Many analyses of data by multiple regression and related methods (e.g., logistic regression) involve interpreting coefficients of predictors. Unfortunately, the most common interpretation is oversimplified and often incorrect. A more complicated interpretation correctly covers all situations. In what follows, I briefly review the multiple regression model and some purposes of regression, discuss the common interpretation, present the proper interpretation and illustrate it in an example, and give some technical background.

**Equations for Multiple Regression**

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

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$  

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 $\epsilon$ is chance variation (or noise, disturbance, or error). In the present discussion we assume that any appropriate transformations of response and predictors have already been settled, including the possibility that some of the predictors are functions of the same underlying variable (as in a polynomial or a linear spline). Usually $X_i \equiv 1$. I deliberately avoid referring to predictors as “independent variables,” because they are generally not independent in any of the usual senses.

Fitting the above multiple-regression model to a set of data yields estimates $b_1, \ldots, b_p$ of the regression coefficients $\beta_1, \ldots, \beta_p$. We denote the $n$ observed values of $Y$ by
y_1, \ldots, y_n$, an individual value by $y_i$, the corresponding given values of $X_1, \ldots, X_p$ by $x_{i1}^{\cdot} \ldots, x_{ip}^{\cdot}$, and the corresponding residual by $e_i$. Thus, the fitted equation is

\[ y_i = b_1 x_{i1} + \ldots + b_p x_{ip} + e_i. \]

When we interpret regression coefficients in an actual analysis, we are interpreting the $b$'s.

It is important to recognize that the definition of a regression coefficient includes the set of other predictors in the equation; that is, their names are part of the definition. In his column in the August 2012 issue of the *IMS Bulletin*, Terry Speed discusses the usefulness of the notation introduced by G. Udny Yule in 1907, which makes the role of the other predictors explicit. For example, the coefficient of $X_2$ in the first equation would be $\beta_{y13^{\cdot}p^{\cdot}}$, and its estimate (in the second equation) would be $b_{y23^{\cdot}p^{\cdot}}$. The first subscript denotes the response variable, the second subscript denotes the predictor to which the coefficient is attached, and the subscripts after the $\bullet$ denote the other predictors. For compactness I have not used that notation, but the appendix uses different letters to preserve the distinction. The tendency of many presentations to use the same letters in models that involve different sets of other predictors makes it easy to overlook the differences in definition of the coefficient of a particular predictor. Mosteller and Tukey (1977, Chapter 13) give a number of instructive examples.

**Some Purposes of Regression**

For perspective I note that analysts use regression for a variety of purposes. To illustrate, Mosteller and Tukey (1977) list six:

- to get a summary
- to set aside the effect of a variable that might confuse the issue
- contributions to attempts at causal analysis
• to measure the size of [an] effect through a regression coefficient (This use is fraught with difficulties when there are multiple causes and when various noncausal variables are associated with other causal ones. Mosteller and Tukey give examples in Chapter 13, “Woes of Regression Coefficients.”)

• to try to discover a mathematical empirical law

• for prediction.

These are distinct purposes, and differences among them may shape the development of the regression models. A fuller discussion of each, with examples, would be instructive, but it would take us away from the theme of this note. Interpreting a regression coefficient is one aspect of using that coefficient to measure the size of an effect.

**The Most Common Interpretation**

In the equation

\[ y = b_1 x_1 + \cdots + b_p x_p + e \]

it is common to interpret a fitted regression coefficient, say \( b_k \), as telling us about the change in \( Y \) corresponding to an increase of 1 unit in \( X_k \) when the other \( X \)'s are held constant. A number of textbooks give this interpretation, without qualification, and many reports of analyses rely on it.

The oversimplification arises in giving the impression that one can hold all the other \( X \)'s fixed, for any desired value of \( X_k \). What one can actually do depends on the data. Often, for certain (perhaps even many) combinations of values of the other \( X \)'s, one can hold those variables constant and vary \( X_k \). It should be all right to do this, because we are assuming that we have a good model. But one must pay attention to what the data will support. When the other \( X \)'s are held constant, even at their means, some changes of “1 unit” in \( X_k \) could stray into a region of “predictor space” that has little or no representation in the data. Extrapolation beyond one’s data is often risky. (Of course, when \( X_k \) is an indicator or “dummy” variable, an increase of 1 unit is the only increase possible.) On the other hand, various designed
experiments collect data to study the effect of some variables when other variables are held constant (more in the appendix).

A minor change in wording would make the interpretation more accurate. Because $b_k$ is a slope of $Y$ against $X_k$, it summarizes change in $Y$ per unit change in $X_k$.

Whether a change of 1 unit is meaningful depends on the units (i.e., the scaling) of $X_k$.

For some units it would be massive, and for other units it would be insignificant.

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 $x^2$. It is not possible to change $x$ while holding $x^2$ constant (unless one includes the trivial change from $x$ to $-x$). This example may seem to have more relevance to science and engineering than to medicine or social science, but analysts often mechanically add one or more squared terms to models, as a way of summarizing nonlinearity in the relation between $Y$ and the particular predictor. (I generally advise against this approach, because the nonlinearity is seldom actually quadratic. It’s more fruitful to examine the nonlinearity, using one of a number of analytic strategies, to uncover an appropriate functional form.)

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 $x_1 x_2$.

**The Proper Interpretation**
The appropriate, fully general interpretation of $b_k$ is that it
tells us how $Y$ responds to change in $X_k$ after allowing for
simultaneous change in the other predictors in the data at hand.

This set of words is longer and more complicated than “held constant,” but in
general nothing less will do. Indeed, we would often do well to add “on average”
after “responds.” The additional complexity is an inherent feature of multiple
regression by least squares. Its validity is a matter of straightforward mathematics
(discussed further in the appendix). In reporting on some applications, it may be
enough to say “after allowing for the contributions of.”

Instead of “allowing for” we could say “adjusting for,” because we can arrive at $b_k$ by
regressing $Y$ on the other predictors and saving the residuals, separately regressing
$X_k$ on the other predictors and saving those residuals, and then fitting a straight line
through the origin with the $Y$ residuals as the response and the $X_k$ residuals as the
predictor. The slope of that line is $b_k$. Thus, I avoid the common usage “controlling
for” in describing analyses of observational data, because the variables being
“controlled for” are generally not “under control” in the way they would be in a
randomized controlled trial or in a designed experiment.

What if the data contain two observations (say, two subjects) that differ by 1 unit on
$X_k$ and have the same values on all the other predictors? What does $b_k$ tell us about
their respective predicted values of $Y$? In this instance we are estimating the value
of $Y$ at values of the predictor variables that actually occur in the data. The
regression model has adjusted for the contributions of the other predictors, and $b_k$ is
the difference in the two subjects’ predicted values of $Y$, taking those adjustments
into account.

Suppose that, instead of two individual subjects, we had two groups of subjects
whose values of $X_k$ differed by 1 unit and whose values on the other predictors were
the same. Then, although it may not be what users of regression would have in
mind, the “held constant” wording readily suggests using the difference between the
means of $Y$ in the two groups, without regard to any regression model. That simple
difference would usually not be the same as \( b_k \), which reflects the adjustment for the regression of \( Y \) on the other predictor variables in the full set of data. In the earlier example, in order to focus on the effect of a difference between the two subjects of 1 unit in \( X_k \), it is not necessary to require that they have the same values on all the other predictors; they need only have the same value of the regression function, skipping the term \( b_k x_k \):

\[
b_1 x_1 + \cdots + b_{k-1} x_{k-1} + b_{k+1} x_{k+1} + \cdots + b_p x_p
\]

Frank Harrell (personal communication) has suggested a parallel with predictive mean matching: \( b_k \) could be interpreted as the difference in predicted values for two subjects who are matched on the expression above and whose values of \( X_k \) differ by 1 unit. In effect, the matching amounts to adjusting for the contributions of the other predictors.

**Example**

The data for this example come from the 1984 fixture list of the Scottish Hill Runners Association. The response variable is record time for the race (Time, in minutes), and the predictors are the constant, Distance (in miles), and Climb (in feet). Further information, including the names of the races, appears in the appendix.

Looking at the plots (below) of Time vs. Distance and Time vs. Climb, it seems reasonable to consider a multiple regression in which Distance and Climb enter linearly:

\[
\text{Time} = b_1 + b_2 \times \text{Distance} + b_3 \times \text{Climb} + \text{residual}.
\]

The result is

\[
\text{Time} = -10.4459 + 6.8282 \times \text{Distance} + 0.007634 \times \text{Climb} + \text{residual}.
\]

Because I am focusing on the coefficients, I omit the usual details in the regression output, except for the fact that each of the three estimated coefficients has a large t-statistic, and \( R^2 = 0.97 \). 
The plot of Climb vs. Distance shows that these two predictors are related \((r = 0.82)\), but neither is a substitute for the other.

We interpret \(b_3 = 0.007634\) minutes per foot as the slope of record time against Climb, allowing for the contribution of Distance. In this example 1 unit is not a useful amount of change in Climb. We would do better to restate \(b_3\) as 7.634 minutes per 1000 feet. To explain the basis for the proper interpretation, we look at the sequence of regressions mentioned above.

When we regress Time on Distance alone (and the constant), the result is

\[
\text{Time} = -11.5250 + 8.8664 \times \text{Distance} + \text{residual},
\]

which we can rearrange to put the residual on the left-hand side:

\[
\text{residual} (\text{Time}) = \text{Time} + 11.5250 - 8.8664 \times \text{Distance}.
\]

Similarly, regressing Climb on distance produces the residuals

\[
\text{residual} (\text{Climb}) = \text{Climb} + 141.3408 - 266.9715 \times \text{Distance}.
\]

Now both Time and Climb have been freed of their regression on Distance. In the plot of residual(Time) vs. residual(Climb) the pattern shows a definite slope, and fitting a straight line through the origin yields

\[
\text{residual} (\text{Time}) = 0.007634 \times \text{residual} (\text{Climb}) + \text{residual}.
\]

As expected, the slope of that line, 0.007634, equals \(b_3\), providing numerical confirmation of the proper interpretation.

We continue by replacing residual(Time) and residual(Climb) by their definitions:

\[
\text{Time} + 11.5250 - 8.8664 \times \text{Distance} = 0.007634 \times [\text{Climb} + 141.3408 - 266.9715 \times \text{Distance}] + \text{residual} = 0.007634 \times \text{Climb} + 1.0790 - 2.0381 \times \text{Distance} + \text{residual}.
\]

When we rearrange to leave only Time on the left-hand side, we have

\[
\text{Time} = (1.0790 - 11.5250) + (-2.0381 + 8.8664) \times \text{Distance} + 0.007634 \times \text{Climb} + \text{residual}.
\]

or

\[
\text{Time} = -10.4459 + 6.8282 \times \text{Distance} + 0.007634 \times \text{Climb} + \text{residual}.
\]

We would have arrived at the same result if we had taken Distance and Climb in the other order, removing the regression on Climb from Time and Distance, and so on.
What about "held constant"? In the plot (below) of Climb vs. Distance we see that, for a given value of Distance, Climb can vary substantially less than its overall range in these data. A similar statement applies to holding Climb constant and varying Distance.

Adding a predictor that represents an interaction between Distance and Climb improves the fit of the multiple regression somewhat. After centering Distance at 6 miles and Climb at 1000 feet, the new predictor is $(\text{Distance} - 6)\times(\text{Climb} - 1000)$, and the expanded model is

\[
\text{Time} = 37.3330 + 5.8481 \times (\text{Distance} - 6) + 0.006885 \times (\text{Climb} - 1000) \\
+ 0.0004188 \times (\text{Distance} - 6) \times (\text{Climb} - 1000) + \text{residual} .
\]

The $t$-statistic for the interaction is 2.58, $R^2 = 0.9791$ (versus 0.9744), and the residual $s^2$ is 5.57 (versus 6.07).
Appendix

The Most Common Interpretation

Some may ask, “What’s the problem with the most common interpretation? Doesn’t it have a simple mathematical derivation?” I don’t recall where I first encountered a derivation of that interpretation, but an econometrics text published in 1976 provides a good example: “In the three-variable model,

\[ Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \]

the coefficient \( \beta_2 \) measures the change in \( Y \) associated with a unit change in \( X_2 \) on the assumption that all other values for the remaining explanatory variables are held constant. Likewise the coefficient \( \beta_3 \) measures the change in \( Y \) associated with a unit change in \( X_3 \). In both cases the assumption that the values of the remaining explanatory variables are constant is crucial to our interpretation of the coefficients. On the basis of this assumption we call the unknown slope parameters partial regression coefficients, since they correspond to the values of the partial derivatives of \( Y \) with respect to \( X_2 \) and \( X_3 \).” Earlier in this note, when I discussed the proper interpretation, I focused on the estimated coefficient, \( b_k \); that interpretation applies also to the parameters \( \beta_2 \) and \( \beta_3 \) here.

A key shortcoming of the quotation above is that the actual data are nowhere in sight! The partial derivatives of \( Y \) with respect to \( X_2 \) and \( X_3 \) are purely formal. From looking at the model, it is “obvious” that \( \partial Y / \partial X_2 = \beta_2 \) and \( \partial Y / \partial X_3 = \beta_3 \). In calculus these partial derivatives are defined by a limiting process that explicitly holds all the other \( X \)'s constant. In general, however, if the data were consulted, they would often say that the other \( X \)'s cannot be held constant. The authors of that statement are correct that their assumption is crucial. Whether it is tenable and, if so, to what extent are determined by the data. In a particular application the analyst has the responsibility for checking (1) that the data underlying the model support situations
in which the variable changes and other variables do not (at least approximately) and (2) that the variables can be handled the same way in applying the results of the analysis.

A fundamental problem with the “held constant” interpretation is that it does not reflect the way least-squares regression works.

The Proper Interpretation

The proper interpretation emphasizes the adjustment inherent in the fitting process and avoids making assumptions about holding other predictors constant. (Analysts are still responsible for justifying any substantial extrapolation.) To show how the adjustment enters in, we trace the steps in obtaining an estimated regression coefficient constructively.

Suppose the fitted model

$$\hat{y} = b_1X_1 + \cdots + b_{p-1}X_{p-1} + b_pX_p$$

is an adequate summary of the regression of $Y$ on $X_1, \ldots, X_p$, and focus (without loss of generality) on $b_p$. We can arrive at this fit by first fitting $X_1, \ldots, X_{p-1}$ and then bringing in $X_p$. We can write the first intermediate result as

$$\hat{y} = a_1X_1 + \cdots + a_{p-1}X_{p-1},$$

and the residuals $y - \hat{y}$ as

$$y - (a_1X_1 + \cdots + a_{p-1}X_{p-1}).$$

These residuals are what remains of $y$ after we have removed the combined contributions of $X_1, \ldots, X_{p-1}$. In parallel, we remove from $X_p$ the variation that can be accounted for by regression on $X_1, \ldots, X_{p-1}$, leaving the $X_p$-residuals

$$X_p - (c_1X_1 + \cdots + c_{p-1}X_{p-1}).$$

Then it is straightforward to show that $b_p$ is the slope of the regression line through the origin when the $y$-residuals are the dependent variable and the $X_p$-residuals are
the predictor. (The corresponding scatterplot is an example of an “added-variable plot.”) The result is

\[
y - (a_1 X_1 + \cdots + a_{p-1} X_{p-1}) = b_p \left[ X_p - (c_1 X_1 + \cdots + c_{p-1} X_{p-1}) \right] + e_p,
\]

where the \( e_p \) are the residuals in the full multiple-regression model (to keep the equation balanced). This way of writing the model makes clear the interpretation of \( b_p \) as summarizing how \( y \) responds to change in \( X_p \) after allowing for (i.e., adjusting for) simultaneous change in the other predictors in the data at hand. The adjustment to \( y \) is represented by \( a_1 X_1 + \cdots + a_{p-1} X_{p-1} \), and the adjustment to \( X_p \) is represented by \( c_1 X_1 + \cdots + c_{p-1} X_{p-1} \). Logically, this heuristic derivation makes sense, because \( b_p \) summarizes the unique contribution of \( X_p \) to the regression of \( Y \) on \( X_1, \ldots, X_p \), and we can assess that contribution only when we add \( X_p \) to the model.

Rearranging the preceding equation yields

\[
y = (a_1 X_1 + \cdots + a_{p-1} X_{p-1}) + b_p \left[ X_p - (c_1 X_1 + \cdots + c_{p-1} X_{p-1}) \right] + e_p
\]

and then

\[
y = (a_1 - b_p c_1) X_1 + \cdots + (a_{p-1} - b_p c_{p-1}) X_{p-1} + b_p X_p + e_p.
\]

Thus, in the full regression model

\[
\begin{align*}
b_1 &= a_1 - b_p c_1 \\
\vdots \\
b_{p-1} &= a_{p-1} - b_p c_{p-1}.
\end{align*}
\]

We can see how, when \( X_p \) is added to the model, the coefficients of the other predictors (already in the model) change to make room for it. This fact also makes clear that the coefficient of a particular predictor depends on what other predictors are in the model.
In illustrating applications of his “new system of notation,” Yule (1907, Section 9) gives an elegant proof of the fact that $b_p$ is the slope of the regression line through the origin for the relation between the $y$-residuals and the $X_p$-residuals. In his notation $b_p$ becomes $\beta_{yp1:p-1}$, and that regression becomes

$$y_{1:p-1} = \beta_{yp1:p-1} X_{p1:p-1}.$$  

Mosteller and Tukey (1977, Section 14K) also give a proof.

The Example

The data in the table below were introduced by Atkinson (1986) in a discussion of a paper on regression diagnostics. Each year the Scottish Hill Runners Association publishes a list of hill races that are scheduled for the coming year. In addition to the date, location, starting time, and organizer’s name and address for each race, this fixture list gives the distance, climb, and record time. The data in the table came from the 35 races on the 1984 fixture list.

Basic plotting and diagnostics quickly cast suspicion on the data for the Knock Hill Race: A record time of 1 hour 18 minutes 39 seconds is much too high for a race with a distance of 3 miles and a climb of 350 feet. A copying error seemed likely.

With the help of Geoff Cohen of the University of Edinburgh (Atkinson’s source for the data), I was able to correct this mistake: the 1986 fixture list gave the record for the Knock Hill Race as 16 minutes 7 seconds, still standing from 1979. The present analysis uses the corrected value.

In a plot of Climb vs. Distance (not shown) the points for the Bens of Jura Fell Race and the Lairig Ghru Fun Run stood apart from the rest. After consulting a colleague who had competed in most of the races, Geoff Cohen reported that “the Bens of Jura Fell Race is over exceptionally rough terrain.” For the present example I omitted the data on these two races. When I included them and fitted the model with the interaction term, they still had substantial positive studentized residuals.
<table>
<thead>
<tr>
<th>Name of race</th>
<th>Distance (miles)</th>
<th>Climb (feet)</th>
<th>Record time (hr:min:sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenmantle New Year Dash</td>
<td>2.5</td>
<td>650</td>
<td>16:05</td>
</tr>
<tr>
<td>Carnethy &quot;5&quot; Hill Race</td>
<td>6</td>
<td>2500</td>
<td>48:21</td>
</tr>
<tr>
<td>Craig Dunain Hill Race</td>
<td>6</td>
<td>900</td>
<td>33:39</td>
</tr>
<tr>
<td>Ben Rha Hill Race</td>
<td>7.5</td>
<td>800</td>
<td>45:36</td>
</tr>
<tr>
<td>Ben Lomond Hill Race</td>
<td>8</td>
<td>3070</td>
<td>62:16</td>
</tr>
<tr>
<td>Goatfell Hill Race</td>
<td>8</td>
<td>2866</td>
<td>73:13</td>
</tr>
<tr>
<td>Bens of Jura Fell Race</td>
<td>16</td>
<td>7500</td>
<td>3:24:37</td>
</tr>
<tr>
<td>Cairnpapple Hill Race</td>
<td>6</td>
<td>800</td>
<td>36:22</td>
</tr>
<tr>
<td>Scolty Hill Race</td>
<td>5</td>
<td>800</td>
<td>29:45</td>
</tr>
<tr>
<td>Traprain Law Race</td>
<td>6</td>
<td>650</td>
<td>39:45</td>
</tr>
<tr>
<td>Lairig Ghru Fun Run</td>
<td>28</td>
<td>2100</td>
<td>3:12:40</td>
</tr>
<tr>
<td>Dollar Hill Race</td>
<td>5</td>
<td>2000</td>
<td>43:03</td>
</tr>
<tr>
<td>Lomonds of Fife Hill Race</td>
<td>9.5</td>
<td>2200</td>
<td>65:00</td>
</tr>
<tr>
<td>Cairn Table Hill Race</td>
<td>6</td>
<td>500</td>
<td>44:08</td>
</tr>
<tr>
<td>Eildon Two Hills Race</td>
<td>4.5</td>
<td>1500</td>
<td>26:56</td>
</tr>
<tr>
<td>Cairngorm Hill Race</td>
<td>10</td>
<td>3000</td>
<td>1:12:15</td>
</tr>
<tr>
<td>Seven Hills of Edinburgh Race</td>
<td>14</td>
<td>2200</td>
<td>1:38:25</td>
</tr>
<tr>
<td>Knock Hill Race</td>
<td>3</td>
<td>350</td>
<td>1:18:39</td>
</tr>
<tr>
<td>Black Hill Race</td>
<td>4.5</td>
<td>1000</td>
<td>17:25</td>
</tr>
<tr>
<td>Creag Beag Hill Race</td>
<td>5.5</td>
<td>600</td>
<td>32:34</td>
</tr>
<tr>
<td>Kildoon Hill Race</td>
<td>3</td>
<td>300</td>
<td>15:57</td>
</tr>
<tr>
<td>Meall Ant-Suidhe Hill Race</td>
<td>3.5</td>
<td>1500</td>
<td>27:54</td>
</tr>
<tr>
<td>Half Ben Nevis</td>
<td>6</td>
<td>2200</td>
<td>47:39</td>
</tr>
<tr>
<td>Cow Hill Race</td>
<td>2</td>
<td>900</td>
<td>17:56</td>
</tr>
<tr>
<td>North Berwick Law Race</td>
<td>3</td>
<td>600</td>
<td>18:41</td>
</tr>
<tr>
<td>Creag Dubh Hill Race</td>
<td>4</td>
<td>2000</td>
<td>26:13</td>
</tr>
<tr>
<td>Burnswark Hill Race</td>
<td>6</td>
<td>800</td>
<td>34:26</td>
</tr>
<tr>
<td>Largo Law Race</td>
<td>5</td>
<td>950</td>
<td>28:34</td>
</tr>
<tr>
<td>Griffel Hill Race</td>
<td>6.5</td>
<td>1750</td>
<td>50:30</td>
</tr>
<tr>
<td>Achmony Hill Race</td>
<td>5</td>
<td>500</td>
<td>20:57</td>
</tr>
<tr>
<td>Ben Nevis Race</td>
<td>10</td>
<td>4400</td>
<td>1:25:35</td>
</tr>
<tr>
<td>Knockfarrel Hill Race</td>
<td>6</td>
<td>600</td>
<td>32:23</td>
</tr>
<tr>
<td>Two Breweries Fell Race</td>
<td>18</td>
<td>5200</td>
<td>2:50:15</td>
</tr>
<tr>
<td>Cockleroi Hill Race</td>
<td>4.5</td>
<td>850</td>
<td>28:06</td>
</tr>
<tr>
<td>Moffat Chase</td>
<td>20</td>
<td>5000</td>
<td>2:39:50</td>
</tr>
</tbody>
</table>
Geometry of Least Squares

Some books illustrate the step from simple regression to multiple regression by using the three-predictor model

\[ Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \]

representing the data in three dimensions with \( X_2, X_3, \) and \( Y \) as the axes; and
displaying a plane, whose slopes (\( \beta_2 \) and \( \beta_3 \)) and intercept (\( \beta_1 \)) are estimated by minimizing the sum of squared vertical deviations. In this representation, 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 chosen plane perpendicular to the \( X_3 \) axis.
The geometry of obtaining the estimated coefficients (\( b_1, b_2, \) and \( b_3 \)), however, involves a different representation, applicable to any linear regression. Thus, we return to the multiple regression with \( p \) predictors,

\[ Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \]

for which we have \( n \) observations. To discuss the geometry, we use the customary matrix notation, in which \( y = (y_1, \ldots, y_n)^T \) is the vector of data on \( Y \) and the \( n \times p \) matrix \( X \) contains the data on the predictors:

\[
\begin{align*}
y &= X\beta + \epsilon \\
\hat{y} &= Xb
\end{align*}
\]

The least-squares estimates, \( b \), of the regression coefficients, \( \beta \), minimize

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

the distance (in the usual Euclidean sense) from \( y \) (viewed as a point in \( n \)-space) to the subspace spanned by the columns of \( X \) (the vectors of data on the \( p \) predictors). As usual, we assume that that subspace is \( p \)-dimensional (i.e., the columns of \( X \) are linearly independent). Thus, whereas we usually think of the data on \( Y \) and the predictors as \( n \) points in \((p + 1)\) dimensions, least-squares estimation involves \( n \) dimensions.
The point in the $p$-dimensional subspace that is closest to $y$ is $\hat{y}$, the projection of $y$ onto that subspace. Algebraically, we obtain $\hat{y}$ by applying the “hat matrix,” $H = X(X^TX)^{-1}X^T$ to $y$: $\hat{y} = Hy$. And the vector of estimated regression coefficients is $b = (X^TX)^{-1}X^Ty$. The vector of residuals, $e = y - \hat{y}$, is orthogonal to the $p$-dimensional subspace. Scheffé (1959, Figure 1.3.1) has a diagram that illustrates these relations.

In the geometry of least squares $b_p$, the estimated regression coefficient for $X_p$, comes from the component of $\hat{y}$ that is not accounted for by regression on $X_1, ..., X_{p-1}$. We obtain $b_p$ constructively by projecting $y$ onto the $(p-1)$-dimensional subspace spanned by the first $p-1$ columns of $X$, projecting $X_p$ onto that subspace, and then projecting the vector of $y$-residuals onto the vector of $X_p$-residuals. This construction makes clear the way in which $b_p$ is “allowing for simultaneous change in the other predictors in the data at hand.” If the data do not include subsets of observations in which $X_p$ varies and the other predictors are constant, it is not clear how to impose that assumption in the geometry.

Unless the vectors for all the predictors are orthogonal, the sums of squares for the unique contributions of the predictors (other than the constant predictor) do not add up to the sum of squares for the regression. The same statement applies to the individual predictors’ contributions to $R^2$. If we want pieces that add up to the overall $R^2$, we must add the nonconstant predictors to the model one by one, determine the resulting increases in $R^2$, and report the order. For a model with two nonconstant predictors McCullagh and Nelder (1989, Section 3.6.4) illustrate the results of the two sequences of fitting, both algebraically and in diagrams.

**Designed Experiments**

Many designed experiments explicitly hold some variables constant while systematically varying others. The simplest involve two variables, often called factors, each with a fixed number of values (their “levels” or “versions”). The experimental design randomly assigns one experimental unit to each combination of versions of the two factors (a “completely randomized design”), and the basic model
aims to summarize the contributions of the two factors to the response as the sum of their separate effects:

\[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}. \]

In this notation \( i = 1, \ldots, I \) denotes the version of the first factor, with effects \( \alpha_i \) (usually \( \alpha_1 + \cdots + \alpha_I = 0 \)), \( j = 1, \ldots, J \) denotes the version of the second factor, with effects \( \beta_j \) (usually \( \beta_1 + \cdots + \beta_J = 0 \)), the \( \epsilon_{ij} \) are disturbances (or errors) (independent, with mean 0 and variance \( \sigma^2 \)), and the data \( y_{ij} \) consist of one observation for each combination of \( i \) and \( j \). For each \( i \) and for each \( j \) the data provide information on how \( y \) varies when the other factor is held constant. In the absence of interaction this additive model yields a convenient interpretation: the way in which \( y \) varies with one factor is the same for all versions of the other factor. (The definitions of the effects of each factor, however, include the versions of the other factor that are present in the particular set of data.) In the usual two-way analysis of variance (ANOVA), the fitted effect \( \hat{\alpha}_i \) estimates the average effect of version \( i \) of the first factor (as an offset from the overall level, \( \hat{\mu} \)), and similarly for \( \hat{\beta}_j \). In comparisons among the effects of one factor, the overall level and the effects of the other factor drop out. When we set up this model as a multiple regression, it is more convenient to designate one version of each factor as the reference category. These become part of the definition of \( \mu \), and the nonconstant predictors are indicator variables for the other \( I - 1 \) versions of the first factor and the other \( J - 1 \) versions of the second factor.

If each combination of \( i \) and \( j \) has more than one observation, it is possible to study the interaction of the two factors in detail by using the mean of \( y \) for each combination. The model expands to include \( I \times J \) interaction terms, \( \gamma_{ij} \) (with appropriate constraints):

\[ y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}. \]
Now it is possible to estimate how $y$ changes with one variable when the other variable is constant (within the range of versions of the two factors in the data) and attach standard errors to those estimates. For example, for version $j$ of the second factor we can look at the relation of $y$ to the versions of the first factor via

$$
\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}
$$

(for comparisons we can often look at

$$
\hat{\alpha}_i + \hat{\gamma}_{ij}
$$

because $\hat{\mu}$ and $\hat{\beta}_j$ do not involve $i$). If the interactions cannot be neglected, statements about the effect of each factor must specify the version of the other factor, and the data support holding that version constant.

**References**

*Statistical Science* 1:397-402.

Boca Raton, FL: Chapman & Hall/CRC.


