The Stata Journal (2016)
16, Number 1, pp. 5–22

Regressions are commonly misinterpreted

David C. Hoaglin
Independent consultant
Sudbury, MA
dchoaglin@gmail.com

Abstract. Much literature misinterprets results of fitting multivariable models for linear regression, logistic regression, and other generalized linear models, as well as for survival, longitudinal, and hierarchical regressions. For the leading case of multiple regression, regression coefficients can be accurately interpreted via the added-variable plot. However, a common interpretation does not reflect the way regression methods actually work. Additional support for the correct interpretation comes from examining regression coefficients in multivariate normal distributions and from the geometry of least squares. To properly implement multivariable models, one must be cautious when calculating predictions that average over other variables, as in the Stata command `margins`.

Keywords: st0419, regression models, added-variable plot, multivariate normal distribution, geometry of least squares, margins command

1 Introduction

Despite multiple regression's long history and extensive literature, many articles and books are misleading in reporting and interpreting results of fitting regression models. The problems arise in reporting for ordinary least-squares regression, logistic regression, and other generalized linear models, as well as for survival, longitudinal, and hierarchical regressions. Like many other statistical techniques, regression is susceptible to garden-variety forms of abuse, but its greater complexity leads to other less obvious misunderstandings. In what follows, I focus on a major way in which reports and applications of regression analyses often mislead: interpretation of regression coefficients. The correct interpretation is evident in the added-variable plot and the geometry of least squares, as well as from examining regression coefficients in multivariate normal distributions. The common interpretation, regarding the other predictors as held constant, does not accurately reflect how multiple regression works. The misunderstanding in interpreting regression coefficients suggests caution in calculating predictions that average over other variables and in other applications of the Stata command `margins`.

For perspective, the purposes of regression analyses include

- to get a summary;
- to exclude the effect of a variable that might confuse the issue;
- to measure the size of an effect through a regression coefficient;

- to try to discover an empirical law; and
- to make predictions.

Mosteller and Tukey (1977) discuss these and other purposes.

2 Equations for multiple regression

To discuss multiple regression, we need a little notation. One common way to write the relation between the response (or dependent variable) Y and the predictors $X_1, \ldots, X_p$ in multiple regression is

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \tag{1}$$

(usually $X_1 \equiv 1$). This equation represents the underlying or population model; the regression coefficients $\beta_1, \ldots, \beta_p$ are unknown constants to be estimated from the data, and ε is chance variation (noise, disturbance, or error).

By definition, the regression of Y on a set of variables $Z_1, \ldots, Z_m$ (from which predictors may be derived) is the conditional expectation $E(Y|Z_1 = z_1, \ldots, Z_m = z_m)$. Here we allow the possibility that some of the predictors $X_1, \ldots, X_p$ are functions of the same underlying variable (as in a polynomial or a linear spline), and we assume that any appropriate transformations of response and predictors have already been settled. I deliberately avoid referring to predictors as "independent variables", because they are generally not independent in any usual sense. It is difficult to choose an accurate term that has broad appeal. Some people interpret "predictor" as implying causation. Mosteller and Tukey (1977) referred to the X variables as "carriers", a term that seems quite neutral.

In a multiple regression, the definition of each regression coefficient includes the set of other predictors in the equation; that is, their names are part of the definition. G. Udny Yule (1907) introduced a notation that makes the role of the other predictors explicit. For example, we would denote the coefficient of X_2 in (1) by $\beta_{y2\cdot 13\ldots p}$. The first subscript denotes the response variable, the second subscript denotes the predictor to which the coefficient is attached, and the subscripts after the $\cdot$ denote the other predictors. In less abbreviated form, (1) is

$$Y = \beta_{y1\cdot 2\ldots p} X_1 + \beta_{y2\cdot 13\ldots p} X_2 + \cdots + \beta_{yp\cdot 1\ldots p-1} X_p + \varepsilon \tag{2}$$

Each integer 1 through p is an index in the list of predictors. Sometimes, it may be helpful to use the names of the predictors, as in $\beta_{\texttt{gp100m},\texttt{weight}\cdot\texttt{1},\texttt{displacement}}$ (for example, when comparing models that use the same number of predictors, selected from among $X_1, \ldots, X_p$).

Fitting the multiple regression model in (1) to a set of data yields estimates $b_1, \ldots, b_p$ of the regression coefficients $\beta_1, \ldots, \beta_p$. Under the usual assumptions, each b is an unbiased estimate of the corresponding β. We denote an observed value of Y by y, the corresponding given values of $X_1, \ldots, X_p$ by $x_1, \ldots, x_p$, and the corresponding residual by e. Thus the fitted equation corresponding to (1) is

$$y = b_1 x_1 + \cdots + b_p x_p + e$$

and the less abbreviated form corresponding to (2) is

$$y = b_{y1\cdot 2\ldots p} x_1 + b_{y2\cdot 13\ldots p} x_2 + \cdots + b_{yp\cdot 12\ldots p-1} x_p + y_{\cdot 1\ldots p}$$

(now the notation for the residual, $y_{\cdot 1\ldots p}$, shows explicitly the predictors whose contributions have been removed).

Many presentations tend to use the same letters in models that involve different sets of other predictors, which makes it easy to overlook the role of the other predictors in the definition of the coefficient of each predictor. For example, if $2x + 5t$ is a good fit to the data on y, then $-3x + 5(t + x)$ is also a good fit to those data (it gives exactly the same predicted values). In the first fit, 2 is the coefficient of x when t is the other predictor, whereas in the second fit, -3 is the coefficient of x when $t + x$ is the other predictor. By manipulating the choice of the other predictor, I can make the coefficient of x have any value. Mosteller and Tukey (1977, chap. 13) provide instructive examples.

3 Interpretation of regression coefficients

As the notation suggests, $\beta_{y2\cdot 13\ldots p}$ (for example) summarizes the relation between Y and X_2 when X_1, $X_3, \ldots, X_p$ are the other predictors. More specifically, the interpretation of $\beta_{y2\cdot 13\ldots p}$ (or β_2 for short) is that it "tells us how Y responds to change in X_2 after adjusting for simultaneous linear change in the other predictors in the data at hand" (Tukey 1970, chap. 23). This way of stating the effect of X_2 on Y is a direct consequence of the presence of the other predictors. Because the model describes the regression of Y on X_1, X_2, $X_3, \ldots, X_p$ jointly, the coefficient of each predictor accounts for the contributions of the other predictors; that is, it reflects the adjustment for those predictors. The interpretation includes "in the data at hand" because the nature of the adjustment depends on the relations among the predictors in the particular dataset.

The interpretation of a regression coefficient has a straightforward mathematical derivation. Yule (1907) gives an elegant short proof. For the estimated coefficient $\beta_{y2\cdot 13\ldots p}$, the main idea is illustrated by the partial regression plot (also called the "added-variable plot"—for example, in the Stata postestimation command `avplot` after `regress`), in which the vertical coordinate is the residual from the regression of Y on X_1, $X_3, \ldots, X_p$,

$$y_{\cdot 13\ldots p} = y - (b_{y1\cdot 3\ldots p} x_1 + b_{y3\cdot 14\ldots p} x_3 + \cdots + b_{yp\cdot 13\ldots p-1} x_p)$$

and the horizontal coordinate is the residual from the regression of X_2 on $X_1, X_3, \ldots, X_p$,

$$x_{2\cdot 13\ldots p} = x_2 - (b_{21\cdot 3\ldots p} x_1 + b_{23\cdot 14\ldots p} x_3 + \cdots + b_{2p\cdot 13\ldots p-1} x_p)$$

In the regression line (through the origin) for $y_{\cdot 13\ldots p}$ on $x_{2\cdot 13\ldots p}$, the slope is $b_{y2\cdot 13\ldots p}$ (see Cook and Weisberg [1982], section 2.3.2). That is, in the multiple regression of

Y on $X_1, X_2, X_3, \ldots, X_p$, the coefficient of X_2 summarizes the change in Y per unit increase in X_2 after adjusting for simultaneous linear change in $X_1, X_3, \ldots, X_p$ (in the data at hand). Dempster (1969, 160–161) makes a similar point. The interpretation, which applies in the same way to the βs, is also clear from the geometry of least squares, as discussed in section 6.

I avoid the common usage "controlling for" in describing analyses of observational data because it suggests that the variables being "controlled for" are under some sort of "control" (for example, in the way they would be in a randomized controlled trial or in a designed experiment). Referring to a variable as "controlled" implies that it is being held constant. "Adjusting for" is more accurate and straightforward.

For a concrete example of interpreting regression coefficients, I use the data on the foreign cars in the 1978 auto dataset (accessed in Stata), with gallons per 100 miles as the response variable and `weight` and `displacement` as the predictors.

```
. sysuse auto, clear
(1978 Automobile Data)
. generate gp100m = 100/mpg
. label var gp100m "Gallons per 100 miles"
```

For the 22 foreign cars, the command `graph matrix` produces a scatterplot matrix of the 3 variables (figure 1). Gallons per 100 miles has a fairly strong linear relation with `weight` and `displacement`, and the relation between `weight` and `displacement` is even stronger.

```
. graph matrix gp100m weight displacement if foreign==1
```

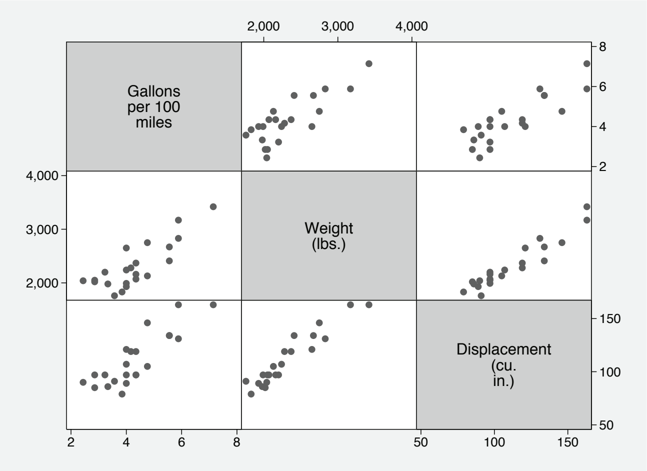

Figure 1. Scatterplot matrix of `gp100m`, `weight`, and `displacement` for the foreign cars in the 1978 automobile data

The command `regress` produces the following results:

```
. regress gp100m weight displacement if foreign == 1
      Source |       SS           df       MS      Number of obs   =        22
-------------+----------------------------------   F(  2,     19)   =     23.86
       Model |  19.6704568         2  9.83522842   Prob > F        =    0.0000
    Residual |  7.83165119        19  .412192168   R-squared       =    0.7152
-------------+----------------------------------   Adj R-squared   =    0.6853
       Total |   27.502108        21  1.30962419   Root MSE        =    .64202

------------------------------------------------------------------------------
      gp100m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0003964   .0010435     0.38   0.708    -.0017877    .0025805
displacement |    .032282   .0181606     1.78   0.091    -.0057286    .0702925
       _cons |   -.195738    .810741    -0.24   0.812    -1.892638    1.501162
------------------------------------------------------------------------------
```

The coefficients, t statistics, and p-values pertain to the contribution of their respective predictors after adjusting for the contributions of the other predictors. Simple regression with `weight` as the predictor yields the expected result from the pattern in figure 1.

```
. regress gp100m weight if foreign == 1
      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  1,    20) =   40.22
       Model |  18.3680109     1  18.3680109           Prob > F      =  0.0000
    Residual |  9.13409716    20  .456704858           R-squared     =  0.6679
-------------+------------------------------           Adj R-squared =  0.6513
       Total |   27.502108    21  1.30962419           Root MSE      =   .6758

------------------------------------------------------------------------------
      gp100m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0021599   .0003406     6.34   0.000     .0014494    .0028703
       _cons |  -.6892425   .8017998    -0.86   0.400    -2.361768    .9832824
------------------------------------------------------------------------------
```

The added-variable plot in figure 2 (produced by the user-written command `favplot`, which can be downloaded from Statistical Software Components using the command `ssc install favplots` and, among other features, allows the user to control the number of decimal places displayed for `b` and `t`) shows the relation of `gp100m` to `displacement` after regression on `weight` has been removed from each. For the line through the origin, the slope (0.0323) and the *t* statistic (1.78) are the same as those for `displacement` in the multiple regression with `weight` and `displacement` as the predictors. Thus 0.0323 gallons per 100 miles per cubic inch summarizes the relation between `gp100m` and `displacement` after adjusting for simultaneous linear change in `weight`. The effect of the adjustment is noticeable. Compare the previous slope with that in the simple regression with `displacement` as the predictor.

```
. favplot displacement, bformat(%7.4f) name(hoaglin_2, replace)
```

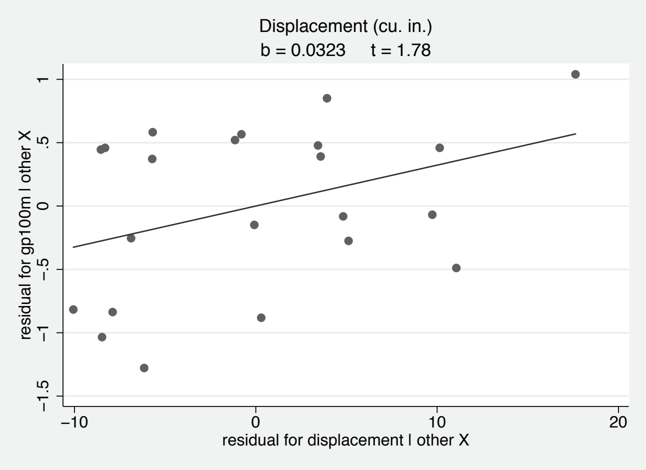

Figure 2. Added-variable plot for displacement in the regression of `gp100m` on weight and `displacement`

```
. regress gp100m displacement if foreign == 1
      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  1,    20) =   49.70
       Model |  19.6109864     1  19.6109864           Prob > F      =  0.0000
    Residual |  7.89112159    20   .39455608           R-squared     =  0.7131
-------------+------------------------------           Adj R-squared =  0.6987
       Total |   27.502108    21  1.30962419           Root MSE      =  .62814

------------------------------------------------------------------------------
      gp100m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
displacement |   .0388401   .0055092     7.05   0.000     .0273482     .050332
       _cons |    -.00723   .6272315    -0.01   0.991    -1.315612    1.301152
------------------------------------------------------------------------------
```

For completeness, figure 3 shows the added-variable plot for `weight`. The adjustment for simultaneous linear change in `displacement` leaves little relation between `gp100m` and `weight`.

```
. favplot weight, bformat(%7.6f) name(hoaglin_3, replace)
```

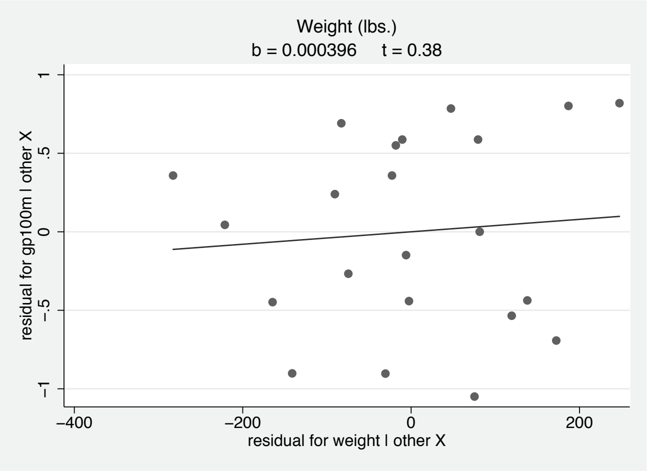

Figure 3. Added-variable plot for weight in the regression of `gp100m` on `weight` and `displacement`

4 A common misinterpretation

In the equation

$$y = b_1x_1 + \cdots + b_px_p + e$$

an estimated regression coefficient (for example, b_2) looks like an ordinary slope, but the reality is more complicated. A common approach interprets b_2 as the average change in Y for a 1-unit increase in X_2 when the other Xs are held constant. A more careful variation recognizes that b_2 is a slope of Y against X_2, so it summarizes change in Y per unit change in X_2. (Of course, when X_2 is an indicator or "dummy" variable, only an increase from 0 to 1 is possible.) Either way, the interpretation is incorrect. It does not reflect the way multiple regression works and should be abandoned. Usually the data were not obtained with the other Xs held constant. And even when some or all other Xs can be held constant, the proper interpretation of b_2 is the one given in section 3.

"Held constant" suggests that one can hold all other Xs fixed for any desired value of X_2. What one can actually do depends on the data. When the other Xs are held constant, even at their means, some changes in X_2 could stray into a region of "predictor space" that is not represented in the data. And when one of the predictors is dichotomous, its mean does not occur in the data. Technically, a point involving such a mean is not in "predictor space" (though it may be surrounded by points that are) because no data can be collected there. On the other hand, various designed experiments collect data to study the effect of some variables when other variables are held constant. Box (1966) discusses examples of passive observation and active (designed) intervention and concludes with the often-quoted remark, "To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)".

The "held constant" interpretation is often justified with a mathematical derivation that uses partial derivatives. If the model is

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

then taking the partial derivative of Y with respect to X_2 yields $\partial Y/\partial X_2 = \beta_2$.

This "proof", however, has two transparent flaws. First, the actual data are nowhere in sight. The partial derivative of Y with respect to X_2 is purely formal. Second, the "proof" is faux mathematics: its assumptions include a key part of the conclusion (holding the other Xs constant). In calculus, the partial derivative is defined by a limiting process that explicitly holds all the other Xs constant and specifies the constant values of those Xs. In general, however, if the data were consulted, they would often say that the other Xs cannot be held constant. For these reasons, taking the partial derivative of the regression function

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p \, (+\varepsilon)$$

with respect to X_2 cannot yield an interpretation of β_2 (or b_2, both of which already reflect the presence of the other predictors). It indicates how the predicted value of Y would change if one could increase X_2 without changing the values of the other predictors. Some such changes in X_2 should generally be possible, because I have assumed that the regression equation is a good fit to the given data, but the justifiable changes are constrained by what the data can support.

We can better understand the situation by recognizing that two distinct purposes of regression are involved. Taking the partial derivative is one aspect of examining the model's use for prediction. Interpreting a coefficient is an aspect of summarizing the effect of that predictor. The partial derivative operates on the model as given, without information on the extent to which the coefficients reflect the contributions of the other predictors.

Prediction that extrapolates substantially beyond the region of predictor space covered by the data is seldom appropriate. And, though less noticeable, interpolation at points that do not occur in the population may not be meaningful. In a particular application, the analyst must check that the data underlying the model support situations in which the variable changes and other variables do not (at least approximately) and check that the variables can be handled in the same way as in applying the results of the analysis. In the example, it is clear from figure 1 that if `displacement` is held constant, the data support changes in `weight` only over a narrow interval, and conversely.

As a simple example in which "held constant" makes no sense, suppose the data come from the model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Here the predictors are 1, x, and x^2, and the subscripts on the βs correspond to the powers of x. It is not possible to change x while holding x^2 constant (except for the trivial change from x to $-x$). This example may seem artificial, but analysts often mechanically add one or more squared terms to models to summarize nonlinearity in the relation between Y and the predictors. I generally advise against this approach because one should not assume that the nonlinearity can be well approximated with a quadratic or higher-order polynomial. It is better to examine the nonlinearity with the aim of uncovering an appropriate functional form. It may be tempting to consider x^2 and x^3 as terms in a Taylor series for a functional relation between Y and x, but the appropriate function may not satisfy the conditions for such an approximation.

In another simple and fairly common example, one predictor uses the product of two other predictors to express their interaction:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

It may be possible to hold $x_1 x_2$ constant while changing x_1, but then x_2 must also change. And changing either x_1 or x_2 while holding the other constant will change $x_1 x_2$.

Both of these examples involve functional dependence of a predictor on one or more other predictors. In the generic regression model, if each of $X_3, \ldots, X_p$ were a function of X_2, then $\partial Y / \partial X_2$ would have the following form:

$$\frac{\partial Y}{\partial X_2} = \beta_2 + \beta_3 \frac{\partial X_3}{\partial X_2} + \cdots + \beta_p \frac{\partial X_p}{\partial X_2}$$

(as before, $X_1 \equiv 1$, so $\partial X_1/\partial X_2 \equiv 0$). Within the limitations of a formal derivative, this gives the correct result for the two examples:

$$\frac{\partial Y}{\partial x} = \beta_1 + 2\beta_2 x$$
$$\frac{\partial Y}{\partial x_1} = \beta_1 + \beta_3 x_2$$

Usually, however, predictors are associated in the data, rather than functionally related. The data supply the information on these associations, and they are accounted for by the interpretation discussed in section 3.

The preceding development applies also when the outcome and the linear predictors are on different scales. In a generalized linear model, for example, the link function, g, relates $\mu_i = E(Y_i)$ to the value of the linear predictor, $\eta : \eta_i = g(\mu_i)$. If h is the inverse of g, so that $\mu_i = h(\eta_i)$, instead of $\partial Y/\partial X_2$ we have

$$\frac{\partial \mu}{\partial X_2} = \frac{dh}{d\eta} \times \frac{\partial \eta}{\partial X_2}$$

Thus, for logistic regression, $g(\mu) = \log_e\{\mu/(1-\mu)\}$, $h(\eta) = 1/(1+e^{-\eta})$, and $dh/d\eta = e^{-\eta}/(1+e^{-\eta})^2 = \mu(1-\mu)$. The coefficients and their interpretation are in the scale of the linear predictor.

5 Regression coefficients in multivariate normal distributions

The interpretation discussed in section 3 also applies to regressions in multivariate normal distributions. This interpretation emphasizes that the coefficients in the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

reflect adjustment for simultaneous linear change in the other predictors. A multivariate normal distribution differs from the usual multiple regression, where the predictors are assumed to be known constants, but the result is the same.

The usual parameters of a multivariate normal distribution are its vector of means ($\boldsymbol{\mu}$) and its covariance matrix ($\mathbf{\Sigma}$). Here it suffices to take each mean equal to 0 and each variance equal to 1 and to focus on the standardized trivariate normal distribution. Thus the three remaining parameters are the off-diagonal elements of $\mathbf{\Sigma}$, which are the pairwise correlations, ρ_{12}, ρ_{13}, and ρ_{23}. We denote the coordinate random variables by X_1, X_2, and X_3.

We regard X_3 as the response variable and X_1 and X_2 as the predictor variables. The regression of X_3 on X_1 and X_2 is linear in X_1 and X_2 and can be written as

$$E\left(X_3|X_1, X_2\right) = \beta_{31\cdot 2} X_1 + \beta_{32\cdot 1} X_2$$

where $\beta_{ij\cdot k}$ is the partial regression coefficient for X_i on X_j when X_k is the other predictor. From the joint density of X_1, X_2, and X_3 and the joint density of X_1 and X_2, it is straightforward to derive the conditional density of X_3 given X_1 and X_2 and to verify that

$$\beta_{31\cdot 2} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{1 - \rho_{12}^2} \quad \text{and} \quad \beta_{32\cdot 1} = \frac{\rho_{23} - \rho_{12}\rho_{13}}{1 - \rho_{12}^2}$$

We arrive at the same expressions if we first adjust for the other predictor. The conditional distribution of X_1 given X_2 has mean $\rho_{12}X_2$ and variance $1 - \rho_{12}^2$, and the conditional distribution of X_3 given X_2 has mean $\rho_{23}X_2$ and variance $1 - \rho_{23}^2$. Then the regression of $X_3 - \rho_{23}X_2$ on $X_1 - \rho_{12}X_2$ has slope

$$\frac{\operatorname{cov}(X_1 - \rho_{12}X_2, X_3 - \rho_{23}X_2)}{\operatorname{var}(X_1 - \rho_{12}X_2)} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{1 - \rho_{12}^2} = \beta_{31\cdot 2}$$

Similarly, the regression of $X_3 - \rho_{13}X_1$ on $X_2 - \rho_{12}X_1$ has slope $\beta_{32\cdot 1}$. Thus the interpretation is that in the regression of X_3 on X_1 and X_2, the coefficient $\beta_{31\cdot 2}$ summarizes the change in X_3 per unit change in X_1 after adjusting for simultaneous linear change in X_2 (that is, after adjusting for the regressions of X_3 and X_1 on X_2).

6 Geometry of least squares

We can also verify the interpretation in section 3 by examining the geometry of least-squares fitting.

Some books illustrate the step from simple regression to multiple regression with the three-predictor model,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

which represents the data in three dimensions, with X_2, X_3, and Y as the axes, and show a plane whose slopes (β_2 and β_3) and intercept (β_1) are estimated by minimizing the sum of squared vertical deviations. Here holding X_3 constant (for example) corresponds to restricting the predicted values of Y to lie on the line formed by the intersection of the fitted plane and the plane perpendicular to the X_3 axis at $X_3 = x_3$. When that line is plotted in the $X_2 - Y$ plane, its slope is b_2 and its intercept is $b_1 + b_3x_3$. If X_2 and X_3 are correlated, the slope of the simple linear regression of Y on X_2 will differ from b_2. The difference between the two slopes is a consequence of their definitions: b_2 reflects the adjustment for X_3, and the slope in the simple regression does not. Thus it is important to look also at the plot of X_3 versus X_2. If the data indicate that a change in X_2 should be accompanied by a corresponding change in X_3, then the predicted value $b_1 + b_2x_2 + b_3x_3$ will change accordingly and will no longer lie on the line in the $X_2 - Y$ plane corresponding to the initial value of X_3.

The geometry of obtaining the estimated coefficients (b_1, b_2, and b_3) by using least squares involves a different representation applicable to any linear regression. Thus we return to the multiple regression with p predictors,

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

in which we have n observations. In the customary matrix notation, $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the vector of data on Y, and the columns of the $n \times p$ matrix $\mathbf{X}$ contain the data on the predictors (considered to be known), as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

If $\mathbf{y}$ contains the true values of Y (that is, $\boldsymbol{\varepsilon} = 0$), then it lies in the subspace spanned by the columns of $\mathbf{X}$ (assumed to have dimension p) and is the linear combination of those columns with coefficients $\beta_1, \ldots \beta_p$. The customary way to recover one of those coefficients (say, β_p) is to change the basis for the subspace, subtracting from X_p the component in the subspace spanned by $X_1, \ldots, X_{p-1}$ and thus replacing X_p as a basis vector by its component orthogonal to that subspace (suitably scaled). Then β_p is the projection of $\mathbf{y}$ on that new basis vector. In the language of multiple regression, β_p is the slope from the regression (through the origin) of y on the residuals from the regression of X_p on $X_1, \ldots, X_{p-1}$ (that is, after adjusting for simultaneous linear change in those other predictors). We get the same β_p by replacing y with the residuals from the regression of y on $X_1, \ldots, X_{p-1}$, so it is appropriate to state the interpretation of β_p in terms of adjusting both y and X_p.

In practice, $\boldsymbol{\varepsilon} \neq 0$, and $\mathbf{y}$ no longer lies in the subspace spanned by the columns of $\mathbf{X}$. The least-squares estimates, $\mathbf{b}$, of the regression coefficients, $\boldsymbol{\beta}$, minimize

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

which is the Euclidean distance from $\mathbf{y}$ to that subspace, yielding

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}$$

To see that the interpretation of β_p applies also to b_p, we can obtain $\widehat{\mathbf{y}}$ by applying the "hat matrix", $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, to $\mathbf{y}$: $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. We can then obtain b_p from $\widehat{\mathbf{y}}$ in the same way as we obtained β_p above.

7 Implications for applications of the Stata command margins

The workings of multiple regression have important implications for use of the Stata command `margins`, which calculates statistics "from predictions of a previously fit model at fixed values of some covariates and averaging or otherwise integrating over the remaining covariates" (StataCorp 2015, 1354). The analyst must demonstrate that the resulting combinations of values of the covariates are meaningful and supported by the data.

To illustrate, we use example 1 in the PDF documentation for `margins`, which involves the regression of `y` on `sex` and `group` in an artificial 3,000-observation dataset. The cross-classification of the two predictors shows different distributions of males and females over the three groups.

```
. webuse margex, clear
(Artificial data for margins)

. tabulate group sex, column

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |          sex
     group |      male     female |     Total
-----------+----------------------+----------
         1 |       215        984 |     1,199 
           |     14.35      65.51 |     39.97 
-----------+----------------------+----------
         2 |       666        452 |     1,118 
           |     44.46      30.09 |     37.27 
-----------+----------------------+----------
         3 |       617         66 |       683 
           |     41.19       4.39 |     22.77 
-----------+----------------------+----------
     Total |     1,498      1,502 |     3,000 
           |    100.00     100.00 |    100.00 
```

The `regress` command yields estimates of the coefficients for `female`, `2.group`, `3.group`, and the constant.

```
. regress y i.sex i.group

      Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  3,  2996) =  152.06
       Model |  183866.077     3  61288.6923           Prob > F      =  0.0000
    Residual |  1207566.93  2996  403.059723           R-squared     =  0.1321
-------------+------------------------------           Adj R-squared =  0.1313
       Total |  1391433.01  2999  463.965657           Root MSE      =  20.076

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
     female  |   18.32202   .8930951    20.52   0.000     16.57088    20.07316
             |
       group |
          2  |   8.037615    .913769     8.80   0.000     6.245937    9.829293
          3  |   18.63922   1.159503    16.08   0.000     16.36572    20.91272
             |
       _cons |   53.32146   .9345465    57.06   0.000     51.48904    55.15388
------------------------------------------------------------------------------
```

With the default response option, `margins` calculates average adjusted predictions (AAPs), treating the sample as if every person were male (respectively, female) as follows:

```
. margins sex

Predictive margins                                Number of obs   =       3000
Model VCE    : OLS

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
       male  |   60.56034   .5781782   104.74   0.000     59.42668    61.69401
     female  |   78.88236   .5772578   136.65   0.000      77.7505    80.01422
------------------------------------------------------------------------------
```

Because the default is `asobserved`, the averaging in this linear regression corresponds to setting `2.group` and `3.group` at their means (0.3727 and 0.2277). The AAPs, 60.56 and 78.88, are meaningful only if it is reasonable to consider an artificial person who is 37.27% in group 2 and 22.77% in group 3 (and, hence, 39.97% in group 1) when data on `y` are available at only six points in "predictor space", corresponding to $\{$male, female$\} \times \{$group 1, group 2, group 3$\}$. Then it must be appropriate to use the same distribution over the three groups for both males and females. Because the data are artificial, I only observe that the combined distribution (39.97%, 37.27%, 22.77%) differs noticeably from the distribution for males and the distribution for females shown in the cross-tabulation. The difference between the AAPs, $78.88 - 60.56 = 18.32$, equals the regression coefficient for `female`. Because the regression is linear in `group`, any distribution over the three groups will, if used for both males and females, yield this same difference.

For an example not based on linear regression, I present one from Williams (2012). Using `nhanes2f.dta` (Second National Health and Nutrition Examination Survey), available from the StataCorp website, Williams (2012) fits a logistic regression model,

```
. webuse nhanes2f, clear
. logit diabetes black female age
  (output omitted)
```

and uses `margins` to obtain adjusted predictions at six values of `age` with `black` and `female` at their means, as follows:

```
. margins, at(age=(20 30 40 50 60 70)) atmeans
  (output omitted)
```

Williams (2012, 313) says, "According to these results, an average 70-year-old (who is again 0.105 `black` and 0.525 `female`) is almost 18 times as likely to have diabetes as an average 20-year-old (11.04% compared with 0.63%)." In practice, an analyst should explain why it is satisfactory to compare an artificial 70-year-old and an artificial 20-year-old who are both 0.105 `black` and 0.525 `female` when data on `diabetes` are available at only four points in the "factor space": $(\texttt{black}, \texttt{female}) = (0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. In `nhanes2f.dta`, the 20-year-olds ($n = 244$) are 0.123 `black` and

0.578 female, and the 70-year-olds ($n = 234$) are 0.064 black and 0.500 female. The overall fractions may be a satisfactory combination for comparisons, but an analyst should first look at 20-year-olds' and 70-year-olds' predicted probabilities of diabetes at each combination of black and female that actually appears in the data. The at() option makes it easy to summarize the predicted probabilities of diabetes at a level of detail that is more relevant to individuals. (As Williams [2012] indicates, "These data were collected in the 1980s. Rates of diabetes in the United States are much higher now.") Thus 70-year-old nonblacks (of both sexes) were nearly 18 times as likely as 20-year-olds to have diabetes (9.60% compared with 0.54% for males and 11.02% compared with 0.63% for females), but the corresponding ratios for blacks were about 16. The ratio for black versus nonblack (of both sexes) was about 2 for 20-year-olds and about 1.85 for 70-year-olds. And females (of both ages and both race categories) were roughly 15% more likely than males to have diabetes. Of course, before embracing predictions from a model, one should check how well it fits. In these data, no 20-year-olds had diabetes, and the highest of the four rates for 70-year-olds was 11.11%.

```
. margins, at(age=(20 70) black=(0 1) female=(0 1))
Adjusted predictions                              Number of obs   =      10335
Model VCE    : OIM
Expression   : Pr(diabetes), predict()
1._at        : black           =           0
               female          =           0
               age             =          20
2._at        : black           =           0
               female          =           0
               age             =          70
3._at        : black           =           0
               female          =           1
               age             =          20
4._at        : black           =           0
               female          =           1
               age             =          70
5._at        : black           =           1
               female          =           0
               age             =          20
6._at        : black           =           1
               female          =           0
               age             =          70
7._at        : black           =           1
               female          =           1
               age             =          20
8._at        : black           =           1
               female          =           1
               age             =          70
```

```
----------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------------------
         _at |
          1  |    .005399   .0009014     5.99   0.000     .0036324    .0071656
          2  |   .0959674   .0071057    13.51   0.000     .0820404    .1098943
          3  |   .0062957   .0010318     6.10   0.000     .0042735    .0083179
          4  |   .1102392   .0073229    15.05   0.000     .0958865    .1245919
          5  |   .0110063   .0020999     5.24   0.000     .0068904    .0151221
          6  |   .1787334    .019682     9.08   0.000     .1401573    .2173095
          7  |   .0128223   .0024099     5.32   0.000     .0080989    .0175456
          8  |   .2025559   .0209683     9.66   0.000     .1614589     .243653
----------------------------------------------------------------------------
```

Setting other variables at their means or averaging over them is also part of calculating marginal effects, elasticity, and semielasticities—the response options `dydx()`, `eyex()`, `dyex()`, and `eydx()`. The logic underlying these options, however, is the same as in the "held constant" interpretation of regression coefficients. Except for interdependencies that are made explicit by using factor-variable notation in the estimation command, the calculations for `dydx()` and the related options lead to interpretations that do not reflect the way multiple regression and other multipredictor analyses actually work.

Although the Stata command `margins` (supported by `marginsplot`) offers great power and flexibility for studying predictions from many models, analysts should not mechanically average over other variables. It is essential to determine the region of "predictor space" covered by the data and examine the associations among the predictors.

8 Many books give the incorrect interpretation

Many books mislead readers by using the "held constant" interpretation. The lowest-numbered page where I have seen this problem is page 2 of Vittinghoff et al. (2012), in an introductory example: "In a sense, multipredictor regression analysis allows us to examine the effect of treatment aggressiveness while *holding the other factors constant* [italics original]."

Out of curiosity, I looked at books that I own by Stata Press that contain material related to multiple regression; these books were by Acock (2010), Kohler and Kreuter (2012), Long and Freese (2006), Mitchell (2012), and Rabe-Hesketh and Skrondal (2012). All of them use the incorrect "held constant" interpretation.

Fortunately, some books use the correct general interpretation. These include the books by De Veaux, Velleman, and Bock (2012), Hastie, Tibshirani, and Friedman (2009), and Weisberg (2014).

9 Conclusion

The interpretation of a coefficient as summarizing the relation between a change in Y and the increase in that predictor after adjusting for simultaneous linear change in the other predictors in the data at hand is an important component of a proper understanding of multiple regression and other multipredictor methods. When one makes explicit the role of the set of other predictors in the definition of each coefficient, this mathematically accurate interpretation is a straightforward consequence of the presence of those other predictors in the model. Applied to the usual tables of estimated coefficients, it helps to clarify the meaning of the t statistics and p-values. It also suggests caution in making predictions and comparisons at combinations of predictor values that do not occur in the data. Appreciation of the proper interpretation should help to avoid common misunderstandings in various applications. However, a challenge is to overcome the barrier created by many publications' use of the "held constant" interpretation, which has no place in a proper understanding of multiple regression.

10 Acknowledgments

I thank Alan Agresti, Jeroan Allison, Nicholas J. Cox, Richard Goldstein, Frank E. Harrell, Jr., Stephen D. Kennedy, Terry Speed, and Paul F. Velleman for discussions and comments on earlier versions. I also thank an anonymous reviewer for thoughtful and extensive comments that substantially improved the presentation. Of course, I am responsible for any opinions expressed.

11 References

Acock, A. C. 2010. *A Gentle Introduction to Stata*. 3rd ed. College Station, TX: Stata Press.

Box, G. E. P. 1966. Use and abuse of regression. *Technometrics* 8: 625–629.

Cook, R. D., and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman & Hall.

De Veaux, R. D., P. F. Velleman, and D. E. Bock. 2012. *Stats: Data and Models*. 3rd ed. Boston: Addison–Wesley.

Dempster, A. P. 1969. *Elements of Continuous Multivariate Analysis*. Reading, MA: Addison–Wesley.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Kohler, U., and F. Kreuter. 2012. *Data Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press.

Long, J. S., and J. Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, TX: Stata Press.

Mitchell, M. N. 2012. *Interpreting and Visualizing Regression Models Using Stata*. College Station, TX: Stata Press.

Mosteller, F., and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison–Wesley.

Rabe-Hesketh, S., and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. 3rd ed. College Station, TX: Stata Press.

StataCorp. 2015. *Stata 14 Base Reference Manual*. College Station, TX: Stata Press.

Tukey, J. W. 1970. *Exploratory Data Analysis*. Limited preliminary ed., vol. 2. Reading, MA: Addison–Wesley.

Vittinghoff, E., D. V. Glidden, S. C. Shiboski, and C. E. McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. New York: Springer.

Weisberg, S. 2014. *Applied Linear Regression*. 4th ed. Hoboken, NJ: Wiley.

Williams, R. 2012. Using the margins command to estimate and interpret adjusted predictions and marginal effects. *Stata Journal* 12: 308–331.

Yule, G. U. 1907. On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London, Series A* 79: 182–193.

About the author

David C. Hoaglin is an independent statistical consultant and an adjunct professor in the Department of Quantitative Health Sciences at the University of Massachusetts Medical School. He received a PhD in statistics from Princeton University in 1971. His current research interests include meta-analysis, biostatistics, exploratory data analysis, and shapes of distributions. He is an associate editor for *Annals of Applied Statistics* and a member of the editorial board for *Research Synthesis Methods*.