13
Grouping variables in regression

Models with 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 β_1 can be interpreted as the differential effect on the outcome (y) of having the dummy variable equal to 1, rather than 0.

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 13.1.

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 13.1 shows the results of one such study. Participants were presented with a natural

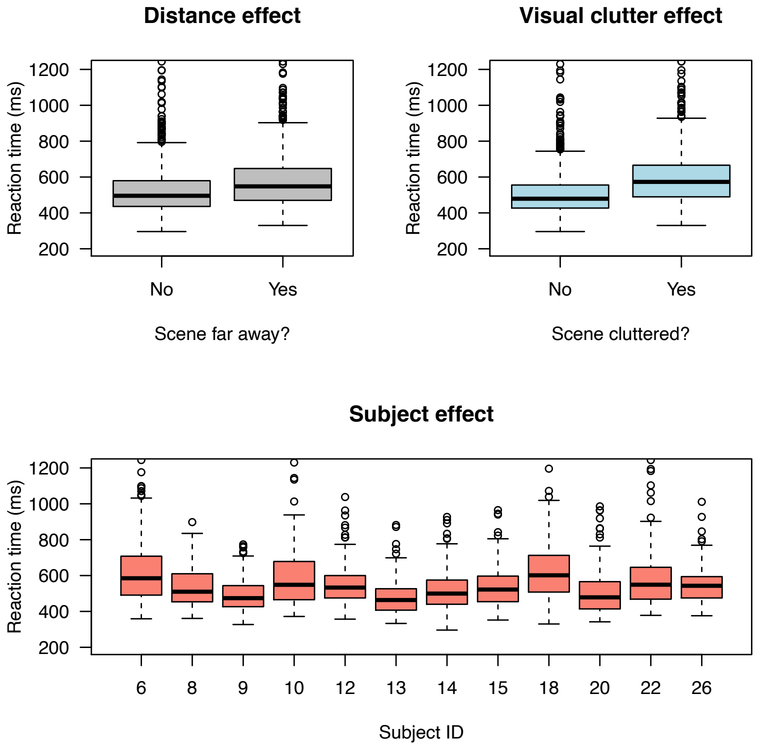

Figure 13.1: Reaction time to visual stimuli in a controlled experiment run by a major video-game maker. Top-left: participants reacted more slowly, on average, when the stimulus was far away within the scene. Top-right: participants reacted more slowly, on average, in a scene with significant visual clutter. Bottom: systematic differences in reaction time across participants in the trial.

scene on a computer monitor, and asked to react (by pressing a button) when they saw an animated figure appear in the scene.[1]

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 Figure 13.1 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 13.1 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 de-

[1] Essentially the company was measuring how quickly people could react to a bad guy popping up on the screen in a video game.

scribing 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 13.1. Clearly the "cluttered + far away" scenes are the hardest, on average.

Table 13.1: 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	491
	Yes	522
Yes	No	559
	Yes	629

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 β_1 and β_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 13.1, we get the equation

$$\text{Reaction} = 482 + 87 \cdot \mathbf{1}_{\{x_{i1}=1\}} + 50 \cdot \mathbf{1}_{\{x_{i2}=1\}} + \text{Residual}\,, \qquad (13.1)$$

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 13.1 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 $\sigma\upsilon\nu\epsilon\rho\gamma$, 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 performance-enhancing drugs.

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.[2]

In statistics, we operationalize the idea of synergy using *interactions among variables.* An interaction is what we get when we multiply two variables together. In the case of two binary categori-

[2] Don't take Tylenol and alcohol together or you'll risk liver damage.

cal 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 β_{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:

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

We interpret this model as follows:

- The baseline reaction time for scenes that are neither cluttered nor far away is 491 milliseconds (ms).
- 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 13.1 on page 127.

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 13.1 on page 126. 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:

$$\begin{aligned} \text{Time} \quad \sim \quad & \text{Clutter effect} + (\text{Distance effect}) \\ + \quad & (\text{Interaction of distance/clutter}) + (\text{Subject effects})\,. \end{aligned} \tag{13.2}$$

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 9.1. The estimated coefficients for this model are in Table 13.2.

Table 13.2: 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 13.1—is the baseline.

Variable	$\hat{\beta}$
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 that variable 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 those two variables.

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: the analysis of variance

The model in Equation 13.2 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?[3] This is usually called the *degrees of freedom* associated with that parameter.

(2) By how much did we improve the predictive power of the model when we added this variable? Anytime we add a variable to a model, R^2 will go up. In ANOVA, we keep track of the precise numerical value of this change in R^2. Larger changes in R^2 mean correspondingly larger improvements in the predictive power of the model.

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 video-games data. We'll add one variable at a time and track how much of the variation in y is predictable versus unpredictable.

Step 1. First, we add an effect due to visual clutter (Time $\sim$ Clutter). The R^2 for this model is 0.094.

Step 2. Next, we add the distance effect to the model already containing the clutter variable (Time $\sim$ Clutter + Distance). The new R^2 for this model is 0.125. Including the distance variable improved R^2 by about 3%.

Step 3. Third, we add the interaction of distance and clutter to the previous model (Time $\sim$ Clutter + Distance + Clutter:Distance). The new model has an R^2 of 0.129, meaning that the clutter:distance interaction improved R^2 by only about half a percent.

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 R^2 is 0.223, meaning that including Subject-level dummy variables improved R^2 by about 10%.

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 $\sim$ Clutter + Distance + Clutter:Distance + Subject) is shown in Table 13.3. The change in predictable variation at each stage gives us a more nuanced picture of the model, compared

[3] 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.

Variable added	# Pars (DF)	R^2	ΔR^2	s_e	Δs_e
Intercept only	1	0.000		142.9	
Clutter	1	0.094	0.094	136.1	6.8
Distance	1	0.125	0.031	133.8	2.3
Clutter:Distance	1	0.129	0.005	133.5	0.3
Subject	11	0.223	0.104	125.6	7.9

Table 13.3: 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. R^2 is tracked at each stage, and s_e is the standard deviation of the model residuals. Remember that R^2 always goes up when we add a variable; the important question in ANOVA is, by how much?

with simply quoting R^2, because it allows us to partition credit among the individual predictor variables in the model.

The most intuitive way to summarize this information is to track the change in R^2 and residual standard deviation (s_e) at each step. For example, in Table 13.3, it's clear that accounting for subject-level variation improves our predictions the most, followed by clutter and then 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. In fact, the distance/clutter interaction looks so negligible that we might even consider removing this effect from the model, just to simplify.

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.

Let's take a simple example involving baseball salaries, plotted in Figure 13.2 on page 135. On the y-axis are the salaries of 142 baseball players, measuring on a logarithmic scale. 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 variables 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 β_0 and β_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 13.2). 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:

$$\begin{aligned}
\text{Regression equation for AA:} \quad & y_i = (\beta_0) & &+ (\beta_2) \cdot AVG & &+ e_i \\
\text{Regression equation for AAA:} \quad & y_i = (\beta_0 + \beta_1^{(AAA)}) & &+ (\beta_2 + \beta_3^{(AAA)}) \cdot AVG & &+ e_i \\
\text{Regression equation for MLB:} \quad & y_i = (\beta_0 + \beta_1^{(MLB)}) & &+ (\beta_2 + \beta_3^{(MLB)}) \cdot AVG & &+ e_i \, .
\end{aligned}$$

Fitting such model produces a picture like the one in Figure 13.3.

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
```

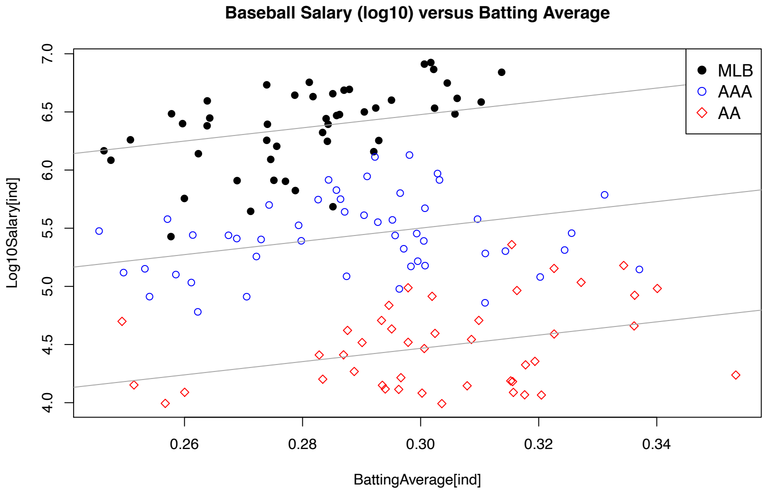

Figure 13.2: Baseball salaries versus batting average for Major League, AAA, and AA players. The lines show a linear fit of log salary versus batting average, with a main effect for league (MLB/AAA/AA) but no interaction. As a result, the three lines for the three different leagues are parallel.

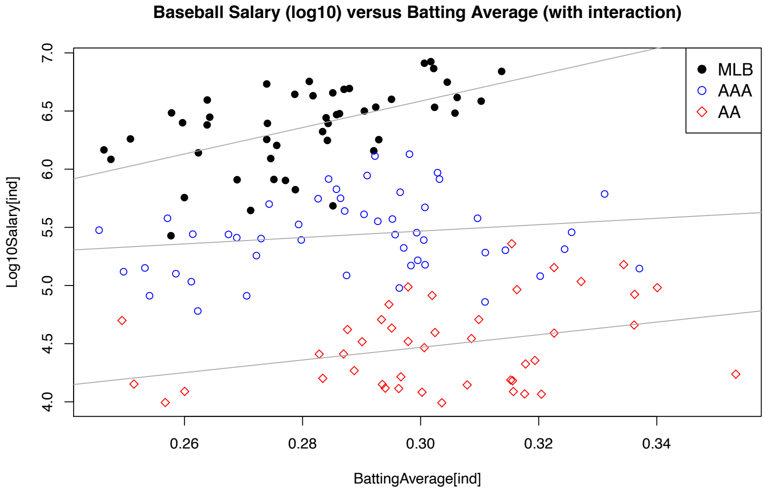

Figure 13.3: Baseball salaries versus batting average for Major League, AAA, and AA players. The lines show a linear fit of log salary versus batting average, with an interaction term between batting average and league. As a result, the three lines for the three different leagues are not parallel.

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
```

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 13.3) 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 improvement to the model's R^2. 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 13.4, which shows a faceted plot of mileage versus horsepower, stratified by vehicle class.

We can see two facts from this figure:

(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

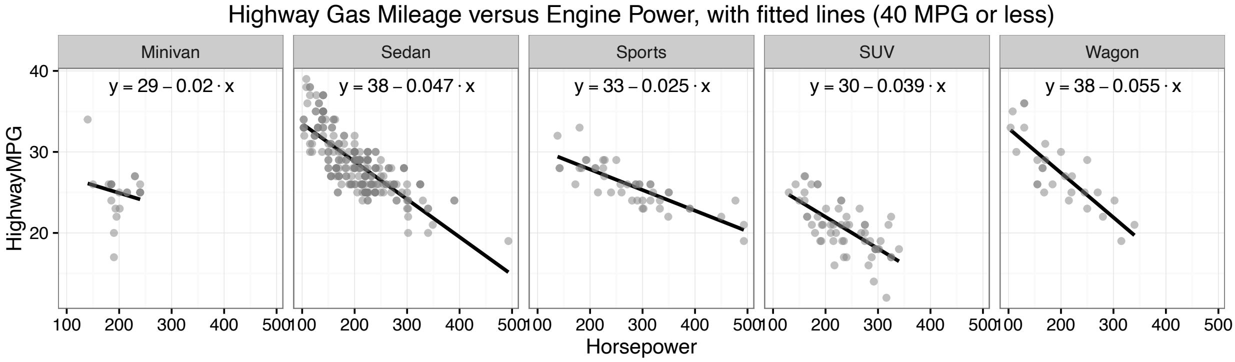

Figure 13.4: A model for the car-mileage data involving an interaction between class and horsepower. Here we've focused only on cars whose gas mileage is less than 40 miles per gallon, where linearity looks like a reasonable assumption.

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.

We now have the right tools—dummy variables and interactions—that allow us to quantify these facts 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. Upon fitting this model by least squares, we get the coefficients in Table 13.4, at right. The corresponding fitted lines within each class are also shown in Figure 13.4. The parameters of this fitted model confirm our informal observations based on the faceted plot: that both the average mileage and the steepness of the mileage/horsepower relationship are affected by vehicle class.

Table 13.4: 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.

Variable	$\hat{\beta}$
Intercept	28.86
Horsepower	-0.02
Sedan	9.28
Sports	4.08
SUV	0.94
Wagon	9.55
Horsepower:Sedan	-0.03
Horsepower:Sports	-0.01
Horsepower:SUV	-0.02
Horsepower:Wagon	-0.04

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

Variable added	# Pars	R^2	ΔR^2	s_e	Δs_e
Intercept only	1	0		4.59	
Horsepower	1	0.426	0.426	3.48	1.11
Class	4	0.725	0.299	2.42	1.06
Horsepower:Class	4	0.743	0.018	2.36	0.07

Table 13.5: 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.

According to this table, we can attribute most of the credit for predicting fuel economy to the horsepower variable ($\Delta R^2 = 0.426$). Most of the remaining credit goes to vehicle class ($\Delta R^2 = 0.299$).

The interaction produces a modest change in R^2; this bears out the visual impression conveyed by Figure 13.4, 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	R^2	ΔR^2	s_e	Δs_e
Intercept only	1	0		4.59	
Class	4	0.397	0.397	3.58	1.01
Horsepower	1	0.725	0.328	2.42	1.16
Class:Horsepower	4	0.743	0.018	2.36	0.07

Table 13.6: 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.

Now we reach the opposite conclusion: that vehicle class contributes more ($\Delta R^2 = .397$) to the predictable variation than does horsepower ($\Delta R^2 = .328$). 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 13.5 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.

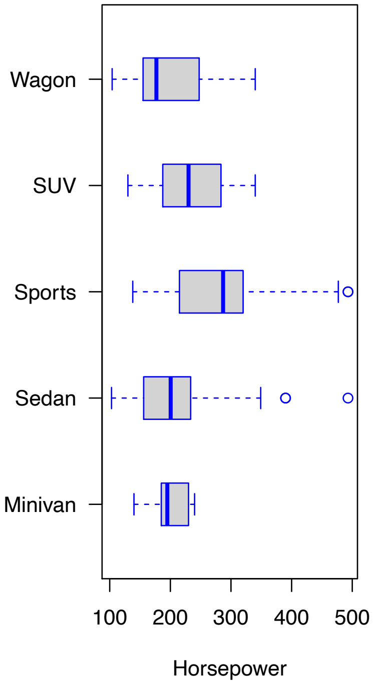

Figure 13.5: Correlation between vehicle class and horsepower.

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 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:

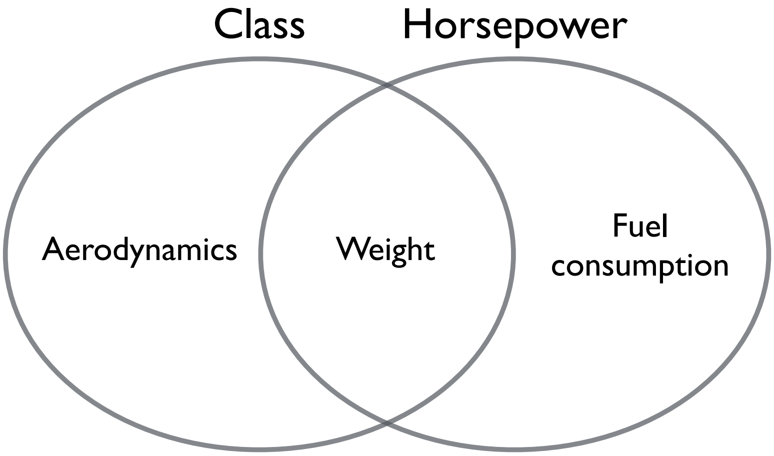

Figure 13.6: The two predictors in the gas-mileage data set provide some information content that is shared between them, in addition to some information that is unique to each one.

In the first analysis of variance (Table 13.5), we added horsepower 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 13.7.

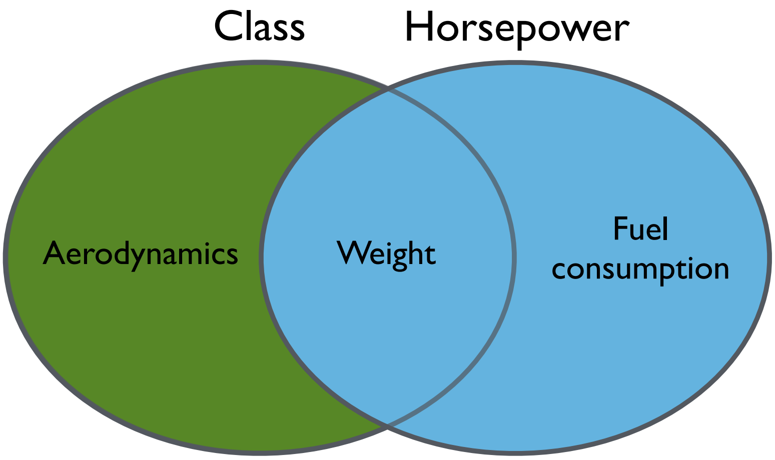

Figure 13.7: Our model for gas mileage includes two variables: engine horsepower and vehicle class. These variables both convey information about a vehicle's size, in addition to some unique information (e.g. class tells us about aerodynamics, while horsepower tells us about fuel consumption). When we add the Horsepower variable first in an analysis of variance (Table 13.5), 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 13.6), this picture changes. 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 13.8.

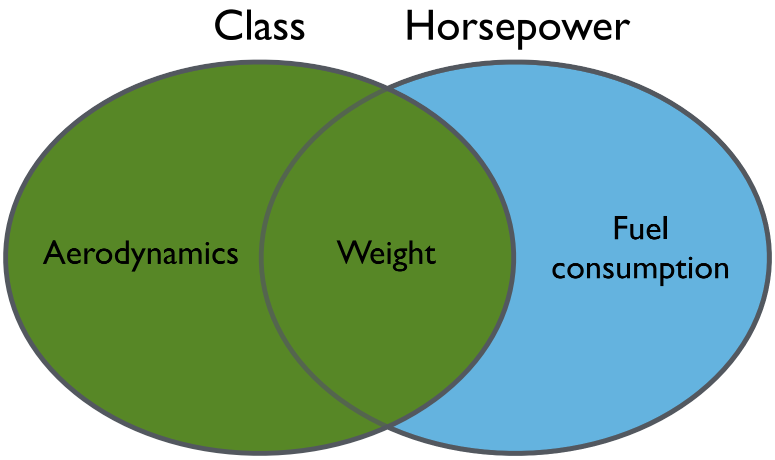

Figure 13.8: (Continued from Figure 13.7.) But when we add the Class variable first in an analysis of variance (Table 13.5), we attribute all of the shared information content to Class, and none to Horsepower, in our ANOVA table.

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 cloves and cinnamon give a little different flavor to what you get from a pumpkin pie. But 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 model 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 ΔPV for each one. Thus in the absence of dependence among the predictors, we can uniquely assign credit for predicting the outcome to each one.[4]

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

[4] 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.