**INFO 2950, Lecture 10, 2023-09-25**

The chalkboard example done in class was in response to this question: why do we interpret dummy coefficients in multivariable regression *relative to the reference variable?*

Remember that our interpretation of a coefficient on an input dummy variable is: (all else equal,) having dummy variable = YES corresponds to a $\beta_B$ increase in output y, relative to the *reference level* = YES. If there are only two possible "levels" of a category (e.g. "Is Nose Pack" or "Is Not Nose Pack"), then the default reference level is the one corresponding to dummy variable = NO, i.e. "is not nose pack." Here, we instead have 3 "levels" of a category (Acne, Brow & Lash, and Cream & Lotion).

**Example**: let's say a skincare product can only be one of three different categories: Acne Treatment (A), Brow & Lash Treatment (B), and Cream & Lotion (C). We want to test whether our output, skincare product price (y = USD $) can be predicted by skincare product category.

Let's assume our reference variable is Acne Treatment, so we do not include $x_A$ in our regression (we drop the first dummy, per lecture). In class, our notation used numbers in subscripts; I'm just using letters A, B, and C here for ease of interpretation. We have:

$$\mathbf{y = \alpha + \beta_B x_B + \beta_C x_C}$$

- If a product is Acne Treatment, then $x_A$=1, $x_B$=0, and $x_C$=0. Plugging this into our regression gives us $\hat{\mathbf{y}}_\mathbf{A} = \alpha + \beta_B*0 + \beta_C*0 = \boldsymbol{\alpha}$
- If a product is Brow & Lash Treatment, then $x_A$=0, $x_B$=1, and $x_C$=0. Plugging this into our regression gives us $\hat{\mathbf{y}}_\mathbf{B} = \alpha + \beta_B*1 + \beta_C*0 = \boldsymbol{\alpha + \beta_B}$
- If a product is Cream & Lotion, then $x_A$=0, $x_B$=0, and $x_C$=1. Plugging this into our regression gives us $\hat{\mathbf{y}}_\mathbf{C} = \alpha + \beta_B*0 + \beta_C*1 = \boldsymbol{\alpha + \beta_C}$

Let's look at the differences between our predictions. Our model predicts skincare product price to be $\hat{y}_B = \alpha + \beta_B$ for Brow & Lash Treatment, as opposed to $\hat{y}_A = \alpha$ for Acne Treatment. We notice that the difference, $\hat{y}_B - \hat{y}_A = (\alpha + \beta_B) - (\alpha)$ is simply our coefficient $\beta_B$. So, here we could say that our coefficient $\beta_B$ represents the difference in our output y between scenario (1) if the product is Brow & Lash Treatment, and scenario (2) if the product is Acne Treatment. That's the same as saying that the coefficient is interpreted as the coefficient of $x_B$, relative to the reference variable (Acne Treatment)!

Similarly, We notice that the difference, $\hat{y}_C - \hat{y}_A = (\alpha + \beta_C) - (\alpha)$ is simply our coefficient $\beta_C$. So, our coefficient $\beta_C$ represents the change in output y when going from the model-estimated price of Acne Treatment to the model-estimated price of Cream & Lotion. Again, the coefficient of $x_C$ represents the change in y between if a product is Cream & Lotion, relative to Acne Treatment.

Are you only ever allowed to talk about changes relative to the reference variable (Acne Treatment)? No, it's mostly for ease of reading regression coefficients. For example, you could refer to the fact that output y changes by $\hat{y}_B - \hat{y}_C = (\alpha + \beta_B)-(\alpha + \beta_C) = \boldsymbol{\beta_B - \beta_C}$ if a product is Brow & Lash Treatment, relative to *Cream & Lotion*. But, that requires a little arithmetic (subtracting between different coefficients).