Lecture 19: Interactions

Let

$$m(x) = \mathbb{E}[Y|X = x]$$

where $x = (x_1, \ldots, x_p)$. We say that there is no interaction between X_j and X_k if

$$\frac{\partial m(x)}{\partial x_i}$$

does not depend on x_j.

Consider the linear model

$$m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Then $\frac{\partial m(x)}{\partial x_1} = \beta_1$ and $\frac{\partial m(x)}{\partial x_2} = \beta_2$. There are no interactions.

Now suppose that

$$m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

Then $\frac{\partial m(x)}{\partial x_1} = \beta_1 + \beta_3 x_2$ and $\frac{\partial m(x)}{\partial x_2} = \beta_2 + \beta_3 x_1$. So we say there is an interaction between x_1 and x_2.

If your model does not fit well, then adding interactions is yet another way to improve the fit of the model. You could plot the residuals versus $X_1 X_2$ or just as the interaction to the model.

1 The Conventional Form of Interactions in Linear Models

The usual way of including interactions in a linear model is to add a product term, as, e.g.,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon. \tag{1}$$

Once we add such a term, we estimate β_3 in exactly the same way we'd estimate any other coefficient.

People often call β_1 and β_2 the main effects and they call β_3 the interaction effect. This is not the greatest terminology but it is pretty standard. Usually people don't add interactions into a model without adding the main effects. So it's rare to see a model of the form $Y = \beta_0 + \beta_3 X_1 X_2 + \epsilon$. Adding in the main effects gives a model with more flexibility and generality.

2 Interaction of Categorical and Numerical Variables

If we multiply the indicator variable for a binary category, say X_B, with an ordinary numerical variable, say X_1, we get a different slope on X_i for each category:

$$Y = \beta_0 + \beta_1 X_1 + \beta_{1B} X_B X_1 + \epsilon. \tag{2}$$

When $X_B = 0$, the slope on X_1 is β_1, but when $X_B = 1$, the slope on X_1 is $\beta_1 + \beta_{1B}$; the coefficient for the interaction is the *difference* in slopes between the two categories.

In fact, look closely at Eq. 2. It says that the categories share a common *intercept*, but their regression lines are not parallel (unless $\beta_{1B} = 0$). We could expand the model by letting each category have its own slope and its own intercept:

$$Y = \beta_0 + \beta_B X_B + \beta_1 X_1 + \beta_{1B} X_B X_1 + \epsilon.$$

This model is similar to running two separate regressions, one per category. It does, however, insist on having a single noise variance σ^2 (which separate regressions wouldn't accomplish). Also, if there were additional predictors in the model which were not interacted with the category, e.g.,

$$Y = \beta_0 + \beta_B X_B + \beta_1 X_1 + \beta_{1B} X_B X_1 + \beta_2 X_2 + \epsilon$$

then this would definitely not be the same as running two separate regressions. We can also add categoricla variables and interactions with categorical variables. Just remember that a categorical variable with k levels requires adding only $k-1$ indicator variables.

2.1 Interactions of Categorical Variables with Each Other

Suppose we have two binary categorical variables, with corresponding indicator variables X_B and X_C. If we fit a model of the form

$$Y = \beta_0 + \beta_1 X_B + \beta_2 X_C + \beta_3 X_B X_C + \epsilon$$

then we can make the following identifications:

$$\begin{aligned}
\mathbb{E}\left[Y|X_B = 0, X_C = 0\right] &= \beta_0 && (3)\\
\mathbb{E}\left[Y|X_B = 1, X_C = 0\right] &= \beta_0 + \beta_1 && (4)\\
\mathbb{E}\left[Y|X_B = 0, X_C = 1\right] &= \beta_0 + \beta_2 && (5)\\
\mathbb{E}\left[Y|X_B = 1, X_C = 1\right] &= \beta_0 + \beta_1 + \beta_2 + \beta_3 && (6)
\end{aligned}$$

Conversely, these give us four equations in four unknowns, so if we know the group or conditional means on the left-hand sides, we could solve these equations for the β's.

3 Higher-Order Interactions

Nothing stops us from considering interactions among three or more variables, rather than just two. For example

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \epsilon.$$

As you can see, these models get complicated very quickly. Also, we have to ask ourselves: which interactions should I add? For example, I could have added $X_1^2 X_2$ into the model as well as other terms. We are now entering the realm of model-building and model-selection that we will discuss in a future lecture. For now, we will try to keep our models fairly simple.

4 Interactions in R

The `lm` function is set up to comprehend multiplicative or product interactions in model formulas. Pure product interactions are denoted by `:`, so the formula

```
lm(y ~  x1:x2)
```

tells R to fit the model $Y = \beta_0 + \beta X_1 X_2 + \epsilon$. (Intercepts are included by default in R.) Since it is relatively rare to include just a product term without linear terms, it's more common to use the symbol `*`, which expands out to both sets of terms. That is,

```
lm(y ~ x1*x2)
```

fits the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

This special use of `*` in formulas over-rides its ordinary sense of multiplication; if you wanted to specify a regression on, say $1000 X_2$, you'd have to write `I(1000*x2)` rather than `1000*x2`. Note that `x1:x1` is the same as `x1`; if you want higher powers of a variable, use `I(x1^2)` or `poly(x1,2)`.

The `:` symbol applies will apply to combinations of variables. Thus

```
(x1+x2):(x3+x4)
```

is the same as

```
x1:x3 + x1:x4 + x2:x3 + x2:x4
```

Also,

```
(x1+x2)*(x3+x4)
```

is the same as

```
x1 + x2 + x3 + x4 + x1:x3 + x1:x4 + x2:x3 + x2:x4
```

The reason you can't just write `x1^2` in your model formula is that the power operator *also* has a special meaning in formulas, of repeatedly `*`-ing its argument with itself. That is,

```
(x1+x2+x3)^2
```

is the same as

```
(x1+x2+x3)(x1+x2+x3)
```

which is

```
x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3
```

`poly` **and interactions.** If you want to use `poly` to do polynomial regression and interactions, do this:

```
lm(y ~ poly(x1,x2,degree=2)
```

which fits the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \epsilon.$$

4.1 Example

Let's continue with the mobility data. First, here is a useful trick:

```
x = c("a","b","c","d","e","f")
y = c("a","b")
z =  x %in% y
print(z)
## [1]  TRUE  TRUE FALSE FALSE FALSE FALSE
```

The command

```
%in%
```

is a matching operator.

Let's use this to create a binary variable indicating whether a state was or was not part of the Confederacy in the Civil War.

```
Confederacy = c("AR","AL","FL","GA","LA","MS","NC","SC","TN","TX","VA")
mobility$Dixie = mobility$State %in% Confederacy
out = lm(Mobility ~ Commute*Dixie,data=mobility)
summary(out)
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.01880    0.00683  2.7600 5.95e-03
## Commute             0.19500    0.01340 14.5000 2.93e-42
## DixieTRUE          -0.02120    0.01190 -1.7700 7.64e-02
## Commute:DixieTRUE -0.00131    0.02830 -0.0461 9.63e-01
```

The coefficient for the interaction is negative, suggesting that increasing the fraction of workers with short commutes predicts a smaller difference in rates of mobility in the South than it does in the rest of the country. This coefficient is not significantly different from zero, but, more importantly, we can be confident it is small, compared to the base-line value of the slope on `Commute`:

```
confint(out)
##                       2.5 %  97.5 %
## (Intercept)         0.00543 0.03220
## Commute             0.16900 0.22200
## DixieTRUE          -0.04470 0.00225
## Commute:DixieTRUE -0.05680 0.05420
```

Thus, even if the South does have a different slope than the rest of the country, it is not a very different slope.

The difference in the intercept, however, is more substantial. It, too, is not significant at the 5% level, but that is because (as we see from the confidence interval) it might be quite large and negative or perhaps just barely positive — it's not so precisely measured, but it's either lowering the expected rate of mobility or adding to it trivially. Of course, we should really do all our diagnostics here before paying much attention to these inferential statistics.