**CATEGORICAL INDEPENDENT VARIABLES ACTIVITY GUIDANCE**

**Activity learning goals:**

* Incorporate categorical independent variables into linear regression models as sets of dummy variables.
* Interpret coefficients on dummy variables as expected changes in the conditional mean of the dependent variable relative to a reference category.
* Recognize and avoid the “dummy variable trap” of including dummy variables for every possible value of a categorical independent variable.

**Introducing the activity:**

Imagine that you run a local coffee shop and are trying to understand the determinants of your customers’ demand for coffee. Over the past year you have randomly varied the price you charge for coffee each week (*pi*) and recorded how many cups you sell each week (*qi*). You have also created a variable (*seasoni*) that is coded as 1 for spring, 2 for summer, 3 for fall, and 4 for winter.

**Guiding students during the activity:**

1. *How would you interpret the coefficient on season in the following model?*

At this point in the course, most students can interpret a coefficient on a count variable:   
β2 represents the expected difference in quantity sold between one season and the following season.

1. *What assumption are you making about the effects of the different seasons in this model?*The expected difference between spring and summer is the same as the difference between summer and fall and the difference between fall and winter. These are clearly not reasonable assumptions.
2. *Can you think of a better way to control for season in your model?*Students usually come up with a variety of ideas on their own, but if a group is stuck, you can suggest that they try defining a new variable (or set of variables) based on season and include those variables instead.

**Wrapping up the activity:**

Wrap by first walking through a couple of common “solutions.” Some students will create a single dummy variable for a season. Their model tells them nothing about expected differences in sales between the other seasons, and in essence, this solution throws away important information. Some students will put all four dummy variables in the model. Here we remind them that we often interpret the intercept substantively as the expected outcome holding all the independent variables equal to zero. This interpretation doesn’t make sense in this case because exactly one of the season dummy variables is always equal to one. It’s also difficult to interpret the coefficients on the other dummy variables. This may or may not be an appropriate time to point out that this model suffers from perfect multicollinearity. Finally, we present the expert’s solution: Choose a reference category and include all the other season dummy variables. Now we can clearly interpret all the model coefficients. Finish by showing that the choice of reference category has no effect on predicted differences between categories.