# Grouping variables in regression

## Grouping variables and aggregation paradoxes

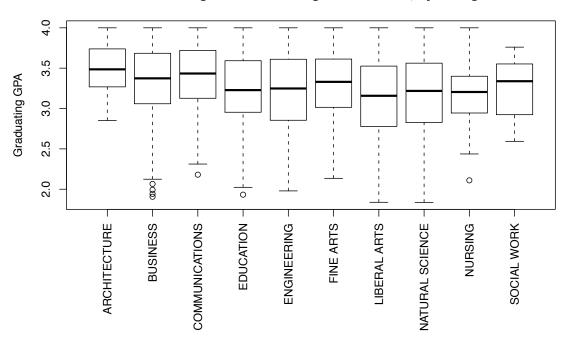
THE previous chapters have taught us to fit equations to data involving a numerical response and a numerical predictor. In this chapter, we'll generalize these ideas to incorporate grouping variables as predictors, too.

It's very common in real-world systems for one variable to modulate the effect of another. For example, a person's overall size and weight modulate the relationship between alcohol and cognitive impairment. A single glass of wine might make a small person feel drunk, but have a negligible effect on a big person.

This phenomenon is easiest to visualize in data when the variable that does the modulating is categorical. To see an example of this, we'll revisit the data set on college GPA versus high-school SAT scores. You'll recall that this data set 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. In Figure 4.2, we notice the expected positive relationship between combined SAT score and final GPA. We also notice the fact that SAT scores and graduating GPAs tend to differ substantially from one college to the next. Figure 4.1 shows boxplots of SAT and GPA stratified by the ten undergraduate colleges at UT.

What we see in Figures 4.2 and 4.1 is an example of 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 4.1 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, 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

# Graduating GPA for Entering Class of 2000, by College



# SAT Scores for Entering Class of 2000, by College

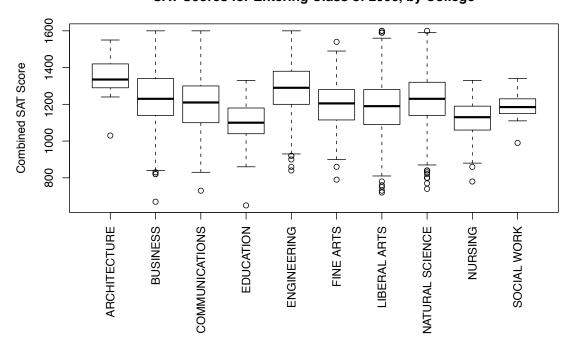


Figure 4.1: GPA and SAT scores stratified by the ten undergraduate colleges at UT.



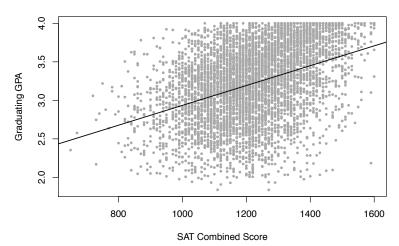


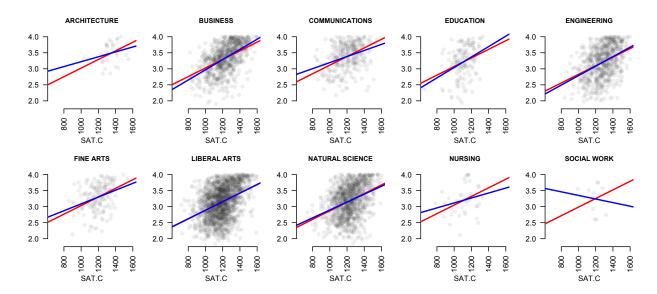
Figure 4.2: Combined SAT scores versus graduating GPA for the entering fall class of 2000 at the University of Texas.

college with high GPAs.

The paradox disappears when we realize the the "College" variable modulates the relationship between SAT score and GPA. A student's college is systematically associated with both SAT and GPA: some degrees are harder than others, and these degrees tend to enroll students with higher test scores.

The right way to proceed here is to disaggregate the data and fit a different regression line within each of the ten colleges, to account for the effect of the modulating variable. 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  $(\beta_1)$ . This would make sense if we thought that the same SAT–GPA relationship 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 4.3. You can see the differences among the red lines if you look carefully at where they hit the y axis in relation to a GPA of 2.5—for example, compare Communications and Engineering.
- 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 fun-



damentally across the colleges. These are the blue lines in Figure 4.3.

But which strategy should we take?

Things get even more complex in the presence of more than one grouping variable. For example, we might want to look at these relationships separately for different years, for men versus women, and for in-state and out-of-state students. To be able to model the effect of all these variables on GPA simultaneously, we will need to introduce some new notation and a few new concepts.

## Models for a single grouping variable

#### Dummy variables

LET's return to a simple scenario where we have numerical data that falls into two groups, and we want to compare the variation between the groups. The dotplot in Figure 4.4 shows the weekly sales volume of package sliced cheese over 61 weeks at a Dallasarea Kroger's grocery store. In 38 of these weeks, the store set up a prominent display near the entrance, calling shoppers' attention to the various culinary adventures they might undertake with the cheese. The data show that, in these 38 weeks, sales were higher overall than when no display was present.

Figure 4.3: 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.

How much higher? The average sales volume in display weeks was 5,577 units (the blue dotted line in Figure 4.4), versus an average of 2341 units in non-display weeks (the red dotted line). Thus sales were 3236 units higher in the display weeks. This difference is depicted in Figure 4.4 as the difference or offset between the dotted lines.

This example emphasizes that in many data sets, we care less about the absolute magnitude of a response under different conditions, and more about the differences between those conditions. We therefore often build our model in such a way that these differences are estimated directly, rather than indirectly (i.e. by calculating means and then subtracting them).

We do this using *indicator* or *dummy* variables. To understand this idea, 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  $\beta_0$  and  $\beta_1$  the *coefficients* of the model. This way of expressing the model implies the following.

Group mean for case where x is off  $= \beta_0$ Group mean for case where x is on  $= \beta_0 + \beta_1$ .

Therefore, we can think of  $\beta_0$  as the baseline (or *intercept*), and  $\beta_1$  as the offset. To see this in action, consult Figure 4.4 again. Here the dummy variable encodes the presence of an in-store display. The red dot at 2341, in the non-display weeks, is  $\beta_0$ . This is the baseline case, when the dummy variable x is "off." The coefficient for the dummy variable,  $\beta_1 = 3236$ , is the vertical distance between the two means. Thus if we wanted to reconstruct the mean for the with-display weeks, we would just add the baseline and the offset, to arrive at 2341 + 3236 = 5577, where the blue dot sits.

As before, we estimate the values of  $\beta_0$  and  $\beta_1$  using the least-squares criterion: that is, make the sum of squared errors,  $\sum_{i=1}^{n}e_i^2$ , as small as possible. This is mathematically equivalent to computing the group-wise means separately, and then calculating the difference between the means.

#### Weekly sales of cheese at a Dallas-area Kroger

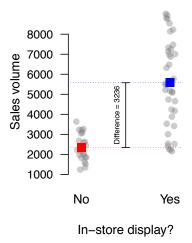
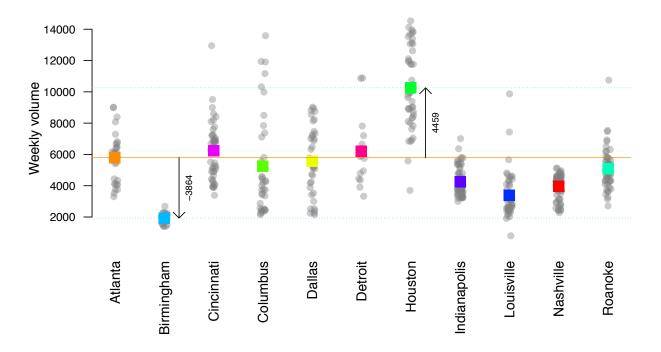


Figure 4.4: Weekly sales of packaged cheese slices at a Dallas-area Kroger's grocery store, both with and without the presence of an in-store display ad for the cheese. The red dot shows the mean of the no-display weeks, and the blue dot shows the mean of the with-display weeks. The estimated coefficient for the dummy variable that encodes the presence of a display ad is 3236, which is the vertical distance between the two dots.

## Weekly cheese sales at 11 Kroger's stores



#### More than two levels

If the categorical predictor x has more than two levels, we represent it in terms of more than one dummy variable. Suppose that x can take three levels, labeled arbitrarily as 0 through 2. Then our model is

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

The dummy variables  $\mathbf{1}_{\{x_i=1\}}$  and  $\mathbf{1}_{\{x_i=2\}}$  tell you which of the levels is active for the *i*th case in the data set.<sup>1</sup>

More generally, suppose we have a grouping variable with K levels. Then  $\beta_1^{(k)}$  is the coefficient associated with the kth level of the grouping variable, and we write the full model as a sum of K-1 dummy-variable effects, like this:

$$y_i = \beta_0 + \sum_{k=1}^{K-1} \beta_1^{(k)} \mathbf{1}_{\{x_i = k\}} + e_i$$
 (4.1)

Figure 4.5: Weekly sales of packaged cheese slices during weeks with an advertising display at 11 Kroger's grocery stores across the country.

<sup>&</sup>lt;sup>1</sup> Normal people count starting at 1. Therefore you might find it strange that we start counting levels of a categorical variable at 0. The rationale here is that this makes the notation for groupwise models a lot cleaner compared to starting at 1.

We call this a *group-wise model*. Notice that there is no dummy variable for the case x=0. This is the baseline level, whose group mean is the intercept  $\beta_0$ . In general, for a categorical variable with K levels, we will need K-1 dummy variables, and at most one of these K-1 dummy variables is ever active for a single observation. The coefficient on each dummy variable  $(\beta_1^{(k)})$  is the differences between the baseline and the mean of group k:

Group mean for case where 
$$(x_i = 0) = \beta_0$$
  
Group mean for case where  $(x_i = k) = \beta_0 + \beta_1^{(k)}$ .

In Figure 4.5, we see an example of a single categorical variable with more than two levels. The figure shows weekly cheese sales (during display-present weeks only) at 11 different Kroger stores in 11 different markets across the country. The grouping variable here is the market: Atlanta, Birmingham, Cincinnati, and so forth. If we fit a model like Equation 4.1 to the data in this figure, choosing Atlanta to be the baseline, we get the set of estimated coefficients in the second column ("Coefficient") of the table below:

Variable	Coefficient	Group mean
Intercept	5796	_
Birmingham	-3864	1932
Cincinnati	427	6223
Columbus	-543	5253
Dallas	-219	5577
Detroit	400	6196
Houston	4459	10255
Indianapolis	-1542	4254
Louisville	-2409	3387
Nashville	-1838	3958
Roanoke	-717	5079

Atlanta is the baseline, and so the intercept is the group mean for Atlanta: 5796 packages of crappy cheese. To get the group mean for an individual market, we add that market's offset to the baseline. For example, the mean weekly sales volume in Houston is 5796 + 4459 = 10255 units. Group mean = baseline + offset.

The figure also shows you two of the offsets as arrows, to give you a visual sense of what these numbers in the above table represent. The coefficient for Houston is  $\beta_6^{(1)} = 4459$ , because the group

mean for Houston (10255) is 4459 units higher than the baseline group mean for Atlanta (a positive offset). Similarly, the coefficient for Birmingham is  $\beta_1^{(1)} = -3864$ , because the group mean for Birmingham (1932) is 3864 units *lower* than the baseline group mean for Atlanta (a negative offset).

The choice of baseline. In the above analysis, we chose Atlanta as the baseline level of the grouping variable. This was arbitrary. We could have chosen any city as a baseline, measuring the other cities as offsets from there instead.

A natural question is: does the model change depending on what level of the grouping variable we choose to call the baseline? The answer is: yes and no. Yes, the estimated model coefficients will change when a different baseline is used; but no, the underlying group means do not change. To see this, consider what happens when we fit another model like Equation 4.1 to the Kroger cheese-sales data, now choosing the Dallas store to be the baseline:

Variable	Coefficient	Group mean
Intercept	5577	_
Atlanta	219	5796
Birmingham	-3644	1932
Cincinnati	646	6223
Columbus	-324	5253
Detroit	619	6196
Houston	4678	10255
Indianapolis	-1323	4254
Louisville	-2190	3387
Nashville	-1619	3958
Roanoke	-498	5079

The intercept is the Dallas group mean of 5577, and the other market-level coefficients have changed from the previous table, since these now represent offsets compared to a different baseline. But the group means themselves do not change. The moral of the story is that the coefficients in a model involving dummy variables do depend upon the choice of baseline, but that the information these coefficients encode—the means of the underlying groups does not. Different choices of the baseline just lead to different ways of expressing this information.

## Multiple grouping variables

We began our discussion of dummy variables by looking at a simple group-wise model with a binary predictor, meaning that  $x_i$  is either 0 or 1. Such a model takes the form

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

We learned something important about this model: that the coefficient  $\beta_1$  can be interpreted as the differential effect of being in group 1, as opposed to the baseline (group o).<sup>2</sup> That's a nice feature of using dummy variables: if we care primarily about the difference in the average response between conditions, we get an estimate of that difference ( $\hat{\beta}_1$ ) directly from the fitted model.

This approach of using dummy variables to encode the grouping structure of our data really comes into its own when we encounter data sets with more than one grouping variable. To see why, we'll spend some time with the data in Figure 4.6.

#### Main effects

Making a best-selling video game is hard. Not only do you need a lot of cash, a good story, and a deep roster of artists, but you also need to make the game fun to play. Take Mario Kart for the Super Nintendo, my favorite video game from childhood. In Mario Kart, you had to react quickly to dodge banana peels and Koopa shells launched by your opponents as you all raced virtual go-karts around a track. The game was calibrated just right. If the required reaction time had been just a little slower, the game would have been too easy, and therefore boring. But if the required reaction time had been a little bit faster, the game would have been too hard, and therefore also boring.

Human reaction time to visual stimuli is a big deal to video game makers. They spend a lot of time studying it and adjusting their games according to what they find. Figure 4.6 shows the results of one such study. Participants were presented with a natural scene on a computer monitor, and asked to react (by pressing a button) when they saw an animated figure appear in the scene.<sup>3</sup>

The experimenters varied the conditions of the natural scene: some were cluttered, while others were relatively open; in some, the figure appeared far away in the scene, while in others it appeared close up. They presented all combinations of these conditions to each participant many times over. The top two panels of

<sup>&</sup>lt;sup>2</sup> Remember, we start counting groups/levels at o.

<sup>&</sup>lt;sup>3</sup> Essentially the company was measuring how quickly people could react to a bad guy popping up on the screen in a video game.

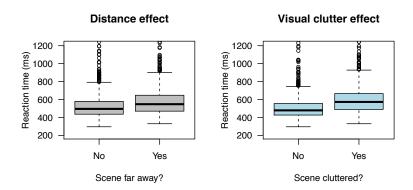


Figure 4.6: Reaction time to visual stimuli in a controlled experiment run by a major video-game maker. Top-left: participants reacted more slowly, on average, when the stimulus was far away within the scene. Top-right: participants reacted more slowly, on average, in a scene with significant visual clutter. Bottom: systematic differences in reaction time across participants in the trial.

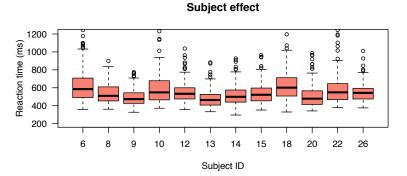


Figure 4.6 show boxplots of all participants' reaction times across all trials under these varying conditions. On average, participants reacted more slowly to scenes that were far away (top left panel) and that were cluttered (top right panel).

We'll return to the bottom panel of Figure 4.6 shortly. For now, let's focus on the "distance effect" and the "clutter effect" in the top two panels. This presents us with the case of two grouping variables,  $x_1$  and  $x_2$ , each of which affects the response variable, and each of which can take the value 0 ("off") or 1 ("on"). To account for this, we need to build a model that is capable of describing the joint effect of both variables at once.

Strategy 1: slice and dice. One approach to modeling the joint effect of  $x_1$  and  $x_2$  on the response y is to slice and dice the data. In other words: take subsets of the data for each of the four combinations of  $x_1$  and  $x_2$ , and compute the mean within each subset. For our video-game data, we get the result in Table 4.1. Clearly the

Table 4.1: Mean reaction time across all trials and participants for the four combinations of the two experimental factors in the video game data.

Cluttered	Far away	Time (ms)
No	No Yes	491 522
Yes	No Yes	559 629

"cluttered + far away" scenes are the hardest, on average.

This slice-and-dice approach is intuitively reasonable, but combinatorially explosive. With only two binary grouping variables, we have four possible combinations—not a big deal. But suppose we had 10 binary grouping variables instead. Then there would be  $2^{10}=1024$  possible subsets of the data, and thus 1024 group-wise means to estimate. For a scenario like this, if you were to take the slice-and-dice approach, you would need a lot of data—and not merely a lot of data overall, but a lot of data for each combination separately.

Strategy 2: use dummy variables. A second strategy is to estimate the effect of  $x_1$  and  $x_2$  by building a model that uses dummy variables. Intuitively, the model we'll fit assumes that the response can be expressed as:

$$y_i = \hat{y}_i + e_i = \text{Baseline} + (\text{Effect if } x_{i1} \text{ on}) + (\text{Effect if } x_{i2} \text{ on}) + \text{Residual}.$$

Notice that we need two subscripts on the predictors  $x_{i1}$  and  $x_{i2}$ : i, to index which case in the data set is being referred to; and 1 or 2, to indicate which categorical predictor is being referred to (e.g. far away versus cluttered).

This notation gets cumbersome quickly. We can write it more concisely in terms of dummy variables, just as we learned to do in the case of a single grouping variable:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_{i1}=1\}} + \beta_2 \mathbf{1}_{\{x_{i2}=1\}} + e_i$$
.

Notice how the dummy variables affect the expected value of  $y_i$  by being either present or absent, depending on the case. For example, if  $x_{i2} = 0$ , then the  $\beta_2 \mathbf{1}_{\{x_2\}}$  term falls away, and we're left with the baseline, plus the effect of  $x_1$  being on, plus the residual. We refer to  $\beta_1$  and  $\beta_2$  as the *main effects* of the model, for reasons that will become clear in a moment.

If we fit this model to the video-game data in Figure 4.1, we get the equation

Reaction = 
$$482 + 87 \cdot \mathbf{1}_{\{x_{i1}=1\}} + 50 \cdot \mathbf{1}_{\{x_{i2}=1\}} + \text{Residual}$$
, (4.2)

where  $x_{i1} = 1$  means that the scene was cluttered, and  $x_{i2} = 1$  means that the scene was far away. This equation says that if the scene was cluttered, the average reaction time became 87 milliseconds slower; while if the scene was far away, the average reaction time became 50 milliseconds slower.

#### Interactions

A key assumption of the model in Equation 4.2 is that the effects of clutter and distance on reaction time are separable. That is, if we want to compute the joint effect of both conditions, we simply add the individual effects together.

But what if the effects of  $x_1$  and  $x_2$  aren't separable? We might instead believe a model like this:

 $y_i = \text{Baseline} + (\text{Effect if } x_1 \text{ on}) + (\text{Effect if } x_2 \text{ on}) + (\text{Extra effect if both } x_1 \text{ and } x_2 \text{ on}) + \text{Residual}$ .

In the context of our video-games data, this would imply that there's something different about scenes that are both cluttered and far away that cannot be described by just summing the two individual effects.

The world is full of situations like this, where the whole is different than the sum of the parts. The ancient Greeks referred to this idea as  $\sigma v \nu \epsilon \rho \gamma$ , or synergia. This roughly means "working together," and it's the origin of the English word "synergy." Synergies abound:

- Neither an actor nor a cameraman can do much individually, but together they can make a film.
- Two hydrogens and an oxygen make water, something completely unlike either of its constituent parts.
- Biking up a hill is hard. Biking in a big gear is hard. Biking up a hill in a big gear is impossible, unless you take drugs.

Alas, examples of the whole being worse than the sum of the parts also abound—groupthink on committees, ill-conceived corporate mergers, Tylenol and alcohol, and so forth.4

In statistics, we operationalize the idea of synergy using *inter*actions among variables. An interaction is what we get when we multiply two variables together. In the case of two binary categorical predictors, a model with an interaction looks like this:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_1=1\}} + \beta_2 \mathbf{1}_{\{x_2=1\}} + \beta_{12} \mathbf{1}_{\{x_1=1\}} \mathbf{1}_{\{x_2=1\}} + e_i$$
.

We call  $\beta_{12}$  an *interaction term*; this term disappears from the model unless  $x_1$  and  $x_2$  are both equal to 1. Fitting this model to the video-games data gives the following estimates:

Reaction = 
$$491 + 68 \cdot \mathbf{1}_{\{x_{i1}=1\}} + 31 \cdot \mathbf{1}_{\{x_{i2}=1\}} + 39 \cdot \mathbf{1}_{\{x_{i1}=1\}} \mathbf{1}_{\{x_{i2}=1\}} + \text{Residual}$$
,

We interpret this model as follows:

• The baseline reaction time for scenes that are neither cluttered nor far away is 491 milliseconds (ms).

<sup>4</sup> Don't take Tylenol and alcohol together or you'll risk liver damage.

- The main effect for the "cluttered" variable is 68 ms.
- The main effect for the "far away" variable is 31 ms.
- The interaction effect for "cluttered" and "far away" is 39 ms.
  In other words, scenes that are both cluttered and far away
  yield average reaction times that are 39 milliseconds slower
  than what you would expect from summing the individual
  effects of the two variables.

From these main effects and the interaction we can use the model to summarize the expected reaction time under any combination of experimental variables:

- $(x_1 = 0, x_2 = 0)$ :  $\hat{y} = 491$  (neither cluttered nor far).
- $(x_1 = 1, x_2 = 0)$ :  $\hat{y} = 491 + 68 = 559$  (cluttered, near).
- $(x_1 = 0, x_2 = 1)$ :  $\hat{y} = 491 + 31 = 522$  (not cluttered, far).
- $(x_1 = 1, x_2 = 1)$ :  $\hat{y} = 491 + 68 + 31 + 39 = 629$  (cluttered, far).

A key point regarding the fourth case in the list is that, when a scene is both cluttered and far away, both the main effects *and* the interaction term enter the prediction. You should also notice that these predictions exactly match up with the group means in Table 4.1 on page 84.

#### Incorporating still more categorical predictors

Once you understand the basic recipe for incorporating two categorical predictors, you can easily extend that recipe to build a model involving more than two. For example, let's return one last time to the video-game data in Figure 4.6 on page 84. So far, we've been ignoring the bottom panel, which shows systematic differences in reaction times across different subjects in the study. But we can also incorporate subject-level dummy variables to account for these differences. The actual model equation starts to get ugly with this many dummy variables, so we often use a shorthand that describes our model intuitively rather than mathematically:

Time 
$$\sim$$
 Clutter effect + (Distance effect) (4.3)  
+ (Interaction of distance/clutter) + (Subject effects).

Here the  $\sim$  symbol means "is modeled by" or "is predicted by." There are 12 subjects in the data set. Thus to model the subject-level effects, we introduce 11 dummy variables, in a manner similar to what was done in Equation 4.1. The estimated coefficients for this model are in Table 4.2.

Table 4.2: Fitted coefficients for the model incorporating subject-level dummy variables into the video-game data. Remember, K levels of a factor require K-1 dummy variables, because one level—in this case, the subject labeled "Subject 6" in Figure 4.6—is the baseline.

Variable	$\hat{eta}$
Intercept	570
Cluttered	68
FarAway	31
Subject 8	-90
Subject 9	-136
Subject 10	-44
Subject 12	-76
Subject 13	-147
Subject 14	-112
Subject 15	-93
Subject 18	-8
Subject 20	-118
Subject 22	-34
Subject 26	-79
Cluttered:FarAway	39

When to include interactions. In the model above, we're assuming that clutter and distance affect all subjects in the same way. Thus we have 15 parameters to estimate: an intercept/baseline, two main effects for Littered and FarAway, one interaction term, and 11 subject-level dummy variables. If instead we were to compute the groupwise means for all possible combinations of subject, clutter, and distance, we'd have 48 parameters to estimate: the group mean for each combination of 12 subjects and 4 experimental conditions. Moreover, we'd be implicitly assuming an interaction between the experimental conditions and the subject, allowing clutter and distance to affect each person's average reaction time in a different way, rather than all people in the same way.

This example should convey the power of using dummy variables and interactions to express how a response variable changes as a function of several grouping variables. This framework forces us to be explicit about our assumptions, but it also allows us to be selective about the complexity of our models. Compare estimating 15 parameters versus estimating 48 parameters in the video-games example—that's a big difference in what we're asking of our data.

The essence of the choice is this:

- If a variable affects the response in a similar way under a broad range of conditions, regardless of what the other variables are doing, then it probably warrants only a main effect in our model.
- But if a variable's effect is modulated by some other variable, we should describe that using an interaction between them.

The choice of which variables interact with which other ones should ideally be guided by knowledge of the problem at hand. For example, in a rowing race, a strong headwind makes all crews slower. But wind affects lighter crews more than heavier crews: weight modulates the effect of wind. Thus if we want to build a model to predict the winner of an important race, like the one between Oxford and Cambridge every spring on the Thames, we should strongly consider including an interaction between wind speed and crew weight. This is something that anyone with knowledge of rowing could suggest, even before seeing any data. But the choice of whether to include an interaction term in a model can also be guided by the data itself. We will now learn about a process called the analysis of variance that can help us address this important modeling question.

Before we get there, however, here's one final generic guideline

about interactions: it is highly unusual to include an interaction in a regression model without also including both corresponding main effects. There are various technical math reasons why most textbooks warn you about this, and why I'm doing so now. But the most important concern is that it is very difficult to interpret a model having interaction terms but no main effects. You should fit such a model only if you have a very good reason.

### ANOVA: bookkeeping for sums of squares

The model in Equation 4 postulates four effects on the reaction time for the video-game data: (1) an effect due to visual clutter; (2) an effect due to distance of the stimulus in the scene; (3) an interaction effect (synergy) of distance and clutter; and (4) effects due to differences among experimental subjects. The  $R^2$  for this model is about 0.23, and the residual standard deviation is about 126 milliseconds. This tells us something about the overall predictive abilities of the model. But can we say something about the predictive abilities of the individual variables within this model?

Yes, we can, by conducting an analysis of variance (ANOVA). An analysis of variance is just a simple book-keeping exercise aimed at attributing credit to individual variables in a model. To run an ANOVA, we build a model one step at time, adding one new variable (or one new interaction among variables) at each step. Every time we do this, we ask two questions:

- (1) How many parameters did we have to add to the model to account for the effects of this variable?<sup>5</sup>
- (2) By how much did we reduce the unpredictable variation in the response when we added this variable? Remember the variance decomposition:

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

$$\text{TV} = \text{PV} + \text{UV}.$$

Every time we add a new variable to a model, the total variation in the response variable (TV) stays the same, but we move some of this variation out of the "unpredictable" column (UV) and into the "predictable" (PV) column. In ANOVA, we keep track of the precise numerical value of this improvement.

<sup>&</sup>lt;sup>5</sup> For example, we needed to add 11 parameters to account for the "Subject" variable in the video-games data, because we needed to represent this information in terms of 11 dummy variables.

The final result of an analysis of variance is a table—called the ANOVA table—that shows the answers to these two questions at each model-building step.

Let's take the specific example of our model for the videogames data, for which TV = 39,190,224. We'll add one variable at a time and track how TV is partitioned among PV and  $UV.^6$ 

Step 1. First, we add an effect due to visual clutter (Time  $\sim$  Clutter). The variance decomposition for this model is

$$39,190,224 = 3,671,938 + 35,518,285$$
.

Thus the clutter effect gets credit for predicting 3,671,938 (out of a possible 39,190,224) units of total variation, at the cost of adding one parameter to the model.

Step 2. Next, we add the distance effect to the model already containing the clutter variable (Time  $\sim$  Clutter + Distance). The new variance decomposition is:

$$39,190,224 = 4,878,397 + 34,311,827.$$
<sub>TV</sub>

The previous PV was 3,671,938, and the new one is 4,878,397. Thus the distance effect gets credit for 4,878,397 - 3,671,938 = 1,206,459 units of total variation.

Step 3. Third, we add the interaction of distance and clutter to the previous model (Time  $\sim$  Clutter + Distance + Clutter:Distance). The new variance decomposition is:

$$39, 190, 224 = 5,062,030 + 34,128,194.$$
 $_{TV}^{PV}$ 

The previous PV was 4,878,397, and the new one is only slightly better at 5,062,030. Thus the interaction effect gets credit for a measly 5,062,030 - 4,878,397 = 183,633 units of total variation.

Step 4. Finally—almost done here—we add the 11 subject-level dummy variables to the previous model (Time  $\sim$  Clutter + Distance + Clutter:Distance + Subject). The new variance decomposition reveals a big bump in PV:

$$39, 190, 224 = 9, 122, 852 + 30, 067, 371.$$

<sup>6</sup> The quantity TV = 39,190,224 highlights one feature that makes ANOVA tricky at first: the units are non-intuitive, since we measure improvement using sums of squares. Here the units are squared milliseconds; when you square a quantity like 1000 ms (1 second), you get 1,000,000 ms<sup>2</sup>, which is why we're seeing numbers in the millions here.

# Pars	$\Delta$ PV	$R^2$
1	3671938	0.094
1	1206459	0.125
1	183633	0.129
11	4060822	0.233
ariation	9122852	
Unpredictable Variation		
Total Variation		
	1 1 1 11 ariation	1 3671938 1 1206459 1 183633 11 4060822 ariation 9122852 ariation 30067371

Table 4.3: The analysis of variance (ANOVA) table for the model incorporating effects due to clutter, distance, and subject, along with an interaction between clutter and distance. In an ANOVA table, we add each variable in stages, one at a time. "# Pars" refers to the number of new parameters added to the model at each stage.  $\Delta$  PV refers to the change in predictable variation at each stage.  $R^2$  is the coefficient of determination for the model at each stage.

The previous PV was 5,062,030, and the new one is only slightly better at 9,122,852. Thus the subject effects get credit for 9,122,852-5,062,030=4,060,822 units of total variation.

*Interpreting the ANOVA table.* As you've now seen, the analysis of variance really is just bookkeeping! The ANOVA table for the final model (Time ~ Clutter + Distance + Clutter:Distance + Subject) is shown in Table 4.3. The chance in predictable variation ( $\Delta$ PV) at each stage gives us a more nuanced picture of the model, compared with simply quoting  $R^2 = 0.23$ , because it allows us to partition credit among the individual predictor variables in the model. In Table 4.3, it's clear that subject-level variation is largest, followed by variation due to clutter and then variation due to distance. The distance-clutter interaction contributes a small amount to the predictive ability of the model, relatively speaking: it improves  $R^2$  by only half a percentage point, from 0.124 to 0.129. This tells us that any modulating effect between distance and clutter is relatively small. In fact, the distance/clutter interaction looks so negligible that we might even consider removing this effect from the model, just to simplify. We'll revisit this question later in the book, when we learn some more advanced tools for statistical hypothesis testing and predictive model building.

Finally, always remember that the construction of an ANOVA table is inherently sequential. For example, first we add the clutter variable, which remains in the model at every subsequent step; then we add the distance variable, which remains in the model at every subsequent step; and so forth. Thus the actual question being answered at each stage of an analysis of variance is: how much variation in the response can this new variable predict, in the context of what has already been predicted by other variables in the

model? This point—the importance of context in interpreting an ANOVA table—is subtle, but important. We'll revisit it soon, when we discuss the issues posed by correlation among the predictor variables in a regression model.

#### Numerical and grouping variables together

Now we are ready to add a continuous predictor into the mix. Start with the simplest case of 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 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:

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

Now each group has its own regression equation:

Regression equation for case where 
$$(x_i = 0)$$
:  $y_i = \beta_0 + \beta_2 x_{i2} + e_i$   
Regression equation for case where  $(x_i = k)$ :  $y_i = (\beta_0 + \beta_1^{(k)}) + \beta_2 x_{i2} + e_i$ .

Each line has a different intercept, but they all have the same slope. These are the red lines in Figure 4.3 back on page 78.

The coefficients  $\beta_1^{(k)}$  are associated with the dummy variables that encode which college a student is in. 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 regression 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 (11 parameters total). This is clearly a compromise solution between two extremes: fitting a single model, with one slope and one intercept common to all colleges (2 parameters); versus fitting ten distinct models for the ten individual colleges, each with their slope and intercept (20 parameters).

Interactions between grouping and numerical variables

We can also have modulating effects between numerical and grouping predictors. 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. Mathematically, this means that College modulates the slope of the linear relationship between GPA and SAT scores.

If this is the case, then we should include an interaction term in the model. Remember, in statistical models, interactions are formed by multiplying two predictors together—in this case, a numerical predictor and a dummy (o–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 in Figure 4.7 on page 95. 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 dummy variables for a player's league. The corresponding model equation looks like this:

$$\hat{y}_i = \beta_0 + \underbrace{\beta_1^{(AAA)} \cdot 1_{AAA} + \beta_1^{(MLB)} \cdot 1_{MLB}}_{\text{Dummy variables}} + \beta_1 \cdot AVG$$

The three lines are parallel: the coefficients on the dummy vari-

ables shift the line up or down as a function of a player's league.

But if we want the slope to change with league as well—that is, if we want league to modulate the relationship between salary and batting average—then we must fit a model like this:

$$\hat{y}_i = \beta_0 + \underbrace{\beta_1^{(AAA)} \cdot 1_{AAA} + \beta_1^{(MLB)} \cdot 1_{MLB}}_{\text{Dummy variables}} + \beta_2 \cdot AVG + \underbrace{\beta_3^{(AAA)} \cdot AVG \cdot 1_{AAA} + \beta_3^{(MLB)} \cdot AVG \cdot 1_{MLB}}_{\text{Interaction terms}}$$

The y variable depends on  $\beta_0$  and  $\beta_2$  for all players, regardless of league. But when a player is in AAA, the corresponding dummy variable ( $1_{AAA}$ ) fires. Before, when a dummy variable fired, the entire line was merely shifted up for down (as in Figure 4.7). Now, an offset to the intercept ( $\beta_1^{(AAA)}$ ) and an offset to slope ( $\beta_3^{(AAA)}$ ) are activated. Ditto for players in the Major League: then the MLB dummy variable ( $1_{MLB}$ ) fires, and both an offset to the intercept ( $\beta_1^{(MLB)}$ ) and an offset to the slope ( $\beta_3^{(MLB)}$ ) are activated:

Regression equation for AA: 
$$y_i = (\beta_0)$$
  $+(\beta_2) \cdot AVG$   $+e_i$   
Regression equation for AAA:  $y_i = (\beta_0 + \beta_1^{(AAA)})$   $+(\beta_2 + \beta_3^{(AAA)}) \cdot AVG$   $+e_i$   
Regression equation for MLB:  $y_i = (\beta_0 + \beta_1^{(MLB)})$   $+(\beta_2 + \beta_3^{(MLB)}) \cdot AVG$   $+e_i$ .

Fitting such model produces a picture like the one in Figure 4.8. 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 ***
```

- - -

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	

- - -

Multiple R-squared: 0.8514, Adjusted R-squared: 0.846

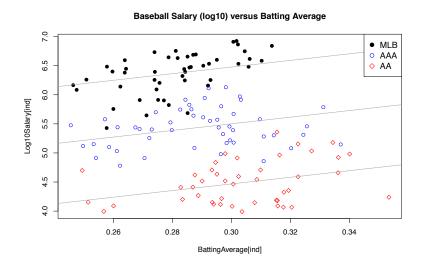


Figure 4.7: Baseball salaries versus batting average for Major League, AAA, and AA players.

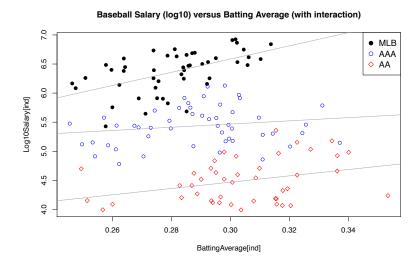
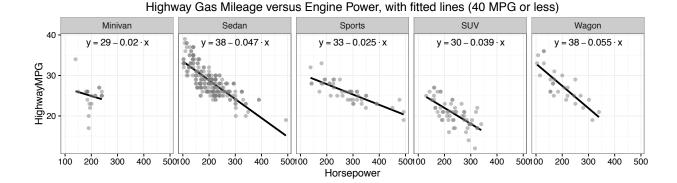


Figure 4.8: Baseball salaries versus batting average for Major League, AAA, and AA players. The fitted lines show the model with an interaction term between batting average and league.



# Dependence among predictors

In this section, we'll discuss the issue of how to interpret an analysis of variance for a model where the predictors themselves are correlated with each other. (Another term for correlation among predictors is *collinearity*.) This discussion will expand upon a point raised before—but only briefly—about the importance of context in the sequential construction of an ANOVA table.

Let's briefly review the analysis of variance (ANOVA). You'll recall that, in our look at the data on reaction time in video games, we ran an ANOVA (Table 4.3) of a regression model that predicted variation in human reaction time in terms of distance, visual clutter, subject-level variation, and a distance/clutter interaction. Our goal was to apportion credit among the individual parts of the model, where "credit" was measured by each variable's contribution to the predictable variation in the model's variance decomposition (TV = PV + UV). This led us, for example, to the conclusions that subject-level variation was large relative to the other effects, and that the distance/clutter interaction contributed only a modest amount to the predictive abilities of the model.

We can also run an analysis of variance on models containing numerical predictors. To see this in action, let's revisit the data on the gas mileage of cars from Figure 1.13, back on page 30. Recall that this data set involved 387 cars and three variables: gas mileage, engine horsepower, and vehicle class (minivan, sedan, sports car, SUV, or wagon). We can see this data once more in Figure 4.9, which shows a lattice plot of mileage versus horsepower, stratified by vehicle class.

Figure 4.9: A model for the car-mileage data involving an interaction between class and horsepower. Here we've focused only on cars whose gas mileage is less than 40 miles per gallon. For this subset of the data, linearity looks like a reasonable, if imperfect, assumption.

In our earlier discussion of this data, we noted two facts:

- (1) The classes exhibit systematic differences in their typical mileages. For example, sedans have better gas mileage, on average, than SUVs or minivans.
- (2) Vehicle class seems to modulate the relationship between MPG and engine power. As engine power increases, mileage gets worse on average, regardless of vehicle class. But this drop-off is steeper for wagons than for sports cars.

Previously, we described these facts only informally. But we now have the right tools—dummy variables and interactions—that allow us to quantify them in the context of a regression model. Specifically: point (1) suggests that we need class-level dummy variables, to move the intercepts up and down as appropriate for each class; while point (2) suggests that we need an interaction between class and horsepower, to make the slope of the regression line get steeper or shallower as appropriate for each class. Using our informal notation from earlier, our regression model should look like this:

 $MPG \sim Horsepower + Class + Class: Horsepower \,.$ 

Upon fitting this model by least squares, we get the coefficients in Table 4.4, at right. The corresponding fitted lines within each class are also shown in Figure 4.9. The parameters of this fitted model confirm our earlier informal observations based on the lattice plot: that both the average mileage and the steepness of the mileage/horsepower relationship are affected by vehicle class.

An analysis of variance table for this model looks like this.

 $R^2$ Variable added # Pars  $\Delta$  PV Horsepower 1 3372 0.43 Class 2372 0.73 Horsepower:Class 143 0.74 Predictable Variation 5887 Unpredictable Variation 2036 **Total Variation** 7923

According to this table, we can attribute most of the credit for predicting fuel economy to the horsepower variable ( $\Delta PV = 3372$ ,

Table 4.4: Fitted coefficients (rounded to the nearest hundredth) for the model that predicts car gas mileage in terms of engine horsepower, vehicle class, and a class/horsepower interaction.

β̂	Variable
28.86	Intercept
-0.02	Horsepower
9.28	Sedan
4.08	Sports
0.94	SUV
9.55	Wagon
-0.03	Horsepower:Sedan
-0.01	Horsepower:Sports
-0.02	Horsepower:SUV
-0.04	Horsepower:Wagon

Table 4.5: An analysis of variance (ANOVA) table for the model that predicts highway gas mileage in terms of a car's engine power and vehicle class, including both main effects and an interaction term. In this ANOVA table, the horsepower variable has been added first, followed by vehicle class.

or equivalently  $\Delta R^2 = 0.43$ ). Most of the remaining credit goes to vehicle class ( $\Delta PV = 2372$ , or equivalently  $\Delta R^2 = 0.30$ ). The interaction produces a modest change in  $R^2$ ; this bears out the visual impression conveyed by Figure 4.9, in which the slopes in each panel are clearly different, but not dramatically so.

But this conclusion about the relative importance of horsepower and vehicle class involves a major, even deal-breaking, caveat. Remember that an analysis of variance is inherently sequential: first we add the horsepower variable, then we add vehicle class, and then we add the interaction, tracking the variance decomposition at each stage. What happens if we build an ANOVA table by adding vehicle class before we add horsepower?

Variable added	# Pars	Δ PV	$R^2$
Class	4	3144	0.40
Horsepower	1	2601	0.73
Horsepower:Class	4	143	0.74
Predictable V	5887		
Unpredictable Variation		2036	
Total Variation		7923	

Now we reach the opposite conclusion: that vehicle class contributes more ( $\Delta PV = 3144$ ) to the predictable variation than does horsepower ( $\Delta PV = 2601$ ). Why does this happen? How could our conclusion about the relative importance of the variables depend upon something so arbitrary as the order in which we decide to add them?

Shared versus unique information. Figure 4.10 provides some intuition why this is so. In our data on gas mileage, the two predictors (horsepower and vehicle class) are correlated with each other: vehicles in certain classes, like SUVs and sports cars, have more powerful engines on average than sedans, wagons, and minivans.

To understand why this correlation between predictors would matter so much in an analysis of variance, let's consider the information provided by each variable. First, a vehicle's class tells us at least two important things relevant for predicting gas mileage.

- 1) Weight: for example, SUVs tend to be heavier than sedans, and heavier vehicles will get poorer gas mileage.
- 2) Aerodynamics: for example, minivans tend to be boxier than

Table 4.6: A second analysis of variance (ANOVA) table for the model that predicts highway gas mileage in terms of a car's engine power and vehicle class, including both main effects and an interaction term. In this ANOVA table, vehicle class has been added first, followed by horsepower.

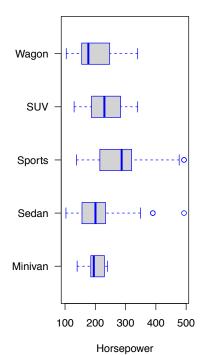


Figure 4.10: Correlation between vehicle class and horsepower.

sports cars, and boxier cars will get poorer gas mileage due to increased drag at highway speeds.

Similarly, the horsepower of a vehicle's engine also tells us at least two important things relevant for predicting gas mileage.

- 1) Weight: more powerful engines are themselves heavier, and tend to come in cars that are heavier in other ways, too.
- 2) Fuel consumption: a smaller engine consumes less fuel and typically has better mileage than a bigger engine.

Notice that both variables provide information about a vehicle's weight; let's call this the shared information. But each also provides information on something else specific to that variable; let's call this the unique information. The shared information between the predictors manifests itself as correlation: bigger cars tend to have both bigger engines, and they also to be in certain classes. We can use a Venn diagram to represent both the shared and the unique information provided by the predictors in a stylized (i.e. non-mathematical) way:

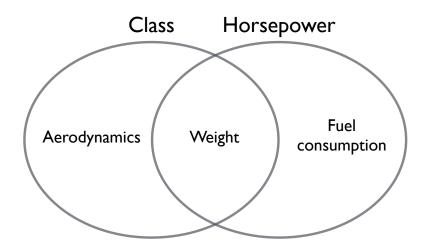


Figure 4.11: The two predictors in the gas-mileage data set provide some information content that is shared between them, in addition to some information that is unique to each one.

In the first analysis of variance (Table 4.5), we added horse-power first. When we did so, the regression model greedily used all the information it could from this predictor, including both the "shared" and "unique" information. As a result, when we added the class variable second, the shared information is redundant—it was already accounted for by the model. We therefore end up giving the class variable credit only for its unique information

content; all the information content it shares with horsepower was already counted in step 1. This is illustrated in Figure 4.12.

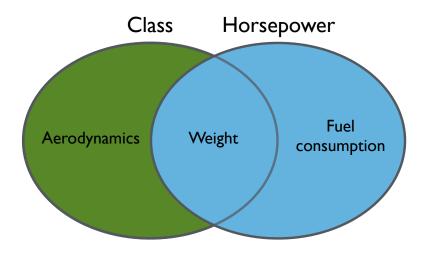


Figure 4.12: Our model for gas mileage includes two variables: engine horsepower and vehicle class. These variables both convey information about a vehicle's size, in addition to some unique information (e.g. class tells us about aerodynamics, while horsepower tells us about fuel consumption). When we add the Horsepower variable first in an analysis of variance (Table 4.5), we attribute all of the shared information content to Horsepower, and none to Vehicle class, in our ANOVA table.

But when we flip things around and add vehicle class to the model first (Table 4.6), this picture change. We end up giving the class variable credit both for its unique information content and for the information it shares with Horsepower. This leaves less overall credit for Horsepower when we add it in step 2 of the ANOVA. This is illustrated in Figure 4.13.

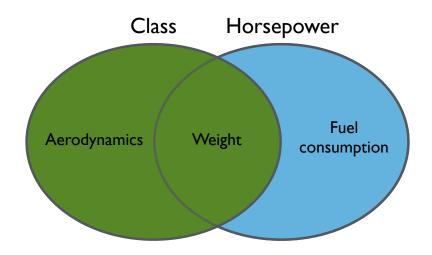


Figure 4.13: (Continued from Figure 4.12.) But when we add the Class variable first in an analysis of variance (Table 4.5), we attribute all of the shared information content to Class, and none to Horsepower, in our ANOVA table.

This example highlights an unsatisfying but true feature of the analysis of variance: when the variables are correlated, their ordering matters when you build the ANOVA table.

This feature of an ANOVA table at first seems counterintuitive, even disturbing. Yet similar phenomena occur all the time in everyday life. A good analogy here is the dessert buffet at Thanksgiving dinner. Imagine two different versions of dessert.

Version 1: After dinner, your aunt offers you apple pie, and you eat your fill. The apple pie is delicious—you were really looking forward to something sweet after a big Thanksgiving meal. It makes you very happy.

Next, after you've eaten your fill of apple pie, your aunt offers you pumpkin pie. Pumpkin pie is also delicious you love it just as much as apple. But your dessert tummy is pretty full already. You eat a few bites, and you enjoy it; that spicy pumpkin flavor is a little different to what you get from an apple pie. But of course, pumpkin pie is still a dessert, and you don't enjoy it as much as you might have if you hadn't eaten so much apple pie first.

Version 2: After dinner, your aunt offers you pumpkin pie, and you eat your fill. The pumpkin pie is delicious—all that whipped cream on top goes so well with the nutmeg and earthy pumpkin flavor. It makes you very happy.

Next, after you've eaten your fill of pumpkin pie, your aunt offers you apple pie. Apple pie is also delicious—you love it just as much as pumpkin. But your dessert tummy is pretty full already. You eat a few bites, and you enjoy it; those tart apples with all the sugar and cinnamon give a little different flavor to what you get from a pumpkin pie. But of course, apple pie is still a dessert, and you don't enjoy it as much as you might have if you hadn't eaten so much pumpkin pie first.

That evening, which pie are you going to remember? In version 1, you'll attribute most of your Thanksgiving dessert afterglow to the apple pie; while in version 2, you'll attribute most of it to pumpkin pie. Context matters, even if in the abstract you like both pies the same amount.

An analysis of variance is like the one-at-a-time dessert eater at Thanksgiving. Whatever variable we add to the model first, the ANOVA greedily eats its fill of that, before turning to the second variable. This affects how credit gets attributed. In our ANOVA tables for the gas mileage data, our two variables (horsepower and vehicle class) are like apple and pumpkin pie. Yes, they each offer something unique, but they also share a lot of their information content (just like the pies are both desserts). Because of this, the order in which they are added to the ANOVA table—or equivalently, the context in which each variable's marginal contribution to the model is evaluated—matters a lot.

The moral of the story is that it rarely makes sense to speak of "the" ANOVA table for a model—only "an" ANOVA table. Thus there is no unique way to partition credit among multiple variables for their shared information content in a regression model. We must make an arbitrary choice, and in an ANOVA table, that choice is "winner take all" to the first variable added to the model.

Final thoughts on ANOVA. There are two further points to bear in mind about the analysis of variance. First, the ANOVA table is not the model itself, only an attempt to partition credit for predicting the outcome among the variables in the model by adding those variables one at a time. And while the ANOVA table is order-dependent, the model itself isn't. Regardless of the order in which you add variables, you will always get the same model coefficients, fitted values, and residuals at the end.

Second, we've discussed the subtleties of interpreting an ANOVA table in the presence of correlation among the predictors. However, if the variables in the model are independent of one another, then they have no shared information content, and the ANOVA table does not depend upon the ordering of the variables.

This is why we ignored the issue of variable ordering when building an ANOVA table for our model of reaction time in video games versus distance, clutter, and subject-level effects. For that data set, the predictor variables were independent with each other: the experimental design was perfectly balanced, with each subject sitting for exactly 40 trials for each pairwise combination of the cluttered and distance variables. Regardless of the order in which we add the variables, we will always get the same  $\Delta PV$  for each one. Thus in the absence of dependence among the predictors, we can uniquely assign credit for predicting the outcome to each one.<sup>7</sup>

Regression models, just like Thanksgiving diners, thrive on variety—that is, on multiple independent sources of information.

<sup>&</sup>lt;sup>7</sup> For this reason, ANOVA is a commonly used tool in the analysis of designed experiments, when we can ensure that the predictors are independent of one another. It is less common in the analysis of observational studies, where the inevitable presence of collinearity significantly weakens the conclusions that we can draw from an ANOVA.