Section 3H. More Categorical Inputs

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Qualitative Inputs: More than two levels

Example: Consider a categorical value with more than two levels, e.g., **Ethnicity**$\in$ {Caucasian, African American, Asian}. Analyze the differences in credit card balance between ethnicities, ignoring other variables. The output variable y_i represents the credit card balance of individual i. We will consider the ethnicity of individual i as the only input x_i. How do we build a linear model? Steps:

Step 1) Choose a baseline for ethnicity, e.g., African American

Step 2) Since we have three (3) possible ethnicity categories, we create two (3-1) dummy variables:

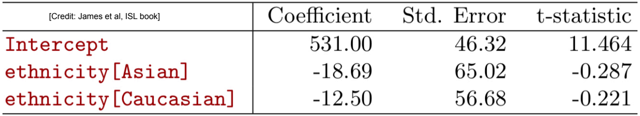
$$x_{i1} = \begin{cases} 1 & \text{if individual } i \text{ is Asian} \\ 0 & \text{if individual } i \text{ is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if individual } i \text{ is Caucasian} \\ 0 & \text{if individual } i \text{ is not Caucasian} \end{cases}$$

Qualitative Inputs: More than two levels

Step 3) Resulting model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if individual } i \text{ is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if individual } i \text{ is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if individual } i \text{ is AA} \end{cases}$$



[Credit: James et al, ISL book]	Coefficient	Std. Error	t-statistic
Intercept	531.00	46.32	11.464
ethnicity[Asian]	-18.69	65.02	-0.287
ethnicity[Caucasian]	-12.50	56.68	-0.221

Figure: Analysis of the Intercept (β_0), the Asian coefficient (β_1), and the Caucasian coefficient (β_2) in the credit card dataset.

Penn Engineering