6
Multiple Regression: the Basics

From lines to planes

Simple linear regression, as we've learned, is a powerful tool for understanding relationships in data. Yet the method suffers from one crucial setback: it can only be used to model the dependence of y upon a single predictor x.

What if, instead, the phenomenon we're interested in depends upon two explanatory factors? For example:

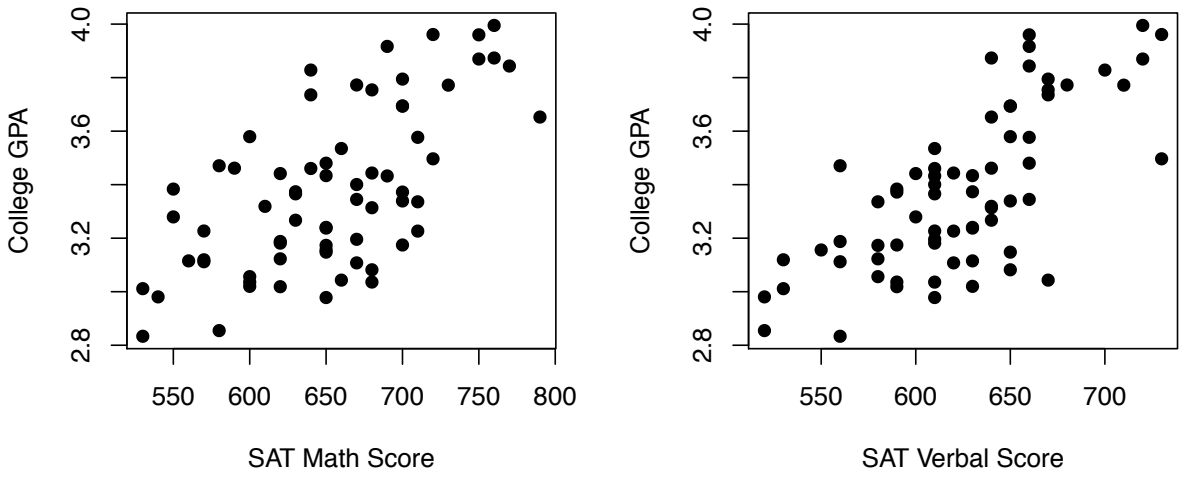

Figure 6.1: Data from a sample of 68 college students. Left: college GPA versus SAT Math score. Right: those same values of college GPA versus SAT Verbal score.

The figure above shows a random sample of 68 college students. On the left, we have plotted each student's college GPA versus his or her SAT Math score. And on the right, we have plotted those same GPA's against SAT Verbal score. Clearly there is a positive

association between college GPA and each of the two components of the SAT.

As in all of our previous examples, of course, neither association is perfect. There is still plenty of residual variation left over, even after accounting for either of the two predictors. Yet it stands to reason that we can predict college GPA better using both parts of the SAT, rather than using either one of them by itself. While the math and verbal scores are not entirely independent of one another, they do measure different skills, and both kinds of skills are important for success in college.

To use the technical term, the three variables—GPA, math score, and verbal score—have a *joint distribution* in three dimensions. Our sample of 68 students is but one sample, out of an enormous number of possible samples, that we might have observed from this joint distribution. The sample is best viewed in all three dimensions of those dimensions, as on the previous page.

The way forward here is simply to add a second term to our regression function. This gives us a linear equation in two variables:

$$E(y_i \mid x_{i,1}, x_{i,2}) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} \,.$$

The two predictors, $x_{i,1}$ and $x_{i,2}$, are the student's SAT Math and Verbal scores, respectively. The response, y_i, is the student's college GPA. Each of these three quantities is specific to an individual, which is why they have subscript i's. The coefficients β_0, β_1, and β_2 are shared among the whole sample.

This specifies the equation of a plane: a two-dimensional linear surface embedded in three dimensions, one which we can imagine slicing roughly through the middle of the point cloud in Figure 6.2. This plane has the same interpretation that the line had in a simple one-dimensional linear regression. If you read off the height of the plane along the y axis, then you know where the predictor variable should be, on average, for a particular point in predictor space, by which we mean a particular pair of values (x_1, x_2).

Of course, in principle, there's no reason to stop at two predictors! We could easily build a regression equation using p different predictors $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$:

$$E(y_i \mid \mathbf{x}_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \,.$$

This is the equation of a p-dimensional hyperplane embedded

We use a bolded $\mathbf{x}_i$ as shorthand to denote the whole vector of predictor values for observation i. That way we don't have to write out $(x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ every time. When writing things out by hand, a little arrow can be used instead, since you obviously can't write things in bold: $\vec{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$. By the same logic, we also write $\vec{\beta}$ for the vector $(\beta_0, \beta_1, \ldots, \beta_p)$.

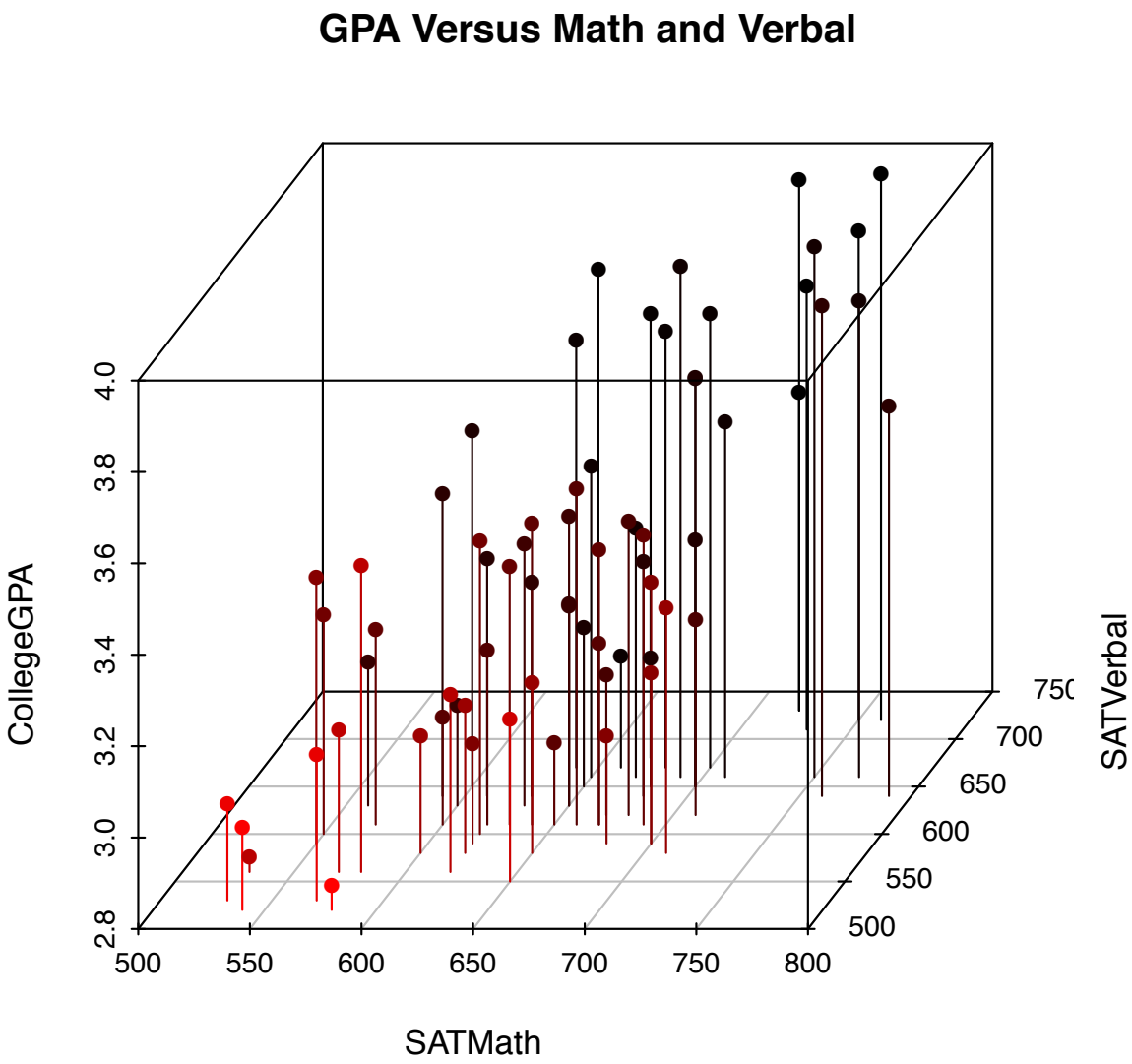

Figure 6.2: A three-dimensional point cloud showing the joint association between college GPA, SAT Math score, and SAT Verbal score.

in $p + 1$-dimensional Euclidean space—impossible to visualize beyond $p = 2$, but straightforward to describe mathematically.

From simple to multiple regression: what stays the same

In this jump from the familiar (straight lines in two dimensions) to the foreign (hyperplanes in arbitrary dimensions), it helps to start out by cataloguing the features that remain the same.

First, we can still fit parameters of the model using the principle of least squares. As before, we will denote our estimates by $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, and so on. For a given configuration of choices for these values, and a given point in predictor space, the fitted value of y is

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \widehat{\beta}_2 x_{i,2} + \cdots + \widehat{\beta}_p x_{i,p} \,.$$

This is a one-dimensional quantity, even though the regression parameters describe a p-dimensional hyperplane. Therefore, we can define the residual sum of squares in the same way as before, as the sum of squared differences between fitted values and observed values:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left\{ y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \widehat{\beta}_2 x_{i,2} + \cdots + \widehat{\beta}_p x_{i,p}) \right\}^2 .$$

The principle of least squares prescribes that we should choose the estimates so as to make the residual sum of squares as small as possible, thereby distributing the "misses" among the observations in a roughly equal fashion. Just as before, the little e_i is the amount by which the fitted plane misses the actual observation y_i. These residuals still have the same interpretation as before: as the part of y that is unexplained by the predictors. Indeed, for a least-squares fit, the residuals will be pairwise-uncorrelated with each of the original predictors.

Second, we still summarize preciseness of fit using R^2, which has the same definition as before:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{UV}{TV} = \frac{PV}{TV} \,.$$

The only difference is that $\hat{y}_i$ is now a function of more than just an intercept and a single slope. Also, just as before, it will still be the case R^2 is the square of the correlation coefficient between y_i and $\hat{y}_i$. It will not, however, be expressible as the correlation between y and any of the original predictors, since we now have more than one predictor to account for. (Indeed, R^2 is the natural

generalization of Pearson's r for measuring correlation between one response and a whole basket of predictors.)

Third, we will still make extensive use of the assumption of normally distributed residuals. This is the so-called *multiple regression model*, where

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \,. \end{aligned}$$

An equivalent way of writing this is:

$$(y_i \mid x_i, \beta_0, \beta_1, \sigma^2) \overset{iid}{\sim} \mathrm{N}\Big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j},\ \sigma^2\Big) \,.$$

This is just the linearity assumption, extended to more than one predictor. The same auxiliary assumptions are needed, as well:

(1) *Independence of the residuals:* no residual provides any information about another residual.

(2) *Normality of the residuals:* the residuals ϵ_i come from a normal distribution with mean 0 and variance σ^2.

(3) *Homoskedasticity:* σ^2 is the same for all observations.

These should all look familiar, and so should the rationale for introducing them—namely, to allow us to quantify our uncertainty about parameters, predictions, and the linear model itself.

Fourth, it remains important to respect the distinction between the true model parameters (σ, β_0, β_1, and so forth) and the estimated parameters ($\hat{\sigma}$, $\widehat{\beta}_0$, $\widehat{\beta}_1$ and so forth). When using the multiple regression model, we imagine that there is some true hyperplane described by β_0 through β_p, and some true residual variance σ^2, that gave rise to our data. We can infer what those parameters are likely to be on the basis of observed data, but we can never know their values exactly.

From simple to multiple regression: what changes

Not everything about our inferential process stays the same, of course. We will focus more on some of the differences later, but for now, we'll mention two major ones.

First of all, the interpretation of each β coefficient is no longer quite so simple. The best way to think of $\widehat{\beta}_j$ is as an estimated

partial slope: that is, the change in y associated with a one-unit change in x_j, holding all other variables constant. In other words, it represents the linear change in y that we can predict using x_j, after adjusting for all the other changes in y that can be predicted in terms of changes in predictor variables. One very important fact worth mentioning is the following: the magnitude of this change (that is, β_j) does not depend upon the particular values at which the other predictor variables are fixed. The effects of different predictors are, in other words, completely separable from one another.

It is important to keep in mind that this adjustment is statistical in nature, rather than experimental, since the whole system is passively observed. We do not, and typically cannot, actively manipulate the values of the other predictors to see how these changes affect y. Still, this is often the best we can do when investigating certain questions that, for whatever reason, just aren't amenable to experimentation.

Second, although we will still have estimated coefficients $\widehat{\beta}_j$ and estimated standard errors $\hat{\sigma}_j$, we no longer have simple formulas for these quantities. So while it remains valid to quote a confidence interval for β_j as $\widehat{\beta}_j \pm t^\star \hat{\sigma}_j$, the quantities themselves must typically be calculated using computer software. Practically speaking, of course, this is no different than the one-predictor case, where most of us would fit the line using software, anyway. It's just that here, we lack even a simple expression for $\widehat{\beta}_j$ and $\hat{\sigma}_j$ that doesn't involve $n \times p$ different terms—that is, every single one of the x_j's for every single one of the observations.

Finally, we estimate the residual variance σ^2 in a slightly different way:

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

where p is the number of predictor variables. The intuition here is that, since we must estimate more parameters compared to the one-variable case, we use up additional degrees of freedom in the data.[1]

[1] Note that if $p = 1$, we recover the original formula from the previous chapter.