Regression Anatomy, Revealed

Filoso, Valerio

University of Naples "Federico II"

10 June 2010

The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–16

Regression Anatomy, Revealed

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

Abstract. The Regression Anatomy (RA) theorem (Angrist and Pischke 2009) is an alternative formulation of the Frisch-Waugh-Lovell (FWL) theorem (Frisch and Waugh 1933; Lovell 1963), a key finding in the algebra of OLS multiple regression models. In this paper, we present a command, `reganat`, to implement graphically the method of RA. This addition complements the built-in Stata command `avplot` in the validation of linear models, producing bidimensional scatterplots and regression lines obtained controlling for the other covariates, along with several fine-tuning options. Moreover, the article provides (1) a fully worked-out proof of the RA theorem and (2) an explanation of how the RA and FWL theorems relate to partial and semipartial correlations, whose coefficients are informative when evaluating relevant variables in a linear regression model. [1]

Keywords: `reganat`, Regression anatomy, Frisch-Waugh-Lovell theorem, Linear models, Partial correlation, Semipartial correlation.

JEL Codes: C13, C5, C51, C52.

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

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

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 econometrics 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 black box, since 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 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 in the world of 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 k-th variable can be written as

$$\beta_k = \frac{\text{cov}(y_i, \tilde{x}_i^k)}{\text{var}(\tilde{x}_i^k)} \tag{1}$$

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

The result 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 purely numerical 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, 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 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 modelling process has been overlooked, Davidson and MacKinnon (1993) say, most probably because the original articles had a limited scope, but it nonetheless illuminated a very general property of the OLS estimator. Hopefully, the introduction of a Stata command which implements it will help spreading its use in econometric practice.

2 The Frisch-Waugh-Lovell theorem

The regression anatomy 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 X. Originally, Frisch and Waugh (1933) tackled a confusing issue in time-series econometrics. Since many temporal series exhibit a common temporal trend, during the early days of econometrics it was common 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 using only the residual variance in the regression model of interest.

In practice, when an analyst was faced with the problem of estimating a model of the following type

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + e_i \tag{2}$$

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

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

plus an analogous regression for the dependent variable

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

and then used the residuals from these models to build an analogue to model (2)

$$\tilde{y}_i = \beta'_0 + \beta'_1 \tilde{x}_{1i} + \ldots + \beta'_k \tilde{x}_{ki} + \ldots + \beta'_K \tilde{x}_{Ki} + e'_i. \tag{3}$$

Alternately, other analysts entered directly the time variable in (2) and estimated the full model

$$y_i = \beta^*_0 + \beta^*_1 x_{1i} + \ldots + \beta^*_k x_{ki} + \ldots + \beta^*_K x_{Ki} + dt + e^*_i. \tag{4}$$

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, viz. they provide exactly the same results, namely

$$\beta'_k = \beta^*_k \tag{5}$$

and

$$e'_i = e^*_i. \tag{6}$$

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

$$y = X'_1 \beta_1 + X'_2 \beta_2 + r. \tag{7}$$

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

$$M_1 y = M_1 X_2 \beta_2 + e \tag{8}$$

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 (8) are identical to those calculated for the model (7). A complete proof can be found in advanced econometrics textbooks like Davidson and MacKinnon (1993, p. 19–24) or Ruud (2000, p. 54–60).

3 The regression anatomy theorem

A straightforward implication of the FWL theorem states that the β_k coefficient can be also estimated *without* partialling the remaining variables out of the dependent variable y_i. This is exactly the *regression anatomy theorem* (RA) which has been advanced by Angrist and Pischke as a fundamental tool in applied econometrics. In this subsection, for the sake of simplicity and relevance to our Stata command `reganat`, we 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 1 (Regression anatomy) *Given the regression model*

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + e_i \tag{9}$$

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

$$x_{ki} = \gamma_0 + \gamma_1 x_{1i} + \ldots + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \ldots + \gamma_K x_{Ki} + f_i \tag{10}$$

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

$$\beta_k = \frac{\operatorname{cov}(y_i, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})} \tag{11}$$

Proof. To prove the theorem, plug (9) and the residual $\tilde{x}_{ki}$ from (10) into the covariance $\operatorname{cov}(y_i, \tilde{x}_{ki})$ from (11) and obtain

$$\begin{aligned} \beta_k =& \frac{\operatorname{cov}(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})} \\ =& \frac{\operatorname{cov}(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + e_i, f_i)}{\operatorname{var}(f_i)} \end{aligned} \tag{12}$$

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

$$\begin{aligned} E[e_i f_i] &= E[e_i f_i] \\ &= E[e_i \tilde{x}_{ki}] \\ &= E[e_i (x_{ki} - \hat{x}_{ki})] \\ &= E[e_i x_{ki}] - E[e_i \hat{x}_{ki}] \end{aligned} \tag{13}$$

   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 (10) 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 (10) such that

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

   This gives

$$\begin{aligned} E\left[\beta_k x_{ki} \tilde{x}_{ki}\right] &= \beta_k E\left[\tilde{x}_{ki} \left(E\left[x_{ki} | X_{-k}\right] + \tilde{x}_{ki}\right)\right] \\ &= \beta_k \left\{E\left[\tilde{x}_{ki}^2\right] + E\left[\left(E\left[x_{ki} | X_{-k}\right] \tilde{x}_{ki}\right)\right]\right\} \\ &= \beta_k \mathrm{var}(\tilde{x}_{ki}) \end{aligned} \tag{14}$$

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

From previous derivations we finally get

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

which completes the proof. ■

4 A comparison between `reganat` and `avplot`

To sum up our results so far, the value of the coefficient β_k can be obtained alternately by the FWL theorem and the RA theorem. While the FWL states that

$$\beta_k = \frac{\mathrm{cov}(\tilde{y}_i, \tilde{x}_i^k)}{\mathrm{var}(\tilde{x}_i^k)} \tag{15}$$

regression anatomy states that

$$\beta_k = \frac{\mathrm{cov}(y_i, \tilde{x}_i^k)}{\mathrm{var}(\tilde{x}_i^k)}. \tag{16}$$

There are good reasons to use both formulations during the process of building a multivariate model, since 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
   $$\text{var}\left(y_i\right) \geq \text{var}\left(\tilde{y}_i\right).$$
   This holds true since the variance of y can be written, in the simple case of a univariate model $y_i = \alpha + \beta x_i + \epsilon_i$, as
   $$\sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_\epsilon^2 \tag{17}$$
   where $\beta^2\sigma_x^2$ is the variance of $\tilde{y}$.

2. **Partial and semipartial correlations**
   In a regression model with just one independent variable, the OLS estimator can be written as
   $$\beta = \frac{\text{cov}(y_i, x_i)}{\text{var}(x_i)} = \rho_{yx}\frac{\sigma_y}{\sigma_x}$$
   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 or the RA methods. In the case of the FWL method we have
   $$\beta_k = \frac{\text{cov}(\tilde{y}_i, \tilde{x}_i^k)}{\text{var}(\tilde{x}_i^k)} = \rho_{\tilde{y}\tilde{x}}\frac{\sigma_{\tilde{y}}}{\sigma_{\tilde{x}}}, \tag{18}$$
   while in the case of the RA theorem we have
   $$\beta_k = \frac{\text{cov}(y_i, \tilde{x}_i^k)}{\text{var}(\tilde{x}_i^k)} = \rho_{y\tilde{x}}\frac{\sigma_y}{\sigma_{\tilde{x}}}. \tag{19}$$
   The term $\rho_{\tilde{y}\tilde{x}}$ is the *partial correlation* coefficient while $\rho_{y\tilde{x}}$ is the *semipartial correlation* coefficient. Since 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
   $$\rho_{y\tilde{x}} = \frac{\sigma_{\tilde{y}}}{\sigma_y}\rho_{\tilde{y}\tilde{x}}$$
   from which is evident that $\rho_{y\tilde{x}} \leq \rho_{\tilde{y}\tilde{x}}$ since the variance of y is larger than the variance of $\tilde{y}$.

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

3. **Semipartial correlations and 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 constructing 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 new regressor is its semipartial correlation coefficient: this fact which 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 k-th variable, weighted by the portion of unexplained variance; in formula:

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

Finally, a correspondence between the correlation coefficient and the R^2's from either the FWL and the RA regression can be established. In the case of an univariate model $y_i = \alpha + \beta x_i + \epsilon_i$ the coefficient of determination R^2 is defined as $\beta^2\sigma^2_x/\sigma^2_y$ and is equal to ρ^2_{yx}, i.e., 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.

It must be noted that Stata includes the official command `avplot` which puts on a graph the variable $\tilde{x}_{ki}$ against $\tilde{y}_{ki}$ (the residual of a regression of y on all variables except the k-th). 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, since the former permits an appreciation of the original metrics of y_i, while the latter focuses on $\tilde{y}_{ki}$, whose metrics is 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, net of other competing factors. The same logic relies on the concept of *ceteris paribus*, i.e., the evaluation of a cause all other factors being equal. While the variable $\tilde{x}_{ki}$ is the statistical counterpart of the causing variable, the variable $\tilde{y}_{ki}$ is less informative than the original y_i, since it is constrained to have a zero mean.

In applied statistical practice – econometrics, for example (Feyrer et al. 2008) – it is customary to present, early on in an article, a bidimensional scatterplot of a dependent

variable against an explanator of interest, even though the plot is potentially misleading as 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 seldom any scatterplot of the main relation of interest is presented. This is unfortunate, since the valuable graphical information derived from the Frisch-Waugh-Lovell theorem gets lost. Nonetheless, to be worth the effort, the post estimation 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(vars) l(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 user specifies K covariates, the commands builds a multi-graph made of K bidimensional 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 `fwl` uses Frisch-Waugh-Lovell formulation in place of RA.

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

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.

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 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 are not equal because of different degrees of freedom used in the calculation.

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)

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

Regression Anatomy
------------------------------------------------------------------------------
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. Furthermore, on the same line, the command displays the squared semipartial correlation coefficient. The calculation is obtained using Stata's built-in command `pcorr` command.

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

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

Feyrer, J., B. Sacerdote, and A. Stern. 2008. Will the stork return to Europe and Japan? Understanding fertility within developed nations. *The 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(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.

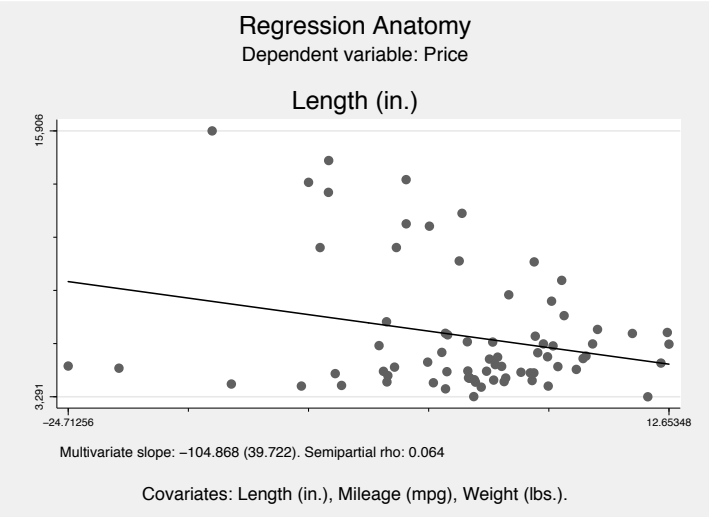

Figure 1: Regression anatomy.

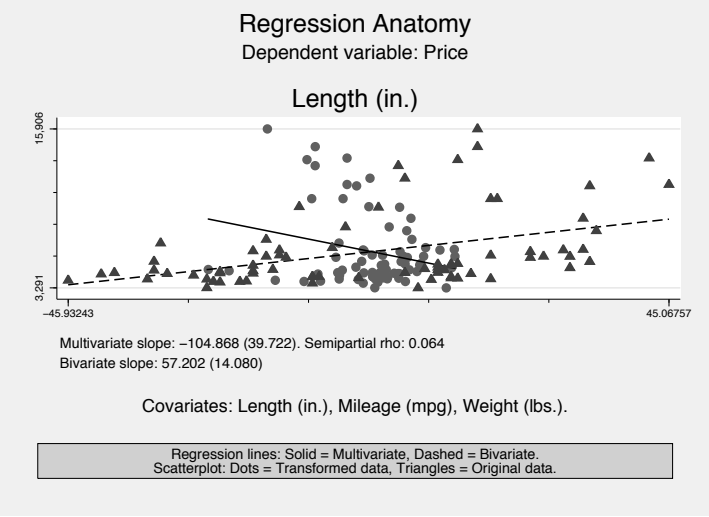

Figure 2: Regression anatomy: original and transformed data.

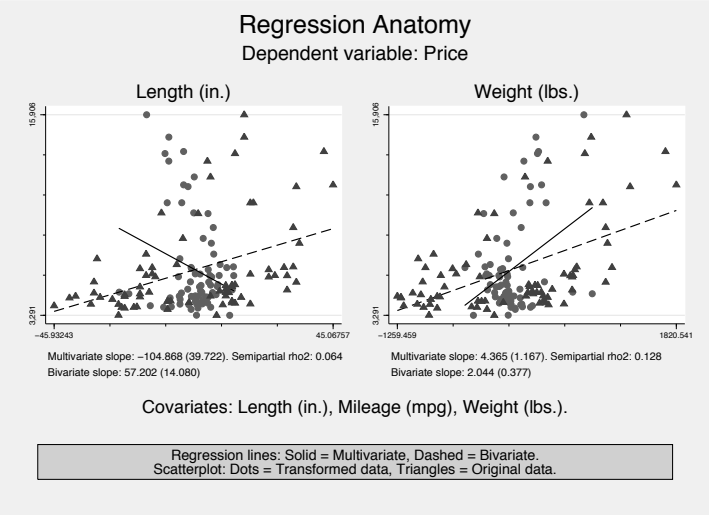

Figure 3: Regression anatomy. Composite graph.