weight
Plotting: length
Label variable: make
```

This particular option produces fig. (4) from which it is evident that the observation for *Cadillac Seville* is a candidate for deletion. Dropping that observation and plotting the resulting scatterplot and regression line

```
. drop if make == "Cad. Seville"
(1 observation deleted)

. reganat price length mpg weight, dis(length) label(make)
Dependent variable: price
Independent variables: length mpg weight
Plotting: length
Label variable: make
```

produces the graph of fig. (5) that shows a significant drop in the estimated value of β.

5 References

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

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

Frisch, R., and F. V. Waugh. 1933. Partial Time Regressions as Compared with Individual Trends. *Econometrica* 1(4): 387–401.

Lovell, M. C. 1963. Seasonal Adjustment of Economic Time Series. *Journal of the American Statistical Association* 58: 993–1010.

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

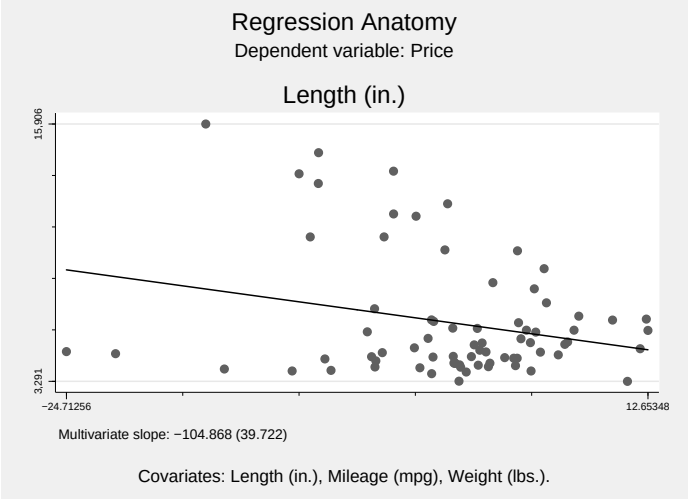

Figure 1: Regression anatomy.

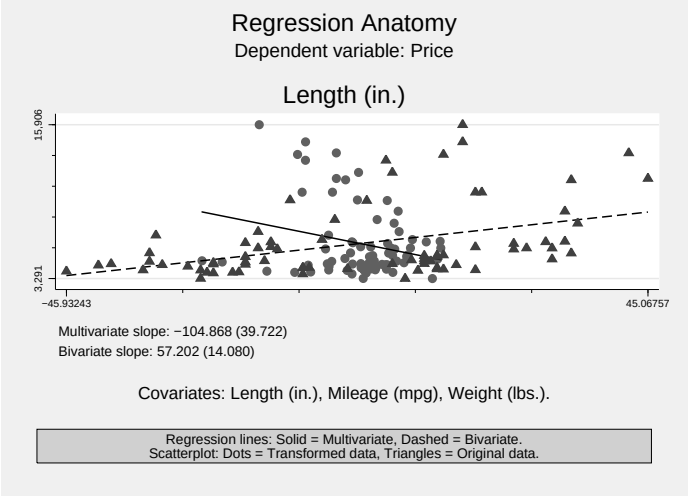

Figure 2: Regression anatomy: original and transformed data.

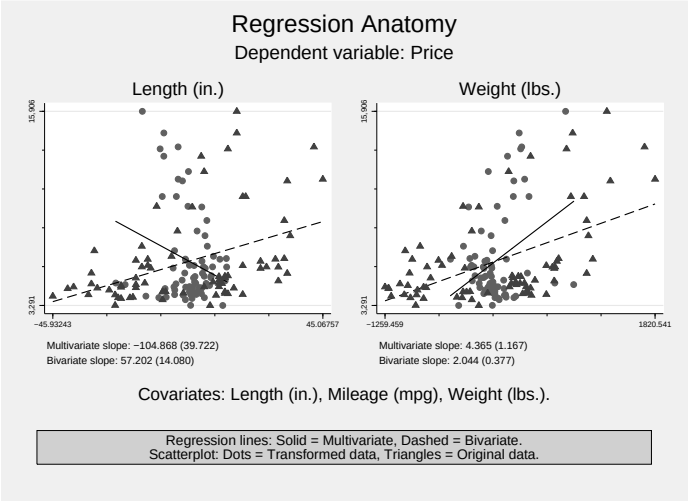

Figure 3: Regression anatomy. Composite graph.

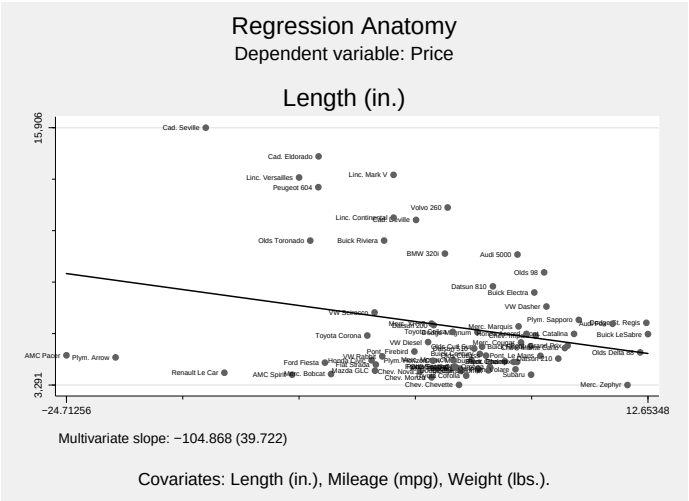

Figure 4: Regression anatomy. A search for outliers.

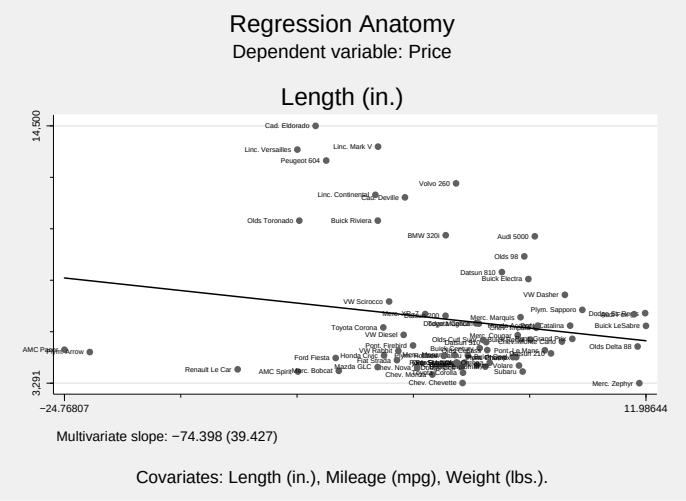

Figure 5: Regression anatomy. The model without an outlier.