Feature Engineering & Design Matrix

Qualitative Predictors

So far, we have assumed that all variables are quantitative. But in practice, often some predictors are qualitative. For example, this credit data set contains information about balance, age, cards, education, income, limit , and rating for a number of potential customers.

Skip to after table

Income	Limit	Rating	Cards	Age	Education	Sex	Student	Married	Ethnicity	Balance
14.890	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.02	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.59	7075	514	4	71	11	Male	No	No	Asian	580
148.92	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

Binary Variables

If the predictor takes only two values, then we create an indicator or dummy variable that takes on two possible numerical values. For example, for gender, we create a new variable: We then use this variable as a predictor in the regression equation.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person otherwise} \end{cases}$$

We then use this variable as a predictor in the regression equation.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is not} \end{cases}$$

Question: What is interpretation of β_0 and β_1?

β_0 is the **average** credit card balance among those who are **not female**, $\beta_0 + \beta_1$, is the **average** credit card balance among those who **are female**, and β_1 is the **average difference** in credit card balance **between the two categories**.

Example: Calculate β_0 and β_1 for the Credit data. you should find $\beta_0 \approx 509$ and $\beta_1 \approx 19$.

More than two values (one-hot encoding)

Often, the qualitative predictor takes more than two values (e.g. ethnicity in the credit data).

In this situation, a single dummy variable cannot represent all possible values. We create *additional* dummy variable as:

$$x_{i,1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

We then use these variables as predictors, the regression equation becomes:

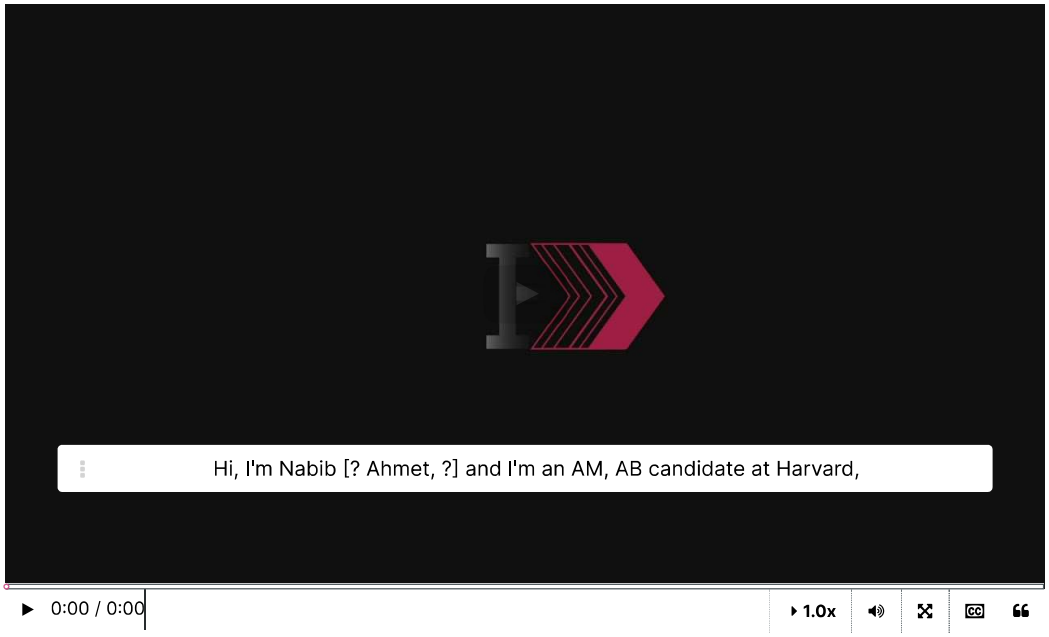
$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \begin{cases} \beta_0 + \beta_1 x_{i,1} + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 x_{i,2} + \varepsilon_i & \text{if } i\text{th person is Caucasian} \end{cases}$$

> **Concept Question**
>
> What is the interpretation of β_0, β_1, and β_2?
>
> B0 is the avarage credit card balance for african americans,
> B1 is the difference between Asians and African americans
> B2 is the diference between Caucasians and African Amer

Using Categorical Variables

How do people use categorical variables? Nabib shows us one example below, in a field you might not normally associate with data science.

Video



Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

The Data Is Not The Reality

Remember that when we talk about categorical measurements, those are statements about *the data we are given to analyze*. Not all data represents reality accurately. Sometimes we have to do our best with the data we currently have, and work to collect better data next time.

Discussion Board (External resource)

Click OK to have your username and e-mail address sent to a 3rd party application.

OK

Previous Next

深圳市恒宇博科技有限公司 粤ICP备17044299号-2