



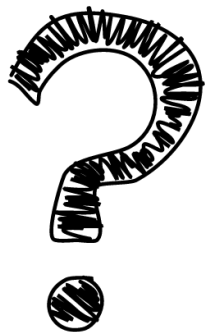
DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

Variável dummy.

Aula 23
15 de junho de 2022

Ana Paula Karruz



Numa regressão, como incorporar a noção de pertencimento a grupos como variável explicativa?

R. Usando dummies.

Variáveis independentes qualitativas

- As **variáveis binárias** (também chamadas de **dicotômicas** ou **dummies**) agregam informação qualitativa a modelos de regressão
- Alguns exemplos de atributos qualitativos são sexo, opinião (sim ou não, concorda ou não concorda) e diferentes categorizações de idade e escolaridade

Definição de variáveis dummy

- É preciso definir a qual evento/condição atribuiremos o valor um, e a qual evento/condição atribuiremos o valor zero
- Essa decisão **não afeta os resultados estimados (\hat{Y}) para cada evento/condição**, mas deve balizar a **escolha de nomes** para a variável dummy em questão. Isto é, devemos nomeá-la de forma a representar o **evento/condição em que a dummy assume o valor um**. Por exemplo, uma dummy para afiliação religiosa com duas opções, cristão ou budista, deve ser nomeada da seguinte forma:
 - “Cristão” se dummy = 1 para cristãos; dummy = 0 caso contrário (i.e., para budistas)
 - “Budista” se dummy = 1 para budistas; dummy = 0 caso contrário (i.e. para cristãos)

“Religião” seria um péssimo nome

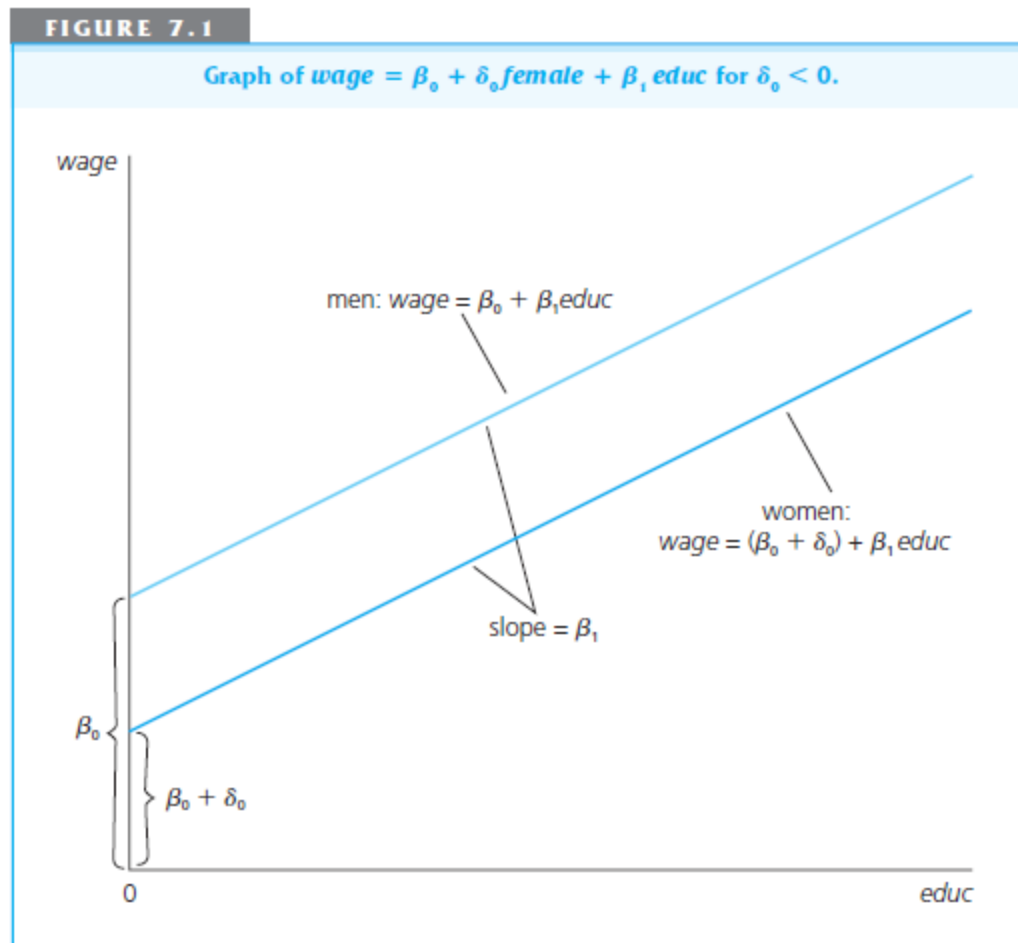
Uma única variável independente binária

- Com somente uma variável dicotômica explicativa (i.e., representando pertencimento a um de dois grupos), simplesmente adicionamos a variável dummy à equação como uma variável independente

$$\text{Salário} = \beta_0 + \beta_1 \text{Escolaridade} + \beta_2 \text{Female} + \varepsilon$$

- Aqui, β_2 é a diferença média no salário entre mulheres e homens, para um mesmo grau de escolaridade
 - Se $\beta_2 < 0$, então as mulheres ganham em média menos que os homens, para o mesmo nível dos outros fatores (aqui, apenas escolaridade)
 - Se $\beta_2 > 0$, então as mulheres ganham em média mais que os homens, para o mesmo nível dos outros fatores (aqui, apenas escolaridade)
 - Se $\beta_2 = 0$, então as mulheres ganham em média o mesmo que os homens, para o mesmo nível dos outros fatores (aqui, apenas escolaridade)
- A diferença salarial entre mulheres e homens pode ser descrita graficamente como um **deslocamento de intercepto** entre as linhas que representam cada sexo

Deslocamento de intercepto: uma linha para cada sexo



A diferença salarial entre sexos não depende do nível de escolaridade, por isso as retas são paralelas

Fonte: Wooldridge (2009: 228).

Precisamos de um grupo de referência

$$\text{Salário} = \beta_0 + \beta_1 \text{Escolaridade} + \beta_2 \text{Female} + \varepsilon$$

- Na equação acima, o intercepto para homens é β_0 e o intercepto para mulheres é $\beta_0 + \beta_2$
- Como o exemplo prevê apenas dois grupos, são necessários apenas dois interceptos diferentes
- Por isso, seria redundante incluir uma variável binária Male, além da variável Female
- O uso de duas variáveis binárias introduziria colinearidade perfeita, porque $\text{Female} + \text{Male} = 1$, o que significa que Male é uma função linear perfeita de Female (e vice-versa)

Neste exemplo, homens foram escolhidos para compor o **grupo de referência (ou grupo base)**, que é o grupo contra o qual as comparações são realizadas. O grupo de referência é **omitido da equação, já que seu intercepto é dado diretamente por β_0 .**

Exemplo: Dummy para estimar diferença de médias

Human Capital and Black-White Earnings Gaps, 1966-2017

Owen Thompson

WORKING PAPER 28586

DOI 10.3386/w28586

ISSUE DATE March 2021

<https://www.nber.org/papers/w28586>

Table 2: Black-White Gaps in Human Capital Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
	Educational Attainment			Test Scores		
	NLS-OC	NSLY-79	NLSY-97	NLS-OC	NSLY-79	NLSY-97
Black	-1.01*** (0.13)	-0.83*** (0.08)	-1.12*** (0.11)	-1.02*** (0.07)	-1.03*** (0.03)	-0.83*** (0.04)
Observations	18,138	42,717	23,660	18,138	42,717	23,660

Notes: The dependent variable for Columns 1-3 is educational attainment measured in years, while the dependent variable for Columns 4-6 is standardized test scores measured in standard deviations. Observations consist of person-years. All samples are restricted to non-Hispanic Black and white men between the ages of 21 and 37 who are not currently enrolled in school. Sampling weights applied. Standard errors are clustered at the individual level and reported in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels, respectively.

Exemplo: Dummy para estimar diferença de médias, com covariáveis

Human Capital and Black-White Earnings Gaps, 1966-2017

Owen Thompson

WORKING PAPER 28586

DOI 10.3386/w28586

ISSUE DATE March 2021

<https://www.nber.org/papers/w28586>

Table 1: Unconditional and Conditional Black-White Earnings Differentials

	(1)	(2)	(3)	(4)	(5)	(6)
	NLS-OC		NLSY-79		NLSY-97	
	Baseline	With Controls	Baseline	With Controls	Baseline	With Controls
Black	-0.960*** (0.144)	-0.866*** (0.151)	-1.681*** (0.100)	-1.407*** (0.111)	-1.974*** (0.135)	-1.420*** (0.136)
Educational Attainment (years)		0.072*** (0.018)		0.157*** (0.021)		0.191*** (0.021)
Test Score (standard deviations)		0.021 (0.045)		0.137** (0.059)		0.410*** (0.063)
Observations	18,138	18,138	42,717	42,717	23,660	23,660
Level Change After Covariates	-0.09		-0.27		-0.55	
Percent Change After Covariates	-9.8%		-16.3%		-28.1%	

Notes: The dependent variable for all models is the inverse hyperbolic sine of total earnings, with zeros included. Observations consist of person-years. All samples are restricted to non-Hispanic Black and white men between the ages of 21 and 37 who are not currently enrolled in school. Sampling weights applied. Standard errors are clustered at the individual level and reported in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels, respectively.

Informações ordinais com variáveis binárias

- As categorias de variáveis ordinais podem ser organizadas em alguma ordem (crescente ou decrescente)
- Sabemos que há diferenças relativas entre os valores dos dados, mas não sabemos as magnitudes das diferenças
- Por exemplo, na escala de frequência “pouco/ médio/ muito”, é possível ordenar os dados, mas não sabemos se a diferença entre “pouco” e “médio” é a mesma existente entre “médio” e “muito”
- Aqui não faz sentido supor que o aumento de uma unidade nessa variável terá um efeito constante sobre outra variável
- Mas é possível criar três variáveis binárias, tomando uma como referência

Exemplo: Uso de dummies para mais de 2 categorias

Population Equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Y= wage in \$

X_1 = high school diploma only

X_2 = some college

(reference group: people without a high school degree)

Estimated Equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$$\hat{Y} = 5.25 + 1.0X_1 + 5.0X_2$$

Slope Coefficients

$\hat{\beta}_1$ Pessoas com diploma do ensino médio, mas que nunca foram para a faculdade, ganham \$1,00 a mais do que aqueles sem diploma do ensino médio

$\hat{\beta}_2$ Pessoas com alguma faculdade ganham \$5,00 a mais do que aqueles sem diploma do ensino médio e ganham \$4,00 a mais do que os graduados do ensino médio que nunca foram para a faculdade

Não caia na pegadinha!
Se $X_2 = 1$,
então $X_1 = 0$



DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

Variável dummy.

Aula 23
15 de junho de 2022

Ana Paula Karruz