Teaching of Multiple Regression Should Reflect the Way It Works

David C. Hoaglin[1]
[1]Independent consultant, 73 Hickory Road, Sudbury, MA 01776
Department of Quantitative Health Sciences, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605
dchoaglin@gmail.com

Abstract

Currently multiple linear regression is usually taught in ways that limit students' understanding and lead to mistakes in applications. The most important shortcoming involves the interpretation of regression coefficients, specifically the contribution of the other explanatory variables. Also, when the definitions of the regression coefficients are presented, the role of those other variables is often overlooked. The workings of least-squares regression are straightforward to understand, and students can be given adequate explanations without technical details. The benefits extend to other regression methods, including logistic regression and survival analysis.

Key Words: Regression coefficients, logistic regression, survival analysis

1. Introduction

In some ways we statisticians have been doing a poor job of teaching regression. It's a problem that we, as a profession, should be embarrassed about. But, once we recognize it, the problem is not hard to fix.

In this paper I focus mainly on two aspects of regression. One, the definition of regression coefficients, is fairly minor. The other, interpreting coefficients in multiple regression, is a big deal.

Nothing is new, but the title implies that something is wrong. The problem is that many textbooks give students incomplete and flawed information on multiple regression.

In the sections that follow, I set up notation for multiple regression models, discuss the definition of coefficients and a notation for them, introduce a flawed interpretation of coefficients in multiple regression, explain the proper interpretation, comment on the connection with output from fitting a multiple regression model, draw support for the proper interpretation from the geometry of least squares, look closer at the flawed interpretation and its "proof," and discuss implications.

Hoaglin (2015) discusses most of these topics in greater detail.

2. Notation for Multiple Regression Models

In a discussion of multiple regression, we need notation for models. One common way of writing the relation between the response (or dependent variable) Y and the predictors X_1, …, X_K (and the constant, $X_0 = 1$) in multiple regression is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon .$$

After estimating the coefficients, we have

Data = Fit + Residual

$$Y = b_0 + b_1 X_1 + \cdots + b_K X_K + e .$$

3. Definition of the Coefficients

Before we turn to interpreting regression coefficients, we should give careful attention to how they are defined. The key idea is that all the predictors are in the model together.

When we focus on a particular regression coefficient, we think mainly about the relation between Y and that predictor. An essential part of the definition of the coefficient is the set of other predictors in the model. The notation introduced by Yule (1907) makes it explicit:

$$\beta_{y1\bullet 02\cdots K} \text{ and } b_{y1\bullet 02\cdots K}$$

In the subscripts of these coefficients for X_1 the first character (y) denotes the response variable, the second character (1) denotes the predictor to which the coefficient is attached, and the characters after the $\bullet$ ($02\cdots K$) denote the other predictors.

In explaining the role of the other predictors in the definition of each coefficient, it is easy to start with the simple regression line:

$$Y = \beta_0 + \beta_1 X + \varepsilon \text{ becomes } Y = \beta_{y1\bullet x} + \beta_{yx\bullet 1} X + \varepsilon$$

(the subscripts show the names of the predictors, and I have taken a relaxed attitude toward lower-case versus upper-case letters). The line through the origin

$$Y = \beta X + \varepsilon \text{ becomes } Y = \beta_{yx} X + \varepsilon ,$$

and the subscripts emphasize the difference in definition between the two coefficients of X.

An aside: It is traditional to write the simple regression line as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

(or in an equivalent form). Everyone does it, but placing the intercept at $X = 0$ makes little sense in most applications, and we have to give excuses when β_0 has no useful interpretation. When I have taught regression in recent years, I have taken the following approach to any numerical variable: After examining the behavior of the variable, and before using it as a predictor in a regression model, *ask where it should be centered*. That is, the predictor should be $X - c$ for some appropriate c, not X (the mean of the data on X may be suitable, but it is not an automatic choice for c). Each numerical predictor deserves some thought, and the result will be many fewer nonsensical intercepts. For example, what use is an intercept at Age = 0 when the data contain no observations with Age < 65?!

Sometimes it is more convenient to write

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

and

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

for the same set of data, but we must keep in mind that the definition of β_0, β_1, and β_2 differs between the two models.

4. A Flawed Interpretation

Many books interpret b_j as the (average) change in Y for an increase of 1 unit in X_j *when the other X's are held constant*. This interpretation is straightforward, but it is *just plain wrong*! As I explain in Section 7, it does not reflect the way regression works. It should be abandoned.

5. The Proper Interpretation

If the interpretation in Section 4 is incorrect, how should we interpret a coefficient in a multiple regression when we seek to summarize the effect of a predictor?

In general, β_j and b_j tell us about the average change in Y per unit increase in X_j *after adjusting for simultaneous linear change in the other X's in the data at hand* (Tukey 1970, Chapter 23). That is, β_j and b_j summarize the effect of X_j, adjusting for (or removing) the contributions of the other X's. (Those adjustments are the ones that, together, can be best combined with a multiple of X_j to give a close fit.)

What is the basis for this interpretation? It's straightforward mathematics. For convenience I focus on the highest-numbered predictor and its coefficient in the multiple regression, b_K.

Regress Y on 1, X_1, …, X_{K-1}, producing residuals that we can denote by $Y_{\bullet 1 \cdots K-1}$.

Regress X_K on 1, X_1, …, X_{K-1}, producing residuals $X_{K \bullet 1 \cdots K-1}$.

Then b_K is the slope of the regression line through the origin for $Y_{\bullet 1 \cdots K-1}$ versus $X_{K \bullet 1 \cdots K-1}$ (in Yule's notation, $b_{YK \bullet 1 \cdots K-1}$).

The scatterplot of $Y_{\bullet 1 \cdots K-1}$ versus $X_{K \bullet 1 \cdots K-1}$ is well known. It is a partial regression plot (or added-variable plot). Some authors call b_j a "partial regression coefficient."

Yule (1907, Section 9) gives an elegant proof of the above result. Cook and Weisberg (1982, Section 2.3.2) also give a proof.

In the following hypothetical example, due to Paul F. Velleman of Cornell University, the proper interpretation of a key regression coefficient agrees with common sense. Suppose we have data on personal income and a model that relates persons' Income to their Age (in years), Education (in years), and Work experience (in years) and includes the usual intercept term. In a simple regression of Income on Age, we would expect the coefficient of Age to be positive. The three predictors, however, are positively correlated:

Age = Education + Work experience + constant .

When Age is adjusted for Education and Work experience, what remains is the years that the person was not in school or working: 5 years (if the person took no time out); plus time in the military, Peace Corps, or other service organization; plus time spent raising children. The coefficient of Age in the multiple regression with Income as the dependent variable summarizes the contribution of the years that the person was not in school or working. We would expect the sign of that coefficient to be negative.

For some audiences it will be more effective to start with simple linear regression, before illustrating a partial regression plot with two non-constant predictors.

The Y-residuals are $Y - \bar{y}$.

The X-residuals are $X - \bar{x}$.

For Y-residual versus X-residual, the line through the origin has slope $b_{YX \bullet 1}$:

$$Y - \bar{y} = b_{YX \bullet 1}(X - \bar{x}) + e .$$

In the usual table of output from fitting a multiple regression model (estimates of the coefficients, standard errors, etc.), each coefficient is a partial regression coefficient. The proper interpretation aids in understanding how the information in the table fits together. For example, the P-value from the t-test on the coefficient for a particular predictor reflects the significance of that predictor's contribution to the model after accounting for the contributions of all the other predictors.

6. Geometry of Least Squares

A geometric approach also leads to the proper interpretation for either *b* or β in multiple regression.

Consider a multiple regression with *p* predictors and *n* observations,

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

In the customary matrix notation, $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ is the vector of data on Y, and the columns of the $n \times p$ matrix X contain the data on the predictors (considered to be known):

$$y = X\beta + \varepsilon$$

If *y* contains the true values of *Y* (i.e., $\varepsilon = 0$), then it lies in the subspace spanned by the columns of *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 *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 $X_1, \ldots, X_{p-1}$ (i.e., 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 $\varepsilon \neq 0$, and *y* no longer lies in the subspace spanned by the columns of *X*. The least-squares estimates, *b*, of the regression coefficients, β, minimize

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

the Euclidean distance from *y* to that subspace, yielding

$$\hat{y} = Xb \ .$$

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

7. The Flaws in the Flawed Interpretation

What's wrong with the flawed interpretation (Section 4)? It gives the impression that one can hold all the other *X*'s constant for any desired value of X_j. What one can actually do depends on the data. In some situations it may be possible to hold the other *X*'s constant at certain (perhaps many) combinations of values of those predictors. (I have assumed that I have a good model.) But "many combinations in some situations" is appropriately restrictive; we have to study the data. Among other concerns, we don't want to stray into a region of "predictor space" where we have little or no data.

A common "proof" of the flawed interpretation starts with the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon$$

and takes the partial derivative with respect to X_j: $\partial Y / \partial X_j = \beta_j$. The problem is that this is faux mathematics. It assumes what is to be proved: the definition of the partial derivative explicitly holds the other variables constant. And, importantly, the actual data are nowhere in sight. Thus, *taking a partial derivative (or difference) cannot yield an interpretation of* β_j.

What's going on? The model uses all the predictors together to produce a good fit to the data on *Y*. The contribution (coefficient) of each predictor already takes into account the contributions of the other predictors. The most that taking a partial derivative can do is indicate how the predicted value of *Y* would change *if one could increase* X_j *without changing the other predictors*.

8. Implications

The points about the definition and interpretation of regression coefficients apply to multivariable models generally: logistic regression and other generalized linear models, survival regressions, longitudinal regressions, hierarchical regressions.

A key message is that students should be able to trust that the authors of their textbooks understand the methods they are writing about. When an instructor shows students that the textbook's interpretation of regression coefficients is incorrect, it undermines their confidence in the book: What else do the authors have wrong? Thus, many authors have some revising to do! Instructors should push to make sure that those changes happen.

Acknowledgements

I am grateful to Alan Agresti for comments and suggestions on a preliminary version of the slides for my talk.

References

Cook RD, Weisberg S (1982). *Residuals and Influence in Regression*. New York: Chapman & Hall.

Hoaglin DC (2015). Regressions are commonly misinterpreted. *Stata Journal*, to appear.

Tukey JW (1970). *Exploratory Data Analysis*, limited preliminary ed. Reading, MA: Addison-Wesley.

Yule GU (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.