5
Grouping Variables in Regression

Qualitative variables as predictors

So far, we've considered two kinds of regression models:

1. A numerical response with a categorical or grouping predictor. Here, we just computed the group means, and visualized the data using boxplots or dot plots.

2. A numerical response with a numerical predictor. Here, we fit least-squares regression lines (perhaps on a transformed scale), and visualized the data using scatter plots.

Now, we consider the case where we both two predictors: one grouping variable, and one numerical variable. For example, let's look at a 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 expected positive relationship between combined SAT score and final GPA in Figure 5.2.

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. However, we might notice the fact that SAT scores and graduating GPA's tend to differ substantially from one college to the next. Figure 5.1 shows boxplots of SAT and GPA stratified by the ten undergraduate colleges at the University of Texas.

What we see in Figures 5.2 and 5.1 is often called an *aggregation paradox*, where the same trend that holds for individuals does not hold for groupings of individuals. Why is this a paradox? Look carefully at the data: figure 5.1 says that students with higher SAT scores tend to have higher GPAs. Yet this trend does not hold at the college level, even broadly. For example, Engineering students (as a group) have among the highest average SAT scores,

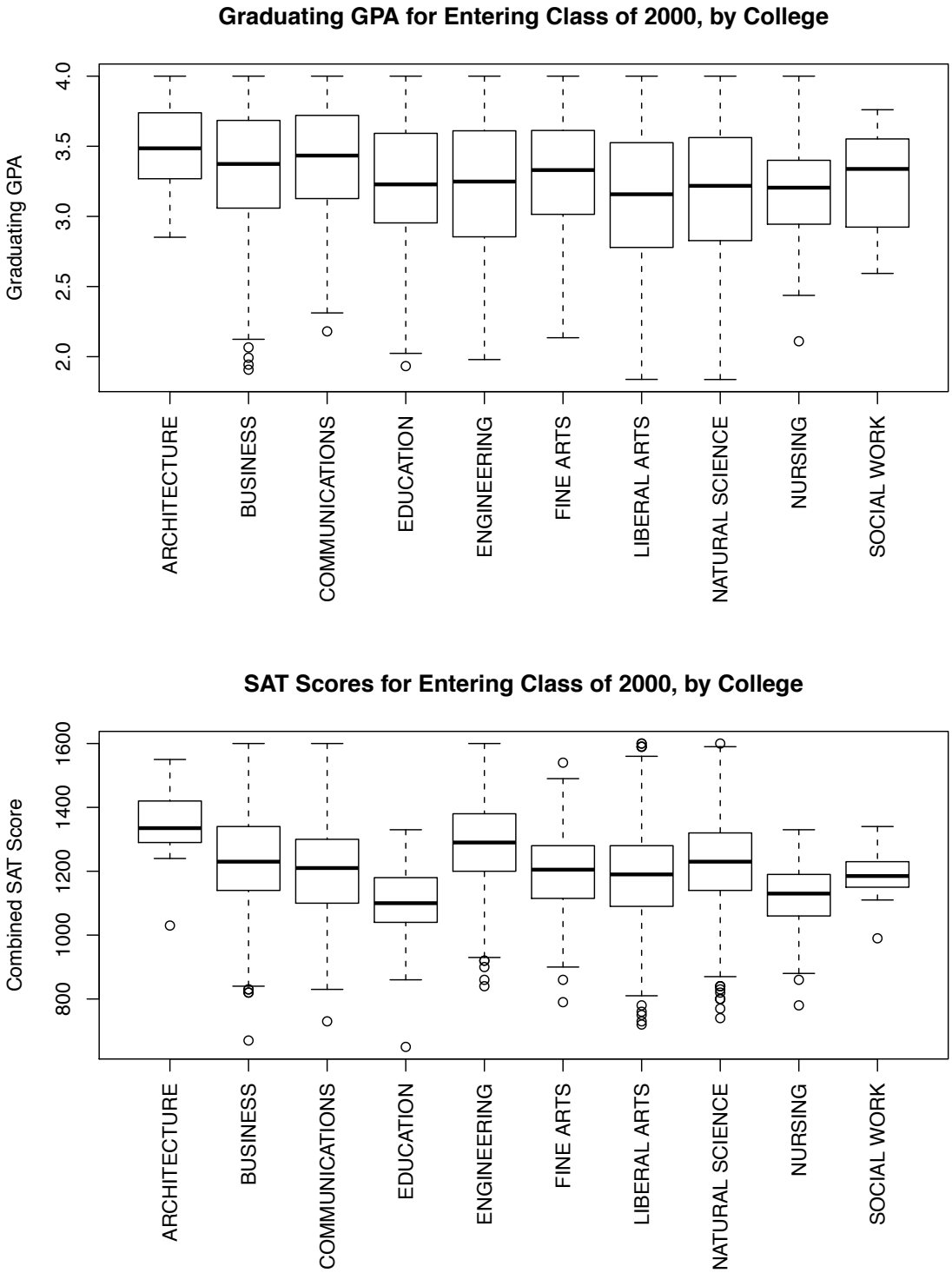

Figure 5.1: GPA and SAT scores stratified by the ten undergraduate colleges at UT.

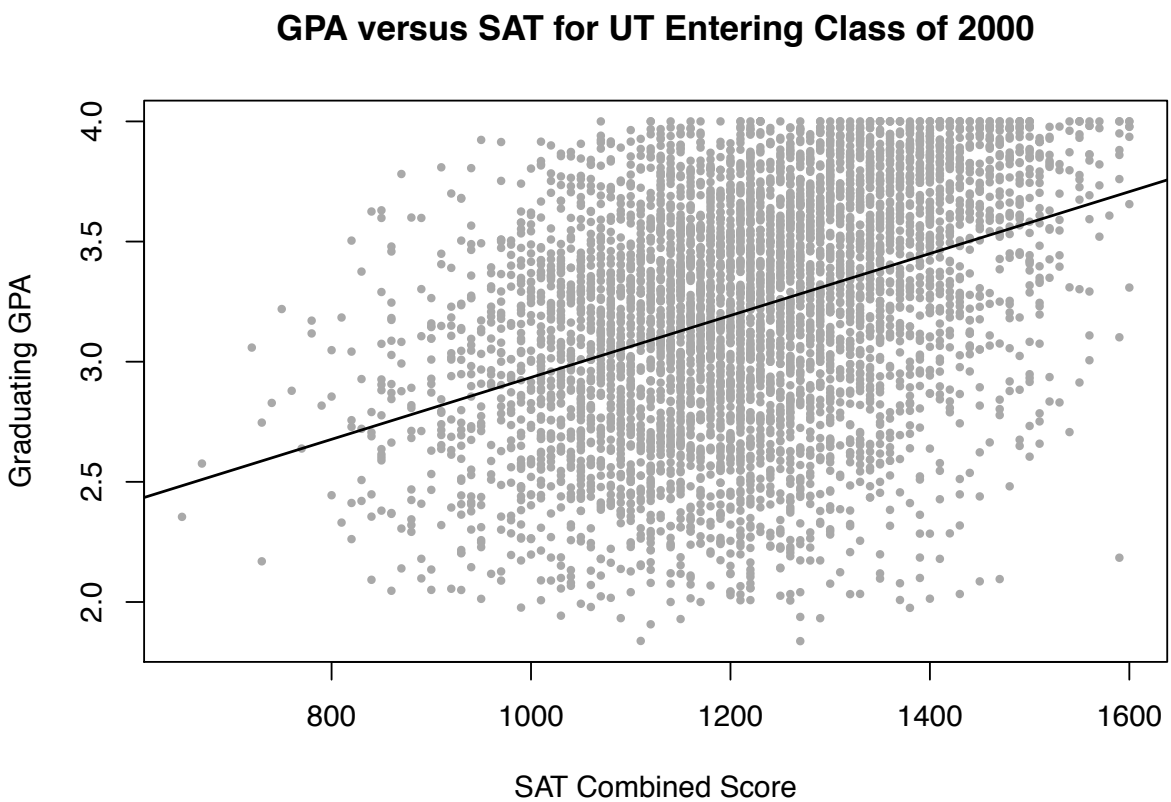

Figure 5.2: Combined SAT scores versus graduating GPA for the entering fall class of 2000 at the University of Texas.

and among the lowest average GPAs. Thus we have a paradox: it looks as though high SAT scores predict high GPAs, but being in a college with high SAT scores does not predict being in a college with high GPAs.

The paradox disappears when we realize the the "College" variable is a confounder for the relationship between SAT score and GPA. That is, a student's college is systematically associated with both SAT and GPA: some degrees are harder than others, and have higher entrance requirements. The right way to proceed here is to disaggregate the data and fit a different regression line within each of the ten colleges. There are two different ways to do this:

1. We could fit ten different lines, each with a different intercept ($\beta_0^{(k)}$), but all with the same slope (β_1). This would make sense if we thought that the same SAT–GPA relationship

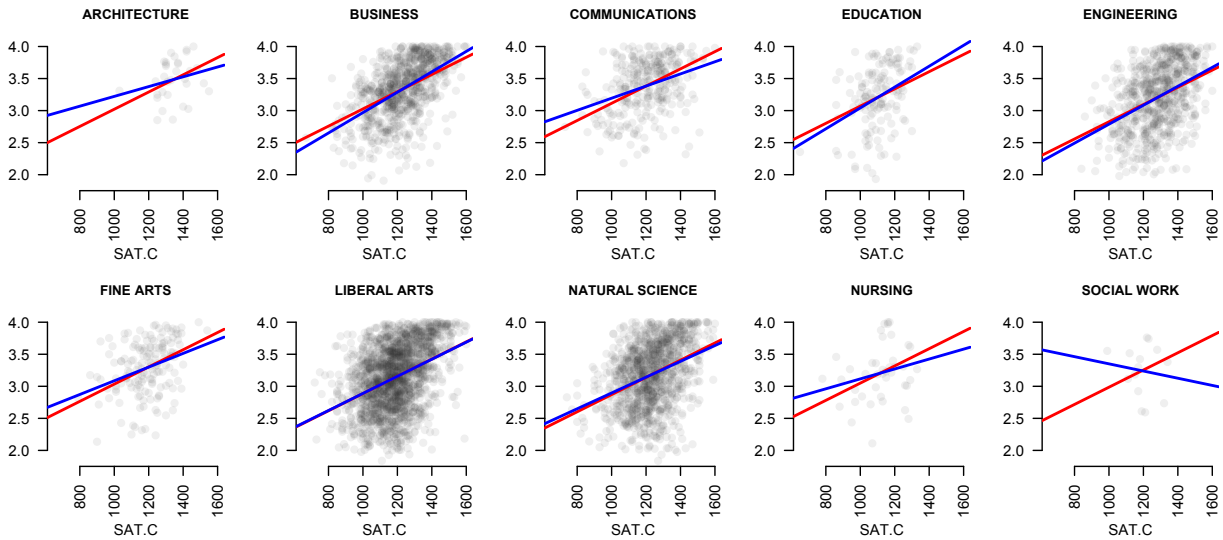

Figure 5.3: Separate regression models fit for GPA versus SAT within each college. The red lines all have the same slope, but a different intercept for each college. The blue lines all have different intercepts and different slopes.

ought to hold within each college, but that each college had a systematically higher or lower intercept (average GPA). These are the red lines in Figure 5.3.

2. We could fit ten different lines, allowing both the slope and the intercept to differ for each college. We would do this if we thought that the SAT–GPA relationship differed fundamentally across the colleges. These are the blue lines in Figure 5.3.

The question of how we make these choices requires a bit of notational background on dummy variables and interaction terms.

Dummy variables. Let's return to a basic group-wise model, where we have a categorical predictor and a numerical response. We often express the estimates of the group means in terms of *indicator* or *dummy* variables. To understand these terms, take the simple case of a single grouping variable x with two levels: "on" ($x = 1$) and "off" ($x = 0$). We can write this model in "baseline/offset" form:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_i=1\}} + e_i \,.$$

The quantity $\mathbf{1}_{\{x_i=1\}}$ is called a dummy variable; it takes the value 1 when $x_i = 1$, and the value 0 otherwise. Just as in an ordinary linear model, we call β_0 and β_1 the *coefficients* of the model. This

way of expressing the model implies the following.

$$\begin{aligned} \text{Group mean for case where } x \text{ is off} &= \beta_0 \\ \text{Group mean for case where } x \text{ is on} &= \beta_0 + \beta_1 \,. \end{aligned}$$

Therefore, we can think of β_0 as the baseline (or *intercept*), and β_1 as the offset.

We estimate the values of β_0 and β_1 using the least-squares criterion: that is, make the sum of squared errors, $\sum_{i=1}^{n} e_i^2$, as small as possible. It turns out that this is mathematically equivalent to computing the group-wise means separately. In light of this, you might wonder: why bother with the baseline/offset form? One reason is simple: we are often interested not in the means themselves, but in the *differences* between the means (in this case, the offset β_1).

More than two levels. If the predictor x has more than two levels, we must expand it in terms of more than one dummy variable. Suppose that x can take four levels, labeled arbitrarily as 0 through 3. Then our model is

$$y_i = \beta_0 + \beta_1^{(1)} \mathbf{1}_{\{x_i=1\}} + \beta_1^{(2)} \mathbf{1}_{\{x_i=2\}} + \beta_1^{(3)} \mathbf{1}_{\{x_i=3\}} + e_i \,.$$

More generally, $\beta_j^{(k)}$ is the coefficient associated with the kth level of the jth variable. Notice that there is no dummy variable for the case $x = 0$: this is the baseline case, whose group mean is described by the intercept β_0. In general, for a categorical variable with K levels, we will have $K - 1$ dummy variables. This is why we sometimes call it "baseline/offset" form: β_0 is the baseline mean, and the other coefficients are the differences between the baseline and the corresponding group:

$$\begin{aligned} \text{Group mean for case where } (x_i = 0) &= \beta_0 \\ \text{Group mean for case where } (x_i = k) &= \beta_0 + \beta_1^{(k)} \,. \end{aligned}$$

Adding a continuous predictor. We're now ready to add a continuous predictor into the mix. Suppose we now have two predictors for each observation: a grouping variable $x_{i,1}$ that can take levels 0 to K, and a numerical predictor $x_{i,2}$. We now start with the regression equation involving a set of K dummy variables, and add the effect of the continuous predictor onto the right-hand side of the regression equation:

$$y_i = \beta_0 + \beta_1^{(1)} \mathbf{1}_{\{x_{i1}=1\}} + \beta_1^{(2)} \mathbf{1}_{\{x_{i1}=2\}} + \cdots + \beta_1^{(K)} \mathbf{1}_{\{x_{i1}=K\}} + \beta_2 x_{i2} + e_i \,.$$

Now we have the following set of regression equations.

$$\begin{aligned}
\text{Regression equation for case where } (x_i = 0): \quad y_i &= \beta_0 + \beta_2 x_{i2} + e_i \\
\text{Regression equation for case where } (x_i = k): \quad y_i &= (\beta_0 + \beta_1^{(k)}) + \beta_2 x_{i2} + e_i \,.
\end{aligned}$$

In words, we have $K + 1$ different regression equations for the $K + 1$ different groups. Each lines has a different intercept, but they all have the same slope. These correspond to the red lines in Figure 5.3.

The first category, in this case Architecture, just gets subsumed into the global intercept term. The coefficients $\beta_1^{(k)}$ are associated with the dummy variables. Notice that only one of these dummy variables will be 1 for each person, and the rest will be zero, since a person is only in one college. Here's the regresison output when we ask for a model of GPA $\sim$ SAT.C + School:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.678365   0.096062  17.472   <2e-16 ***
SAT.C                 0.001343   0.000043  31.235   <2e-16 ***
SchoolBUSINESS        0.004676   0.078285   0.060   0.9524
SchoolCOMMUNICATIONS  0.092682   0.080817   1.147   0.2515
SchoolEDUCATION       0.048688   0.085520   0.569   0.5692
SchoolENGINEERING    -0.195433   0.078460  -2.491   0.0128 *
SchoolFINE ARTS       0.012366   0.084427   0.146   0.8836
SchoolLIBERAL ARTS   -0.134092   0.077629  -1.727   0.0842 .
SchoolNATURAL SCIENCE -0.150631  0.077908  -1.933   0.0532 .
SchoolNURSING         0.028273   0.102243   0.277   0.7822
SchoolSOCIAL WORK    -0.035320   0.139128  -0.254   0.7996
```

There is no dummy variable associated with Architecture, because it is 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. As the math above shows, 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.

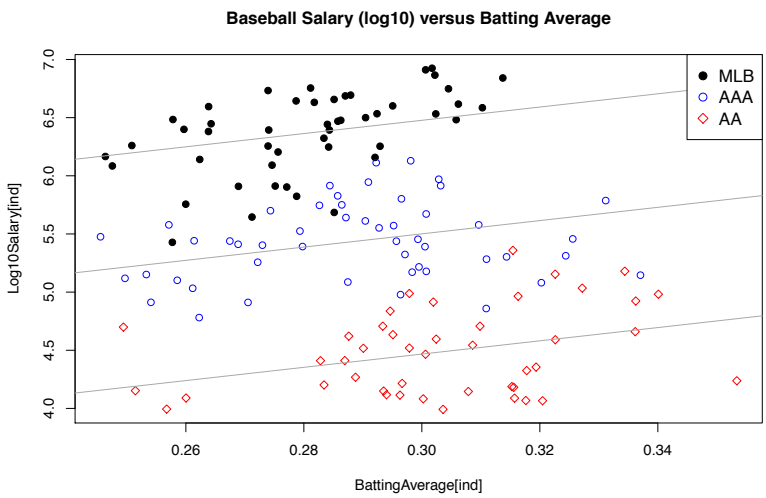

Figure 5.4: Baseball salaries versus batting average for Major League, AAA, and AA players.

Interactions

But 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 like this:

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

The y variable depends on β_0 and β_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 5.4). Now, an extra intercept β_2 *and* an extra slope β_4 are activated. Ditto for players in the Major League: then the MLB dummy variable (1_{MLB}) kicks in, and both β_3 (an extra intercept) and β_5 (an extra slope) are activated. Fitting such model produces a picture like the one above (Figure 5.5).

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
```

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

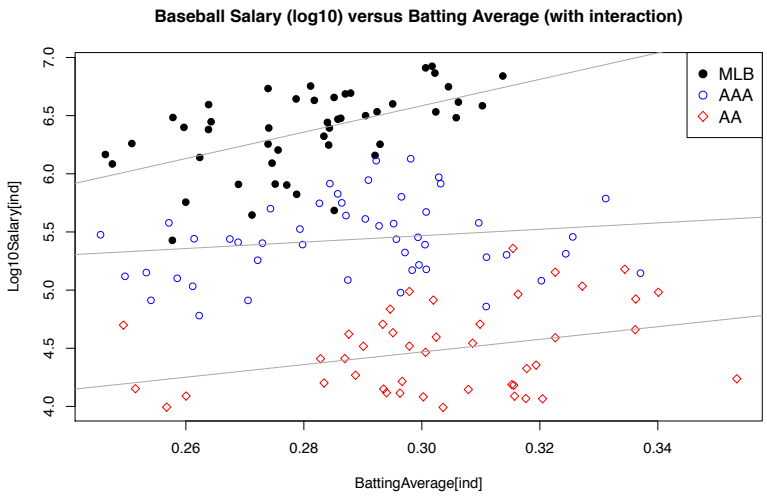

Figure 5.5: Baseball salaries versus batting average for Major League, AAA, and AA players. The fitted lines show the model with an interaction term between batting average and league.

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.