# 6 Multiple Regression

## 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:

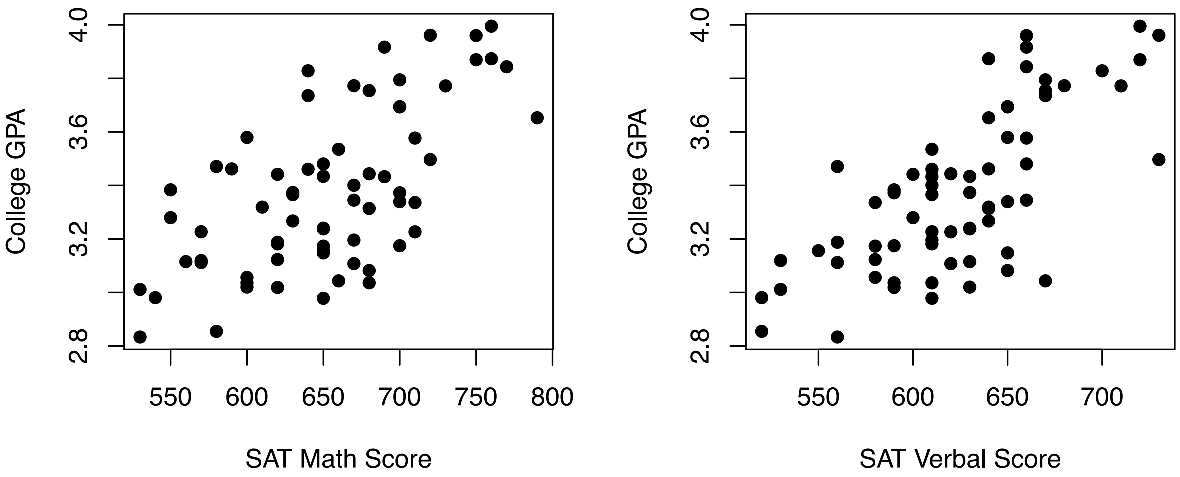

Figure 6.1: Data from a sample of 68 college students. Left: college GPA versus SAT Math score. Right: those same values of college GPA versus SAT Verbal score.

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 $\beta_0$, $\beta_1$, and $\beta_2$ 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 $(\beta_0, \beta_1, \ldots, \beta_p)$.

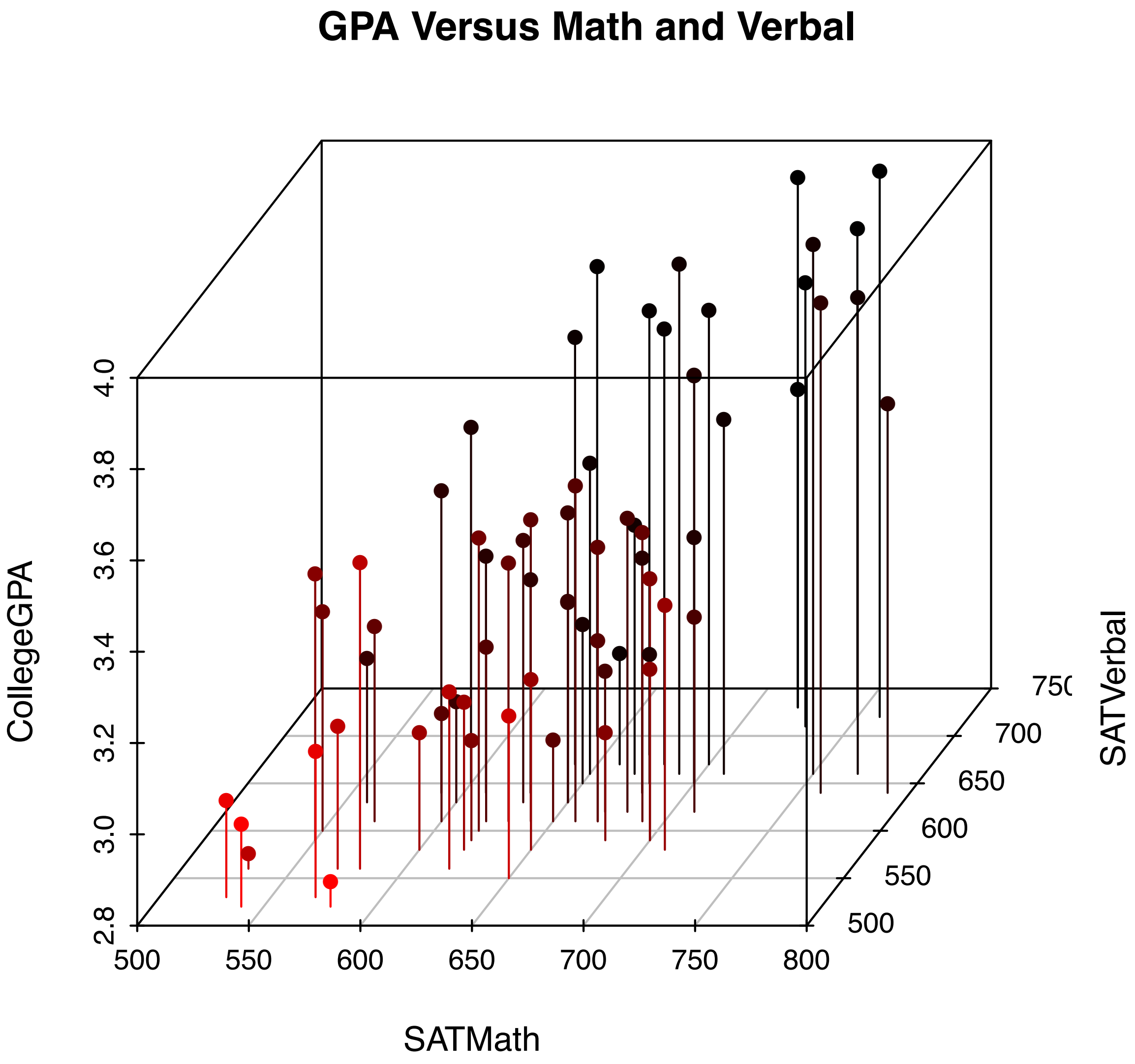

Figure 6.2: A three-dimensional point cloud showing the joint association between college GPA, SAT Match score, and SAT Verbal score.

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 $\epsilon_i$ come from a normal distribution with mean 0 and variance $\sigma^2$.

(3) *Homoskedasticity:* $\sigma^2$ 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 ($\sigma$, $\beta_0$, $\beta_1$, 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 $\beta_0$ through $\beta_p$, and some true residual variance $\sigma^2$, 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 $\beta$ 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, $\beta_j$) 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 $\beta_j$ 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 $\sigma^2$ 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.

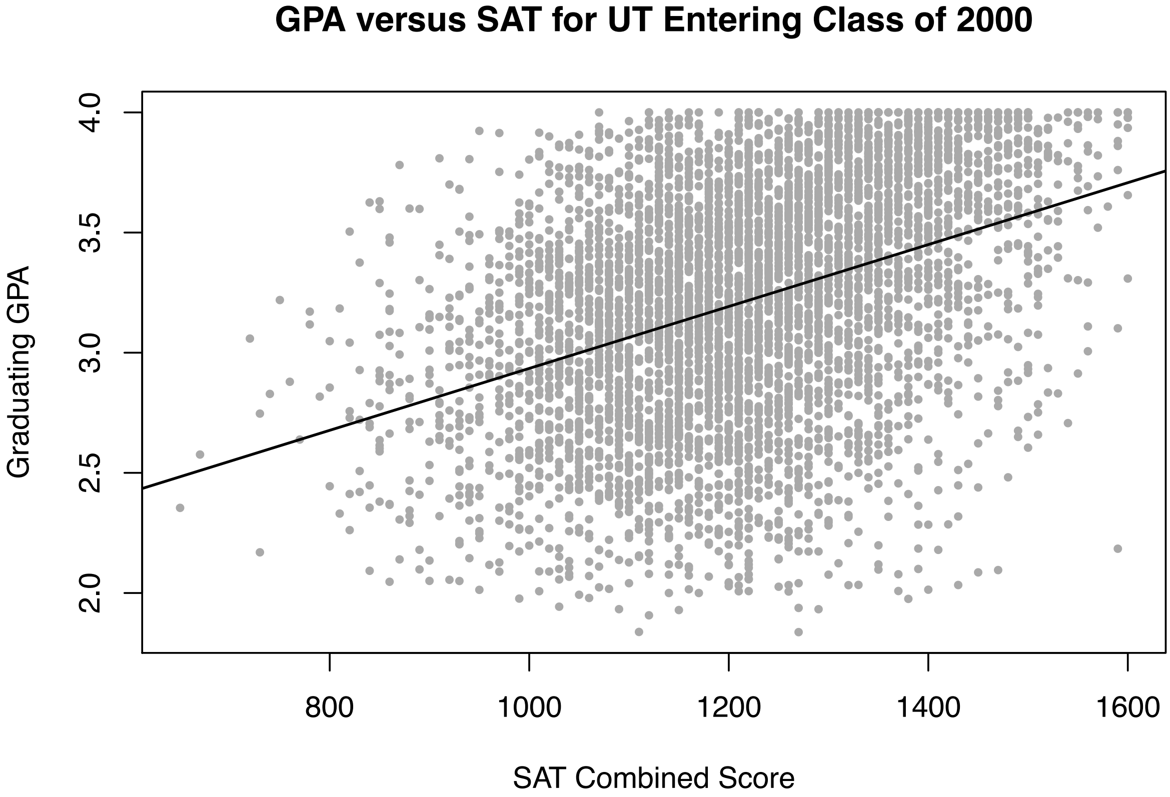

Figure 6.3: Combined SAT scores versus graduating GPA for the entering fall class of 2000 at the University of Texas.

### Qualitative variables as predictors

THE ONLY kind of regression models we've considered so far have been those where all predictors can be expressed on a continuous quantitative scale. But sometimes qualitative predictors can be useful as well. For example, let's take a look a more extensive data set on college GPA versus high-school SAT scores. This one catalogues all 5,191 students at the University of Texas who matriculated in the fall semester of 2000, and who went on to graduate within five years. (Hence those who dropped out or took longer to graduate are not part of the sample.) We notice the familiar positive relationship between combined SAT score and final GPA in the plot above.

So far so good; $R^2$ for this model is only about 15%, but the slope term has a $t$ statistic over 30, so the association looks rock-solid. We could now proceed just as we did before, breaking the SAT score down into its component parts to help us understand the three-variable joint distribution—GPA, SAT Math, and SAT Verbal—a bit more deeply.

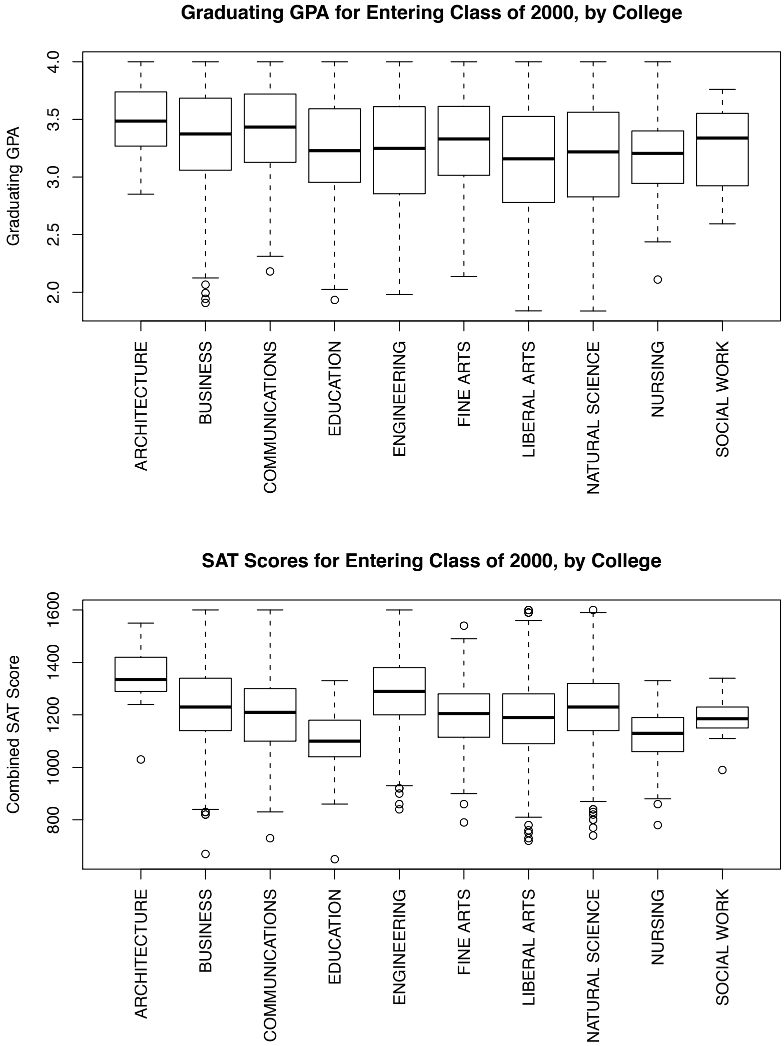

Figure 6.4: GPA and SAT scores stratified by the ten undergraduate colleges at UT.

However, we might notice the fact that SAT scores and graduating GPA's tend to differ substantially from one college to the next. Figure 6.4 shows boxplots of SAT and GPA stratified by the ten undergraduate colleges at the University of Texas. Some major differences are apparent.

It's possible, of course, that all the observed differences in the college-by-college GPA's are explained by corresponding differences in the college-by-college SAT scores. But maybe not. How can we investigate this question?

One possibility is to run a separate regression for each college. But this seems wasteful of information, since we expect the GPA/SAT relationship to at least be qualitatively similar for all ten colleges. An alternative approach is to use the college itself as a predictor variable in the regression.

Since a student's college is a categorical variable rather than a quantitative one, we have to proceed a little bit differently. Let's say that $x_{i,1}$ denotes SAT Math score, and $x_{i,2}$ denotes SAT Verbal score, for student $i$. To use a student's college of enrollment as a predictor, we introduce so-called "dummy variables" that take either the value 0 or 1. For example, to create a dummy variable corresponding to the business school, we create a data column $x_{i,3}$ that is equal to 1 if student $i$ is a business major, and 0 otherwise.

To encode the entire categorical variable, we must create a separate dummy variable for each category. Since the college can be one of ten choices, this means we need nice total dummy variables (i.e. ten new columns in a spreadsheet), in this case numbered $x_{i,3}$ through $x_{i,11}$. Why nine variables for ten categories? Because the first category, in this case Architecture, just gets subsumed into the global intercept term. Notice that only one of these dummy variables will be 1 for each person, since a person is only in one college. (In this data set, a dual-degree recipient is encoded according to the college he or she enrolled in as a freshman.)

The model now has eleven predictors, of which nine can only be 0 or 1. Here's the regression output:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.691e+00  9.624e-02  17.566   <2e-16 ***
SAT.V                  1.486e-03  8.515e-05  17.455   <2e-16 ***
SAT.Q                  1.186e-03  9.098e-05  13.041   <2e-16 ***
SchoolBUSINESS         5.784e-03  7.827e-02   0.074   0.9411
SchoolCOMMUNICATIONS   8.565e-02  8.088e-02   1.059   0.2896
SchoolEDUCATION        4.492e-02  8.552e-02   0.525   0.5994
SchoolENGINEERING     -1.890e-01  7.851e-02  -2.408   0.0161 *
SchoolFINE ARTS        8.423e-03  8.443e-02   0.100   0.9205
SchoolLIBERAL ARTS    -1.374e-01  7.763e-02  -1.770   0.0767 .
```

```
SchoolNATURAL SCIENCE -1.495e-01  7.789e-02  -1.920   0.0549 .
SchoolNURSING          2.423e-02  1.022e-01   0.237   0.8126
SchoolSOCIAL WORK     -3.787e-02  1.391e-01  -0.272   0.7854
```

There is no dummy variable associated with Architecture; think of this as the baseline case, against which the other colleges are compared. The regression coefficients associated with the "School" dummy variables then shift the line systematically up or down relative to the global intercept, but they do not change the slope of the line. Intuitively, we are fitting a model where all colleges share a common slope, but have unique intercepts. This is something of a compromise solution between fitting a single model (as above) and fitting ten distinct models for the ten individual colleges.

### *Interactions*

What if we expect that a categorical variable will result not merely in a change of intercept, but also a change of slope associated with some other continuous predictor? For example, we might expect that, for students in Liberal Arts, GPA's will vary more sharply with SAT Verbal scores, and less sharply with Math scores, than for students in Engineering.

If this is the case, then we should think about including an *interaction term* in the model. Simple interactions are new predictors formed by multiplying a quantitative predictor and a dummy (0–1) variable. When the dummy variable is 0, the interaction term disappears. But when the dummy is 1, the interaction is equal to the original quantitative predictor, whose effective partial slope then changes.

Let's take a simple example involving baseball salaries, plotted above. On the $y$-axis are the log salaries of 142 baseball players. On the $x$-axis are their corresponding batting averages. The kind of mark indicates whether the player is in the Major League, AAA (the highest minor league), or AA (the next-highest minor league). The straight lines reflect the least-squares fit of a model that regresses log salary upon batting average and a couple of dummy variables corresponding to a player's league. The three lines are parallel, since the dummy variable allows only the intercept to change as a function of league.

If we want the slope to change as well, then we must fit a model

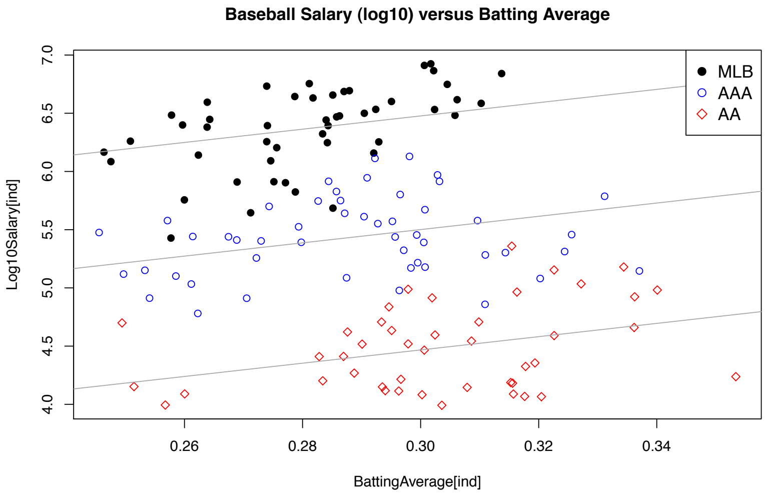

Figure 6.5: Baseball salaries versus batting average for Major League, AAA, and AA players.

like this:

$$E(y_i \mid \mathbf{x}_i) = \beta_0 + \beta_1 \cdot AVG + \underbrace{\beta_2 \cdot 1_{AAA} + \beta_3 \cdot 1_{AA}}_{\text{Dummy variables}} + \underbrace{\beta_4 \cdot AVG \cdot 1_{AAA} + \beta_5 \cdot AVG \cdot 1_{MLB}}_{\text{Interaction terms}}$$

The $y$ variable depends on $\beta_0$ and $\beta_1$ for all players, regardless of league. But when a player is in AAA, the corresponding dummy variable ($1_{AAA}$) kicks in. Before, only an extra intercept term was activated, shifting the entire line up (as in Figure 6.5). Now, an extra intercept $\beta_2$ *and* an extra slope $\beta_4$ are activated. Ditto for players in the Major League: then the MLB dummy variable ($1_{MLB}$) kicks in, and both $\beta_3$ (an extra intercept) and $\beta_5$ (an extra slope) are activated. Fitting such model produces a picture like the one above (Figure 6.6).

Without any interaction terms, the fitted model is:

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.75795    0.41893   6.583 8.88e-10 ***
BattingAverage  5.69745    1.37000   4.159 5.59e-05 ***
ClassAAA        1.03370    0.07166  14.426  < 2e-16 ***
ClassMLB        2.00990    0.07603  26.436  < 2e-16 ***
---
Residual standard error: 0.3324 on 138 degrees of freedom
Multiple R-squared: 0.845,Adjusted R-squared: 0.8416
```

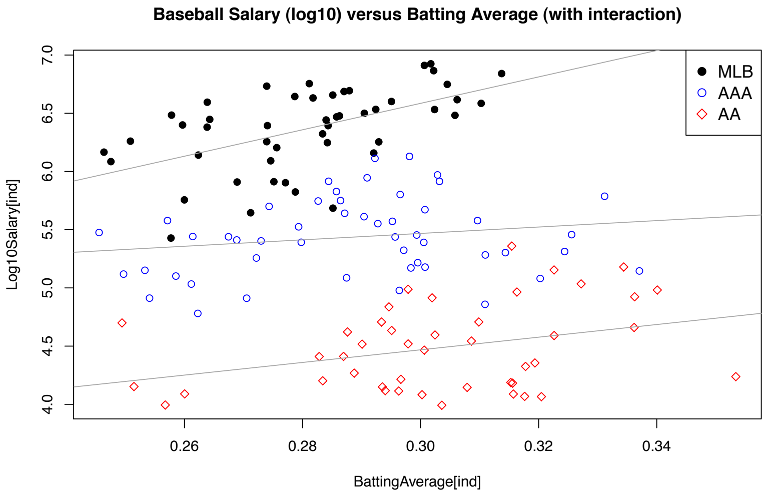

Figure 6.6: Baseball salaries versus batting average for Major League, AAA, and AA players. The fitted lines show the model with an interaction term between batting average and league.

With the interaction terms, we get:

```
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.8392     0.6718   4.227 4.33e-05 ***
BattingAverage             5.4297     2.2067   2.461   0.0151 *
ClassAAA                   1.8024     0.9135   1.973   0.0505 .
ClassMLB                   0.3393     1.0450   0.325   0.7459
BattingAverage:ClassAAA  -2.6758     3.0724  -0.871   0.3853
BattingAverage:ClassMLB   5.9258     3.6005   1.646   0.1021
---
Residual standard error: 0.3278 on 136 degrees of freedom
Multiple R-squared: 0.8514,Adjusted R-squared: 0.846
```

According to these estimates, salaries increase with average fastest in the Major Leagues, and slowest in AAA. Neither of these interaction terms, however, can be estimated very precisely, and the bump in $R^2$ looks pretty small compared to the smaller model. Our reduced model, without the interaction terms, has 3 predictors with $R^2 = 0.8450$.

One question worth asking is: what's the difference between the interaction model, and the process of simply fitting three different regression models to the three cohorts? Here, the only difference is that the interaction model involves one residual variance term, compared to the three we'd have to estimate if we fit three different models.

In more complicated scenarios, however, we might find ourselves with two sets of dummy variables representing two logically different kinds of category. For example, we might introduce another set of dummy variables for a player's position on the field. In that kind of scenario, we might want to fit interactions for only one of the two categories—reasoning, for example, that the premium on good offensive numbers is less for pitchers and catchers than for other positions. In that case, we'd have one set of interactions in, and the other out, which would be a much smaller model than fitting a separate regression for every combination of categories.