<u>Help</u>

Samue|Simao47 >

Next >

Curso Progresso Datas Anotações Discussão Programa de Estudos Perguntas frequentes Related Courses Resources

Resources

↑ Course / Section 2: Multiple and Polynomial Regression / 2.2 Techniques for Multilinear Modeling



☐ Bookmark this page

< Previous

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.

$$x_i = egin{cases} 1 & ext{if ith person is female} \ 0 & ext{if ith person otherwise} \end{cases}$$

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

$$y_i = eta_0 + eta_1 x_i + arepsilon_i igg\{ eta_0 + eta_1 x_i + arepsilon_i & ext{if ith person is female} \ eta_0 + arepsilon_i & ext{if ith person is not} \ \end{array}$$

Question: What is interpretation of β_0 and β_1 ?

 eta_0 is the **average** credit card balance among those who are **not female**, eta_0+eta_1 , is the **average** credit card balance among those who **are female**, and eta_1 is the **average difference** in credit card balance **between the two categories**.

Example: Calculate eta_0 and eta_1 for the Credit data. you should find $eta_0pprox 509$ and $eta_1pprox 19$.

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:

$$m{x_{i,1}} = egin{cases} 1 & ext{if ith person is Asian} \ 0 & ext{if ith person is not Asian} \end{cases}$$

$$x_{i,2} = egin{cases} 1 & ext{if ith person is Caucasian} \ 0 & ext{if ith person is not Caucasian} \end{cases}$$

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

$$y_i = eta_0 + eta_1 x_{i,1} + eta_2 x_{i,2} + arepsilon_i igg\{eta_0 + eta_1 x_{i,1} + arepsilon_i & ext{if ith person is Asian} \ eta_0 + eta_2 x_{i,2} + arepsilon_i & ext{if ith person is Caucasian} \ eta_0 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} eta_0 + eta_1 x_{i,1} + arepsilon_i & ext{otherwise} \ eta_0 + eta_1 x_{i,1} + arepsilon_i & ext{otherwise} \ eta_0 + eta_1 x_{i,1} + arepsilon_i & ext{otherwise} \ eta_0 + eta_1 x_{i,1} + arepsilon_i & ext{otherwise} \ eta_1 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_1 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_1 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_1 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_1 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_1 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + eta_2 x_{i,2} + arepsilon_i & ext{otherwise} \ eta_2 + eta_2 x_{i,2} + eta_2 x_{i,2} + eta_2 x_{i,2} + eta_3 x_{i,$$

② CONCEPT QUESTION

What is the interpretation of $oldsymbol{eta_0}$, $oldsymbol{eta_1}$, and $oldsymbol{eta_2}$?

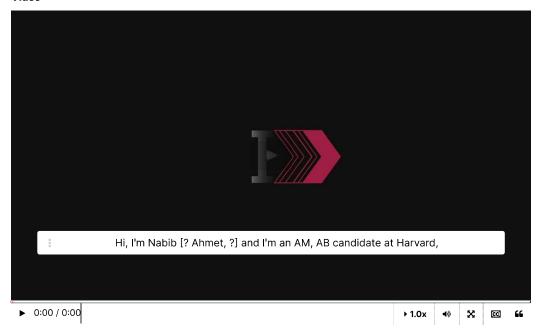
B0 is the avarage credit card balance for african americans, B1 is the difference between Asians and African americans

B2 is the diference between Caucasians and African Amer



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

▲ 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.

C Previous
Next >

© All Rights Reserved



edX

About
Affiliates
edX for Business
Open edX
Careers



Legal

Terms of Service & Honor Code

Privacy Policy

Accessibility Policy

<u>Trademark Policy</u>

<u>Sitemap</u>

Cookie Policy

Your Privacy Choices

Connect

<u>Blog</u>

Contact Us

Help Center

<u>Security</u>

Media Kit















© 2023 edX LLC. All rights reserved.

深圳市恒宇博科技有限公司 <u>粤ICP备17044299号-2</u>