Module 6: Categorical Predictors

Jeffrey Woo

MSDS, University of Virginia

Dummy Coding with Categorical Predictors

Dummy coding uses indicator variables for categorical variables. If the categorical variable has c classes, there will be $c - 1$ indicator variables.

- One class will be 0 for all indicator variables. This class is the **reference class**.
- The coefficients for indicator variables compare the mean response for that class with the reference class.
- To compare the mean response for two non-reference classes, look at the **difference in the coefficients** associated with those classes.
- Choice of reference class does not influence interpretation of coefficients.

Example

Consider the income of Americans, y, with predictors $x_1 =$ years in school and $x_2 =$ political affiliation (Democrat, Republican, Independent). Political affiliation has 3 classes, so we need to create two indicator variables. We could decide to make Independents the reference class

$$I_1 = \begin{cases} 1 & \text{if Democrat} \\ 0 & \text{otherwise;} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{if Republican} \\ 0 & \text{otherwise;} \end{cases}$$

So the regression equation is

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2.$$

Additive Effects

So far, we have looked at multiple linear regression models with **additive effects**. The effect of each predictor is added into the model.

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Additive effects assume that each predictor's effect on the response **does not depend** on the value(s) of the other predictor(s).

Interaction Effects

Interaction effects allow the effect of one predictor on the response to depend on the value(s) of the other predictor(s).

Property Example

In the example from mod6_categorical.pdf in Topic 6.2, we looked at sale prices, y, of homes, with predictors $x_1 =$ square footage of home and $x_2 =$ whether home has air conditioning. We use the indicator variable $I_1 = 1$ if the home has air conditioning, and $I_1 = 0$ if the home does not have air conditioning. So the model is

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 x_1 I_1.$$

where we have

$$\begin{aligned} \text{No air conditioning: } E(Y|x) &= \beta_0 + \beta_1 x_1. \\ \text{Air conditioning: } E(Y|x) &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1. \end{aligned}$$

Result: **different intercepts and different slopes**. The effect of square footage on home price depends on whether the home has air conditioning or not.

Interaction Effects

Interaction can also occur between quantitative predictors. Consider a student's score on an exam, y, with two predictors: amount of time spent in remedial classes, x_1, and student's current GPA, x_2. If we think the impact of time spent in remedial classes on the exam score differs based on the current GPA, an interaction between these predictors should be used.

$$E(Y|x) = 50 + 4x_1 + 0.5x_2 - 0.8x_1x_2.$$

- When $GPA = 2.5$, $E(Y|x) = 51.25 + 2x_1$.
- When $GPA = 4$, $E(Y|x) = 52 + 0.8x_1$.

Hierarchical Principle

Hierachical Principle: If higher order terms are significant, lower order terms must be left in the regression model

Regression with Categorical Predictors using R

It is crucial to check if R views your categorical predictors correctly.

- Sometimes, the data set uses numeric codes for categorical variables and R may not recognize as categorical. You must declare these to be categorical using the `factor()` function.
- If R already views the variable as categorical, you don't have to do anything. Use `is.factor()` to check.
- If the data set uses dummy codes, you don't have to do anything. The `lm()` converts categorical variables into dummy codes "behind the scenes".
- Also check your reference class. Use the `relevel()` function if needed.