

Section 3G. Categorical Inputs

Statistics for Data Science

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

Recap: Linear Regression

- ▶ We consider a model of the form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- ▶ We are provided a dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ with points generated according to the model above; in other words,
 - ▶ For $i = 1$ to N
 - ▶ We draw a random input vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^\top$ from a distribution f_X
 - ▶ We generate an output $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$, where $\varepsilon_i \sim f_\varepsilon$
 - ▶ end For-loop

Recap: Linear Regression (cont.)

- ▶ *Remark:* We do not have direct knowledge about the parameters β_0, \dots, β_p ; these are values that Nature is using to generate the data. Our knowledge is limited to the dataset \mathcal{D} generated by the linear model
- ▶ In the previous slides, we have seen how to:
 1. Compute estimates $\hat{\beta}_0, \dots, \hat{\beta}_p$ from \mathcal{D} alone
 2. Analyze the uncertainty of our estimates using *Confidence Intervals*; in other words, we can claim that the true value of a parameter β_i is within a particular interval with a 95% probability
 3. Determine how likely it is for a particular input X_i to influence the output Y . In this task, we used *Hypothesis Testing*, which allow us to build statistical evidence to reject Null Hypothesis of the form: X_i does not influence Y (legal parable: Bob did not kill Alice)

Qualitative Inputs

Linear models can handle *qualitative* inputs, also called *categorical* variables, taking a discrete set of values. For example:

- ▶ $\text{Gender} \in \{\text{Male}, \text{Female}\}$
- ▶ $\text{Marital status} \in \{\text{Single}, \text{Married}\}$
- ▶ $\text{Ethnicity} \in \{\text{Caucasian}, \text{African American}, \text{Asian}\}$

Example: Analyze the differences in credit card balance between males and females, ignoring other variables. The output variable y_i represents the credit card balance of individual i . We will consider the gender of individual i as the only input x_i . How do we build a linear model?

Steps:

Step 1) Create a dummy variable:

$$x_i = \begin{cases} 1 & \text{if individual } i \text{ is a female} \\ 0 & \text{if individual } i \text{ is a male} \end{cases}$$

Qualitative Inputs

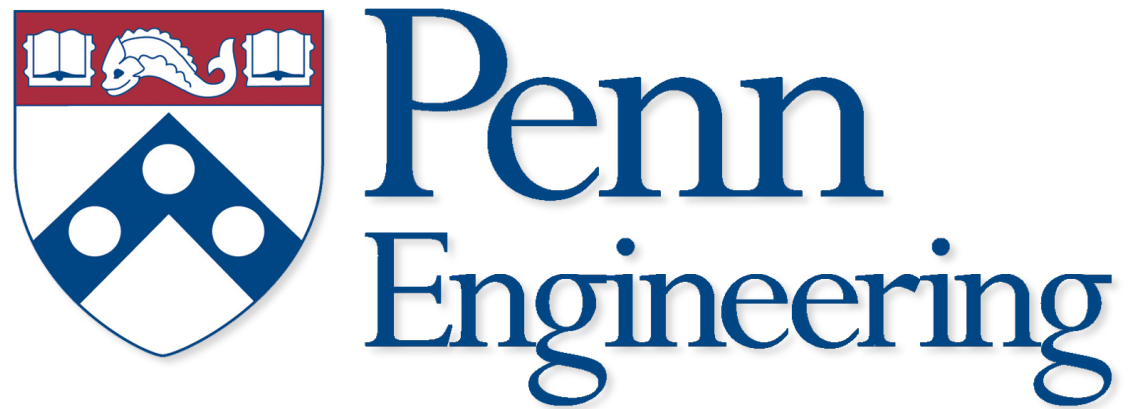
Step 2) Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if individual } i \text{ is a female} \\ \beta_0 + \varepsilon_i & \text{if individual } i \text{ is a male} \end{cases}$$

Hence, β_1 quantifies the increment (decrement) in the Balance of individual i when she is a female.

<small>[Credit: James et al. ISL book]</small>	Coefficient	Std. Error	t-statistic
Intercept	509.80	33.13	15.389
gender[Female]	19.73	46.05	0.429

Figure: Analysis of the Intercept (β_0) and the gender coefficient (β_1) in the credit card dataset.



Copyright 2020 University of Pennsylvania
No reproduction or distribution without permission.