9
Dummy variables

Models for a single grouping variable

Dummy variables

Let's consider 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 9.1 shows the weekly sales volume of package sliced cheese over 61 weeks at a Dallas-area 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.

How much higher? The average sales volume in display weeks was 5,577 units (the blue dotted line in Figure 9.1), 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 9.1 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

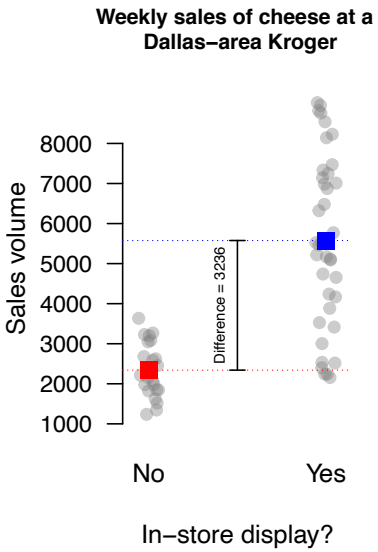

Figure 9.1: Weekly sales of packaged cheese slices at a Dallas-area Kroger's grocery store, both with and without the presence of an in-store display ad for the cheese. The red dot shows the mean of the no-display weeks, and the blue dot shows the mean of the with-display weeks. The estimated coefficient for the dummy variable that encodes the presence of a display ad is 3236, which is the vertical distance between the two dots.

linear model, we call β_0 and β_1 the *coefficients* of the model. This way of expressing the model implies the following.

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

Therefore, we can think of β_0 as the baseline (or *intercept*), and β_1 as the offset. To see this in action, consult Figure 9.1 again. Here the dummy variable encodes the presence of an in-store display. The red dot at 2341, in the non-display weeks, is β_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 β_0 and β_1 using the least-squares criterion: that is, make the sum of squared errors, $\sum_{i=1}^{n} e_i^2$, as small as possible. This is mathematically equivalent to computing the group-wise means separately, and then calculating the difference between the means.

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 ith case in the data set.[1]

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:

$$y_i = \beta_0 + \sum_{k=1}^{K-1} \beta_1^{(k)} \mathbf{1}_{\{x_i=k\}} + e_i \tag{9.1}$$

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 β_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

[1] 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 group-wise models a lot cleaner compared to starting at 1.

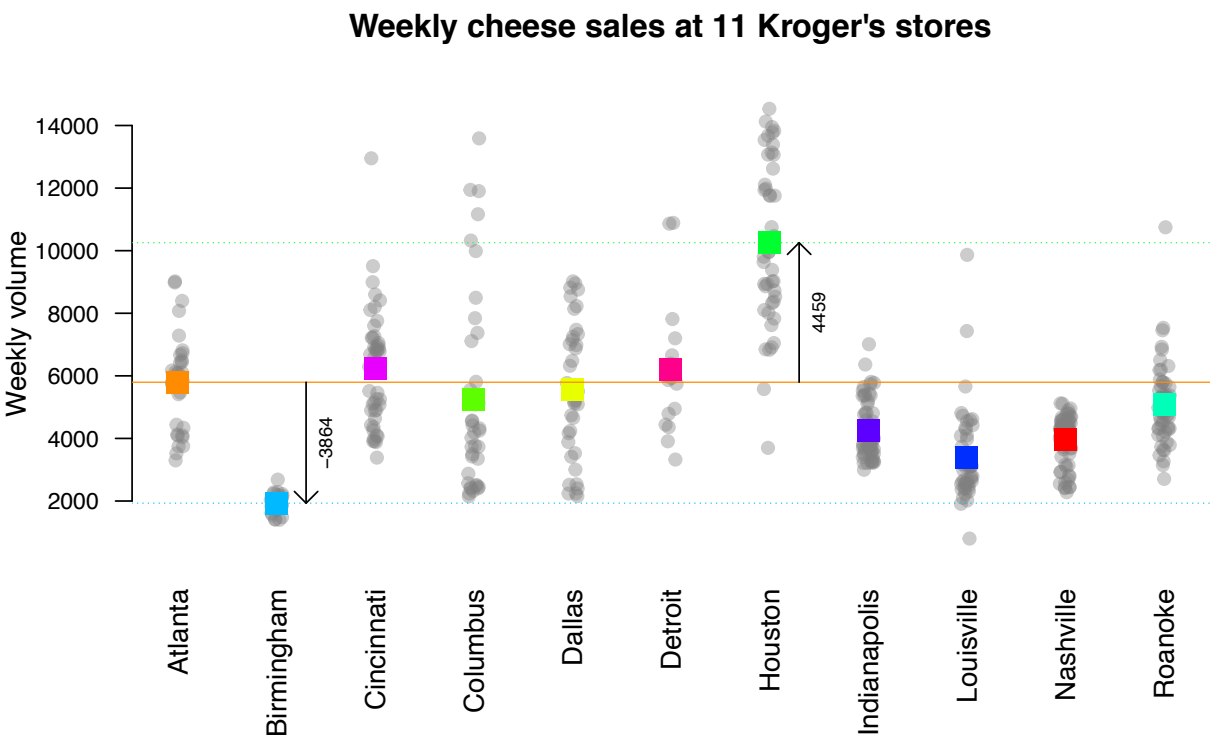

Figure 9.2: Weekly sales of packaged cheese slices during weeks with an advertising display at 11 Kroger's grocery stores across the country.

observation. The coefficient on each dummy variable ($\beta_1^{(k)}$) is the differences between the baseline and the mean of group k:

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

In Figure 9.2, 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 9.1 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:

Atlanta is the baseline, and so the intercept is the group mean for Atlanta: 5796 packages of crappy cheese. To get the group

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

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 $\beta_1^{(6)} = 4459$, 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 9.1 to the Kroger cheese-sales data, now choosing the Dallas store to be the baseline:

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.

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

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.