Lecture 8 Categorical predictors, polynomials, interactions

Binary predictors

So far we have been dealing with predictors that we might call "continuous." Variables like GDP, country size, education, ability are examples of continuous predictors. Many predictors, however, are *binary* in nature: did a person get a treatment or not, is a person a man or a woman, is a person married or not, did a person go to college or not. Binary variables take a central part in causal analysis since treatment variables are often binary, i.e., did a person get a treatment or not.

These binary variables can be included in a regression just like continuous variables, however, they only take two values: 0 or 1. Sometimes we call these variables *indicator* or *dummy* variables. For example, if our variable is whether someone went to college or not, then values of 0 would correspond to people who did not go to college, and values of 1 would correspond to people who did go to college. Suppose our outcome is a person's wage, then our regression would look like

$$\text{Wage} = \beta_0 + \beta_1 \text{Went to college} + U.$$

How do we interpret the slope coefficient β_1 for a binary variable? Let's follow the usual procedure. Start with the CEF of Y:

$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X.$$

Since X takes only two possible values, 0 or 1, we can write down the CEF of Y for each possible value of X

$$\begin{aligned}\mathbb{E}[Y \mid X = 0] &= \beta_0 \\ \mathbb{E}[Y \mid X = 1] &= \beta_0 + \beta_1.\end{aligned}$$

Take the difference to get

$$\mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0] = \beta_1.$$

The slope coefficient β_1 is simply the difference between the conditional means of the outcome when the predictor is "on" ($X = 1$) and "off" ($X = 0$).

In the college example, the slope coefficient would tell us by how much the wage of a person who went to college is different from the wage of a person who did not go to college. The calculations above also tell us the meaning of the intercept term: it is the conditional mean of the outcome when the predictor is "off" ($X = 0$). In this example, for people who did not go to college.

When we work with binary predictors, we only include one variable for each binary predictor, not two. For example, we do not (and cannot) estimate a model like

$$\text{Wage} = \beta_0 + \beta_1 \text{Went to college} + \beta_2 \text{Did not go to college} + U,$$

where Went to college would equal 1 if a person went to college and 0 if they did not, and $\text{Did not go to college}$ would equal 1 if a person did not go to college and 0 if they did. Having those two variables together would violate the assumption of *no perfect collinearity* because $\text{Went to college} + \text{Did not go to college} = 1$. And 1 is our constant variable with the associated coefficient β_0.

Suppose that the mean wage of a person who went to college is 20 euros an hour and the mean wage of a person who did not go to college is 15 euros an hour. Then if include both variables in the model we could get those predictions with $\beta_0 = 15, \beta_1 = 5, \beta_2 = 0$. Or we could get those predictions with $\beta_0 = 20, \beta_1 = 0, \beta_2 = -5$. Or in infinitely many other different ways. The OLS would not be able to give us a unique result.

In other words, when we have a binary predictor we always exclude one category from the model, and this category becomes the *reference* category.

Categorical predictors

A logical extension of a binary variable is a variable that has more than two categories. For example, a person's highest level of education could be "high school", "college", "graduate degree". A person's race or ethnicity could be "black", "white", "hispanic". A person's marital status could be "single", "married", "divorced", "in a relationship". These variables are called *categorical* variables.

The way to include these categorical variables in a regression is to give each category of a variable its own binary variable. Instead of having a variable "highest level of eduction" we will have several binary variables of a kind "whether a person's highest level of education is high school or not", "whether a person's highest level of education is college or not", "whether a person's highest level of education is graduate degree or not", etc. However, just like for binary variables, we always have to exclude one category that would become a *reference* category. All the coefficients on the binary variables associated with a given categorical predictor then will be relative to that reference category.

For example, suppose our outcome is a person's wage and our predictor is the highest level of education. The predictor can take three possible values: "high school", "college", "graduate degree." Our regression model would look like this

$$\text{Wage} = \beta_0 + \beta_1 \text{College} + \beta_2 \text{Graduate} + U.$$

In this model, we exclude the "high school" category, which becomes the reference category. To interpret the coefficient, let's consider the conditional mean of the outcome for each possible level of the predictor.

$$\begin{aligned}
\mathbb{E}[\text{Wage} \mid \text{College} = 0, \text{Graduate} = 0] &= \beta_0 \\
\mathbb{E}[\text{Wage} \mid \text{College} = 1, \text{Graduate} = 0] &= \beta_0 + \beta_1 \\
\mathbb{E}[\text{Wage} \mid \text{College} = 0, \text{Graduate} = 1] &= \beta_0 + \beta_2.
\end{aligned}$$

Then the intercept term is the mean wage of a person with only a high school degree. Taking the difference between the second and first expression, we get

$$\mathbb{E}[\text{Wage} \mid \text{College} = 1, \text{Graduate} = 0] - \mathbb{E}[\text{Wage} \mid \text{College} = 0, \text{Graduate} = 0] = \beta_1.$$

Therefore, β_1 is the difference between mean wages of a person who went to college (but not to graduate school) and a person with only a high school degree.

Taking the difference between the third and first expression, we get

$$\mathbb{E}[\text{Wage} \mid \text{College} = 0, \text{Graduate} = 1] - \mathbb{E}[\text{Wage} \mid \text{College} = 0, \text{Graduate} = 0] = \beta_2.$$

Therefore, β_2 is the difference between mean wages of a person who went to a graduate school and a person with only a high school degree.

A reference category can be anything. You can pick whatever category makes sense for your research question. But then all of the slope coefficients will be interpreted relative to that reference category. For example, if in our example we make "college" the reference category, all our coefficients will change and we will interpret them relative to college.

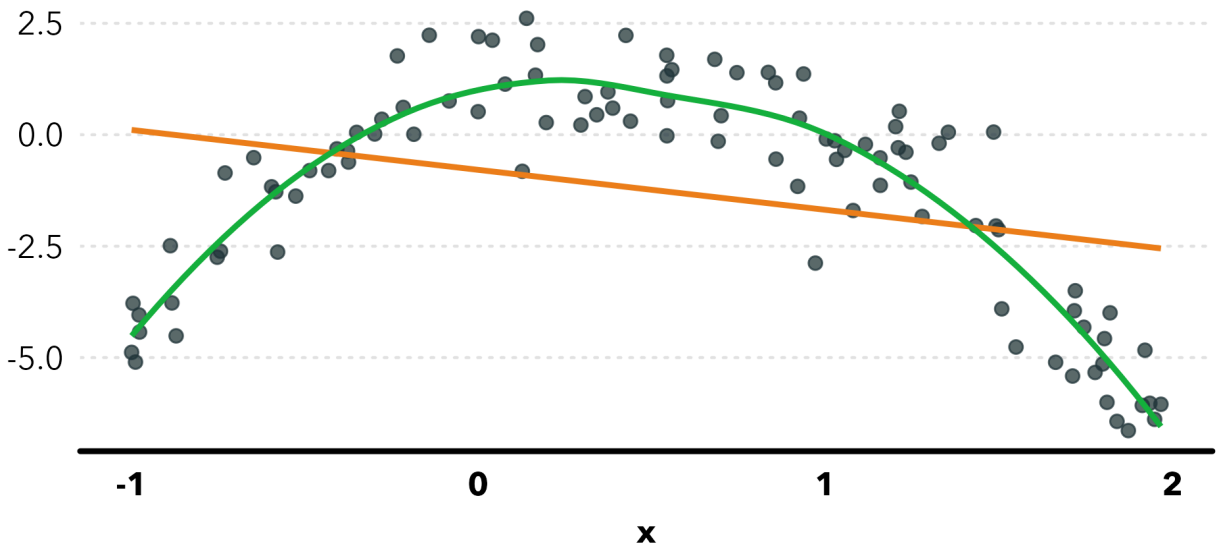
Polynomials

We have seen that by using the logarithmic transformation we can model non-linear relationships between variables. An even more flexible approach is to use *polynomials*. An example of a regression that includes polynomials is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + U.$$

This regression includes the third degree polynomial of X. Each of the three terms is usually referred to as a linear term (X), a quadratic term (X^2), and a cubic term (X^3). When using polynomials of a given degree, we include *all* of the terms up to that degree, without skipping. In economics we usually use either second- or third-degree polynomials, but not higher.

Consider the graph below. The two lines show a linear fit and a quadratic fit. Clearly, a quadratic model fits the data better.



Polynomials allow us to model very flexible relationships. However, the coefficients become trickier to interpret. Recall that previously the interpretation of each individual slope coefficients in a multiple regression was as a *partial* effect. That is, the effect on the outcome of changing one variable while keeping the rest constant. In case of polynomials, this interpretation is no longer possible. If we change X we cannot keep X^2 constant. We have to use another method.

Recall that an alternative derivation of the slope coefficient was to take a partial derivative of the conditional mean of an outcome with respect to the corresponding predictor. We can use this logic with polynomials. For example, for our third-degree polynomial the conditional mean of Y is

$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

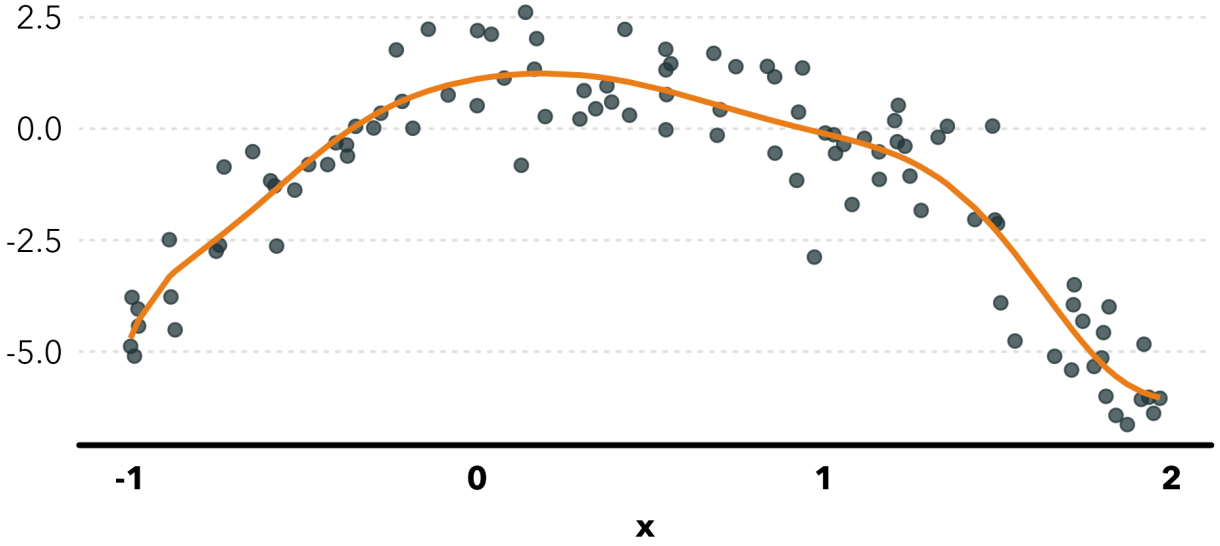
and the derivative of the conditional mean of Y with respect to X is

$$\frac{d\mathbb{E}[Y \mid X]}{dX} = \beta_1 + 2\beta_2 X + 3\beta_3 X^2.$$

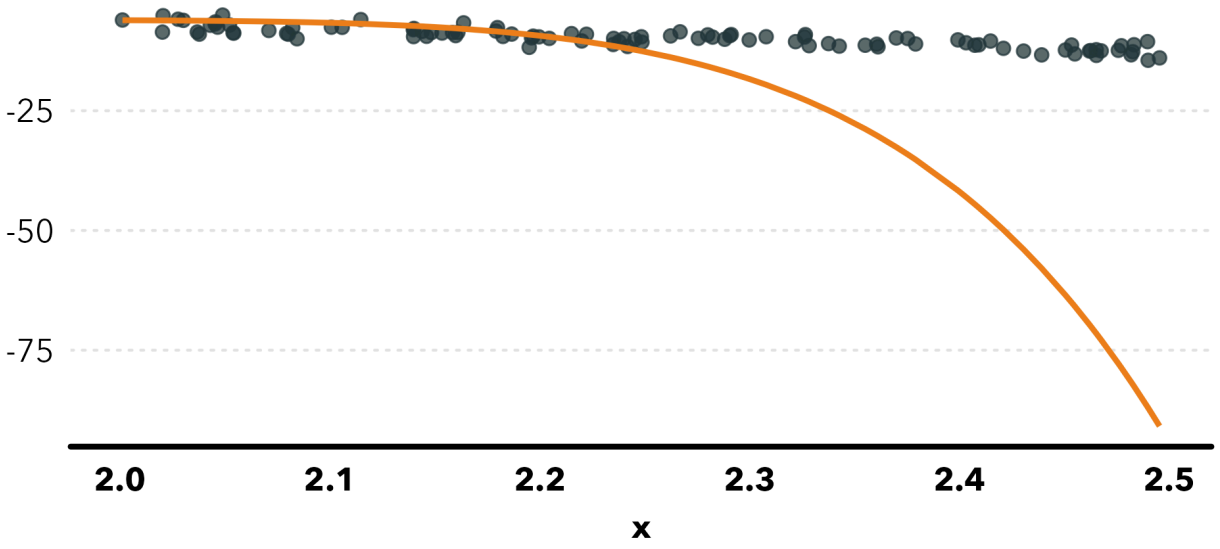
Now the effect of X on the mean outcome depends on the value of X, it is not constant. For example, when $X = 0$, increasing X by a small amount changes the conditional mean of Y by β_1 units. However, when $X = 1$, increasing X by a small amount changes the conditional mean of Y by $\beta_1 + 2\beta_2 + 3\beta_3$ units. Each individual coefficient has little meaning on its own, they have to be interpreted together.

How can we determine the order of polynomials? Why do we have to stop at the cubic term? Why not take higher terms? There are a couple of reasons why in economics we typically stop at the quadratic term or at the cubic term. First, a model with too many polynomial terms becomes *harder to interpret*. The second-order polynomial can have a nice interpretation of a decreasing marginal effect of a predictor, for example, when the linear term has a positive coefficient but a quadratic term has a negative coefficient. But adding a cubic term makes the results much harder to interpret.

Second, even if we do not care too much about interpretation having too many polynomial terms can lead to *overfitting*. We would be adding irrelevant predictors. In the picture below we fit a 10th degree polynomial when the true relationship is quadratic. All those extra polynomial terms are completely irrelevant but it does look like they capture some patterns in the data.



If we try to use this overly complex model on a new data set generated using the same underlying process where the true relationship is quadratic, we will get some bizarre predictions.



Interaction terms

Polynomials let the effect of a predictor depend on the value of that predictor. Interaction terms, on the other hand, let the effect of a predictor depend on the value of *another* predictor. For example, let's say we are interested in the effect of years of schooling on wage. Does this effect depend on a person's gender or race? Modeling an interaction effect would allow us to answer this question. An example of a model with an interaction term is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + U.$$

Here we have two main effects, X and Z, and an interaction term XZ, which is simply a product of the two variables. When we work with interactions, we always include the main

effects in addition to the interaction terms. Notice that polynomial terms can be thought of as interacting a variable with itself: $X^2 = X \times X$.

Including an interaction term in a model complicates the interpretation of individual coefficients. There are two related questions we have to answer. First, what is the effect of X when the model includes an interaction term. And second, what is the interpretation of the interaction term itself.

As was the case with polynomials, we can no longer interpret β_1 as the partial effect of X because changing X also changes XZ. Let's use the same trick by finding the partial derivative of the conditional mean of Y with respect to X.

$$\frac{\partial \mathbb{E}[Y \mid X, Z]}{\partial X} = \beta_1 + \beta_3 Z.$$

Now the partial effect of X depends not only on β_1 but also on the value of Z. For example, when $Z = 0$, the partial effect of X on the conditional mean of Y is β_1, but when $Z = 1$ the partial effect of X on the conditional mean of Y is $\beta_1 + \beta_3$, and so on. Therefore, individual slope coefficients have little meaning by themselves, we have to interpret them together with other coefficients.

What is then the interpretation of the interaction term? One way to deal with it is to take the cross-partial derivative of the conditional mean of Y to get

$$\frac{\partial^2 \mathbb{E}[Y \mid X, Z]}{\partial X \partial Z} = \beta_3.$$

How can we interpret this value? Recall that the partial effect of X on the conditional mean of Y is

$$\frac{\partial \mathbb{E}[Y \mid X, Z]}{\partial X} = \beta_1 + \beta_3 Z.$$

Then we can interpret β_3 as the change in the partial effect of X when Z increases by one unit. The interaction effect is *symmetric*, i.e., it also tells us by how much the partial effect of Z changes when X increases by one unit.

What if one of the variables is binary or categorical? This would be the case in our opening examples of the effect of schooling on wage. Suppose we want to let this effect vary by gender. Our model would look like

$$\mathrm{Wage} = \beta_0 + \beta_1 \mathrm{Education} + \beta_2 \mathrm{Male} + \beta_3 \mathrm{Education} \times \mathrm{Male} + U,$$

where $\mathrm{Education}$ is years of education and Male equals 1 if a person is male and 0 if not.

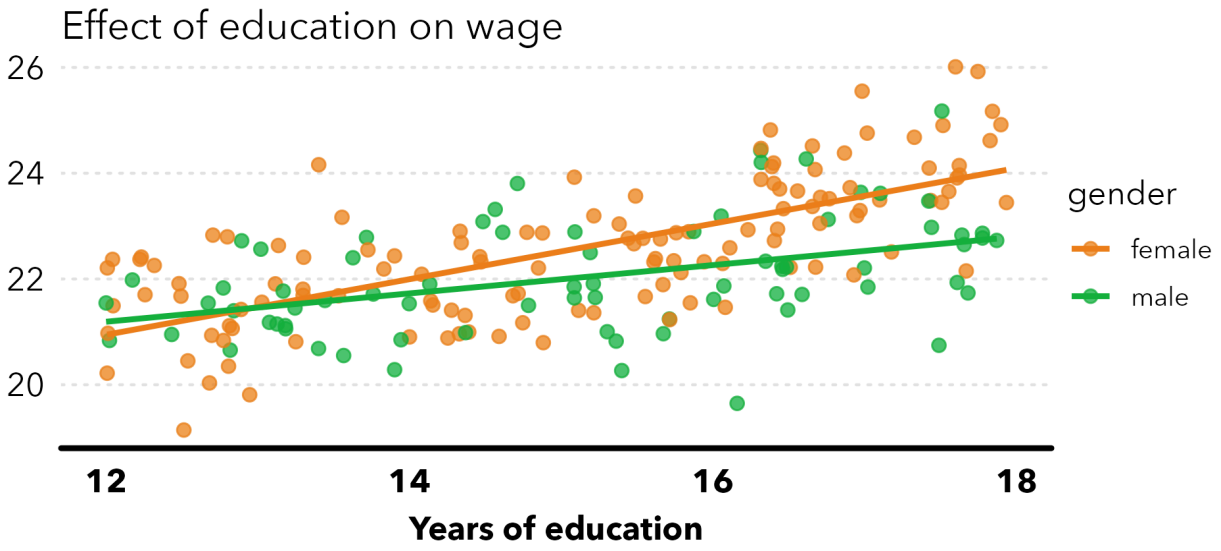
Then the partial effect of education on wage is

$$\frac{\partial \mathbb{E}[\mathrm{Wage} \mid \mathrm{Education}, \mathrm{Male}]}{\partial \mathrm{Education}} = \beta_1 + \beta_3 \mathrm{Male}.$$

Then for males, the effect of schooling on wage is $\beta_1 + \beta_3$ while for females the effect of schooling on wage is β_1. And β_3 is the difference between the effect of schooling on wage

for males vs. females.

A graph below shows an example of such a model using a simulated dataset. Allowing for an interaction term essentially allows us to have two different linear fits of wage on education: one for males and one for females.



Interactions are not limited to two-way interactions, i.e., products between only two variables. You can have three-way interactions and higher, too. E.g.,

$$Y = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 Z + \beta_4 XY + \beta_5 XZ + \beta_5 YZ + \beta_6 XYZ.$$

If we include, say, a three-way interaction term in a model (XYZ above) we typically include all the lower interaction terms, as well as the main effects.

Interactions are a powerful modeling tool. However, it is difficult to estimate the interaction effects in practice. You need a lot of observations to precisely estimate them.