Lecture 19: Interactions

36-401, Fall 2015, Section B

3 November 2015

Contents

When we say that there are no interactions between X_i and X_j, we mean that

$$\frac{\partial \mathbb{E}[Y|X=x]}{\partial dx_i}$$

is not a function of x_i. Said another way, there are no interactions if and only if

$$\mathbb{E}[Y|X=x] = \alpha + \sum_{i=1}^{p} f_i(x_i)$$

so that each coordinate of X makes its own separate, additive contribution to Y. The standard multiple linear regression model of course includes no interactions between any of the predictor variables.

General considerations of probability theory, mathematical modeling, statistical theory, etc., give us no reason whatsoever to anticipate that interactions are rare, or that when they exist they are small. You might be so lucky as to not have any to deal with, but you should not *presume* you will be lucky.

Diagnosing the presence of interactions See Lecture 15 for some ideas about how to do this. One trick not mentioned there is to plot the residuals from an interaction-free model against the product of two predictors, e.g., against $X_1 X_2$. This, however, presumes a particular form for the interaction, gone over in the next section.

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.

Interpretation In the model of Eq. 1, it is no longer correct to interpret β_1 as $\mathbb{E}[Y|X_1 = x_1 + 1, X_2 = x_2] - \mathbb{E}[Y|X_1 = x_1, X_2 = x_2]$. That difference is, rather $\beta_1 + \beta_3 X_2$. Similarly, β_2 is no longer the expected difference in Y between two otherwise-identical cases where X_2 differs by 1. The fact that we can't give one answer to "how much does the response change when we change this variable?", that the correct answer to that question always involves the other variable, is what interaction *means*.

What we can say is that β_1 is the slope with regard to X_1 when $X_2 = 0$, and likewise β_2 is how much we expect Y to change for a one-unit change in X_2 when $X_1 = 0$. β_3 is the rate at which the slope on X_1 changes as X_2 changes, and likewise the rate at which the slope on X_2 changes with X_1 (see Exercise 1 for why it's both).

Diagnostics and inference Diagnostics for a product term goes just like it would for any other: the residuals should have the same distribution no matter what the value of $X_i X_j$ happens to be; all the usual plots can be made using $X_i X_j$ as the predictor variable. Inference, too, works exactly the same way.

Terminology The coefficients which go with the linear terms, β_1 and β_2 above, are often called the "main effects", while β_3 would be an "interaction effect". I think this terminology is misleading in at least two ways. First, by talking about "effects" at all, it carries causal implications which are not usually warranted by a regression. Second, it implies that the linear terms, being "main", are bigger or more important than the interactions, and again there's usually no reason to think that. Why we don't use names like "linear coefficients" and "product coefficients", I couldn't say.

Products without linear terms considered dubious It is very rare to find models where there is a product term $X_i X_j$ without both the linear terms X_i and X_j. If, say, the X_i term was missing, it would mean that Y was completely insensitive to X_i when $X_j = 0$, but only then. This is weird, and indeed flies in the face of one of the best justifications for using product interactions (§1.1). There's no intrinsic reason it couldn't happen, but you should expect models like that to receive additional scrutiny.

1.1 Why Product Interactions?

Most texts on linear regression do not even attempt to justify using interaction terms that look like $X_1 X_2$, as opposed to $\frac{X_1 X_2}{1+|X_1 X_2|}$, or $X_1 H(X_2 - c)$, etc., etc. Here is the best justification I can find.

H is the Heaviside step function, $H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$.

Suppose that the real regression function $\mathbb{E}[Y|X = x] = \mu(x)$ is a smooth function of all the coordinates of x. Because it is smooth, we should be able to do a Taylor expansion around any particular point, say x^*:

$$\mu(x) \approx \mu(x^*) + \sum_{i=1}^{p} (x_i - x_i^*) \left.\frac{\partial \mu}{\partial x_i}\right|_{x=x^*} + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (x_i - x_i^*)(x_j - x_j^*) \left.\frac{\partial^2 \mu}{\partial x_i x_j}\right|_{x=x^*}$$

The first term, $\mu(x^*)$, is a constant. The next sum will give us linear terms in all the x_i (plus more constants). The double sum after that will give us terms for each product $x_i x_j$, plus all the squares x_i^2, plus more constants. Thus, if the true regression function is smooth, and we only see a small range of values for each predictor variable, using product terms is reasonable — provided we also include quadratic terms for each variable. (See Lecture 16 on polynomial regression for how to do that.)

Non-product interactions If have a particular sort of non-product interaction term in mind, say $\frac{X_1 X_2}{1+|X_1 X_2|}$, there is not particular difficulty in estimating

it; just form a new column of predictors with the appropriate values, and estimate a coefficient on it like any other. Interpretation may, however, become even more tricky, and there is also the issue of deciding on what sort of interaction. In 402, we will see ways of discovering reasonable interaction terms automatically, by two-dimensional smoothing.

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. This is just like the way the coefficients on categorical variables back in Lecture 16 ("adjusted effects") were differences between the intercepts for the 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, where "everything is interacted with the category", is *very* close to just 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). It also let you form confidence intervals for β_B and β_{1B}; if one or the other of these is tightly focused around 0, you might consider dropping that term and re-estimating[1]. 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.

As with linear terms for categorical variables ("adjusted effects"), everything works much the same for variables with more than two levels: we add one indicator variable for all but one (reference or baseline) level of the category, we interact the indicators with the other predictor or predictors of interest, and the coefficients are differences to the slopes.

[1] You *could* get the same effect with two separate regressions, by getting a confidence interval for the difference in the two estimates of the slope or the two estimates of the intercept, but the answer would come to the same as what you'd get from the joint regression with full interactions.

2.1 Interactions of Categorical Variables with Each Other

Nothing stops the variable you interact a categorical with from being another categorical. When that happens, you get terms which only apply to individuals which belong to *both* categories, e.g., to plumbers in Ohio.

Categorical interactions vs. group or conditional means 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:

$$\mathbb{E}[Y|X_B = 0, X_C = 0] = \beta_0 \quad (3)$$
$$\mathbb{E}[Y|X_B = 1, X_C = 0] = \beta_0 + \beta_1 \quad (4)$$
$$\mathbb{E}[Y|X_B = 0, X_C = 1] = \beta_0 + \beta_2 \quad (5)$$
$$\mathbb{E}[Y|X_B = 1, X_C = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3 \quad (6)$$

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 βs (Exercise 2).

Notice that if our only predictor variables were these two categorical variables, we'd have one parameter for each distinct value of X — the model is **saturated** — and we'd have very little ability to tell that the model was wrong, regardless of how big n might be. One way we might check it would be to look at the distribution of residuals for each distinct group — by assumption they should all be the same. Of course if we have additional predictor variables, we can check the residuals against them.

3 Higher-Order Interactions

Nothing stops us from considering interactions among three or more variables, rather than just two. Again, the conventional form for this is a product, $X_i X_j X_k$. Again, the best justification for this I've ever seen is a higher-order Taylor expansion, which suggests using terms like $X_i^2 X_j$ and X_i^3 as well. Again, there is nothing special about diagnostics or inference for higher-order interaction terms. Trying to describe their interpretation in words gets extra tricky, however.

4 Product 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)
```

corresponds to 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)
```

is equivalent to

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

and both estimate the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$. This special sense 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`. Also notice that R thinks, not unreasonably, that `x1:x1` is just the same as `x1`; if you want higher powers of a variable, use `I(x1^2)` or `poly(x1,2)`.

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

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

is equivalent to

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

Similarly for `*`. This

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

expands out to this:

```
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, this

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

is equivalent to

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

which in turn is equivalent to

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

(Remember that `x1:x1` is just `x1`.)

I find these operators in formulas most useful when I want to interact lots of variables with a category:

```
lm(y ~ (x1+x2+x3+x5)*xcat + x4)
```

is a lot more compact than writing everything out, as

```
lm(y ~ xcat + x1 + x2 + x3 + x5 + x1:xcat + x2:xcat + x3:xcat + x5:xcat + x4)
```

and it's also something I'm a lot less likely to get wrong. Even writing out the whole formula term by term would be a lot less work, and lead to many fewer errors, than creating all the interacted columns by hand.

`poly` and interactions If you want to use `poly` to do polynomial regression, as in Lecture 16, *and* we want interactions, we can do it:

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

This creates linear terms for both variables (which it gives names ending `1.0` and `0.1`), quadratic terms for both variables (names ending in `2.0` and `0.2`), and their product term (whose name ends in `1.1`). We have to explicitly name the `degree` argument; otherwise, `poly` doesn't know when we've stopped giving it columns we want to interact. If we set `degree` higher than 2, we'll get interactions between powers of the variables, and if we gave it $k > 2$ variables, we'd get all possible $2, 3, \ldots k$-way interactions.

4.1 Economic Mobility vs. Commuting, Again

Let's continue with the data from the first DAP.

```
mobility <- read.csv("http://www.stat.cmu.edu/~cshalizi/mreg/15/dap/1/mobility.csv")
```

As in Lecture 16, on categorical variables, we'll introduce a new binary category, indicating whether each state was or was not a part of the Confederacy in the Civil War. (See that lecture for detailed comments.)

```
# The states of the Confederacy
Confederacy <- c("AR", "AL", "FL", "GA", "LA", "MS", "NC", "SC", "TN", "TX", "VA")
mobility$Dixie <- mobility$State %in% Confederacy
```

In that lecture, we allowed this new indicator variable to change the intercept; you will recall that that term was negative and highly significant. Here, we'll let being in the South affect the slope on `Commute` as well, that is, we introduce an interaction between `Commute` and `Dixie`:

```
mob.dixie <- lm(Mobility ~ Commute*Dixie, data=mobility)
signif(coefficients(summary(mob.dixie)),3)
```

```
##                   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
```

(See also Exercise 3.)

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`:

```
signif(confint(mob.dixie),3)
```

```
##                     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 (-2 percentage points, when the mean is about 10% and the largest value is 47%), 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, but I offer this by way of illustration of the functions. As a further illustraiton, see Exercise 4.

5 Exercises

To think through or to practice on, not to hand in.

1. Consider an apparently different, and perhaps more-interpretable, model than Eq. 1, namely

$$Y = \alpha_0 + (\alpha_1 + \alpha_2 X_2) X_1 + (\alpha_3 + \alpha_4 X_1) X_2 + \epsilon$$

   Show that this can always be re-written in the same form as Eq. 1, and express the latter's $\beta_0, \beta_1, \beta_2$ in terms of the αs of this model. Can models of the form of Eq. 1 always be re-written in this form? If so, express the α parameters in terms of the βs; if not, give a counter-example.

2. Solve Eqs. 3–6 for the βs.
3. Check that we get the same set of terms, with the same coefficients, as in §4.1, if we fit our model with

```
lm(Mobility ~ Commute+Dixie+Commute:Dixie, data=mobility)
```

   Why does this happen?

4. Using the mobility data, regress `Mobility` on
   (a) Latitude and longitude (only);
   (b) Latitude, longitude, and their product (only);
   (c) Latitude, longitude, their product, and their squares (only).

   For each model, make maps[2] of the fitted values and the residuals. Describe the resulting geographic patterns, and compare them (qualitatively) to a map of the actual values of `Mobility`. Can you explain why the maps of fitted values look like they do, based on the terms included in the model?

[2] See the hint on the DAP 1 assignment for help with making such maps.