PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Variável dummy e interação

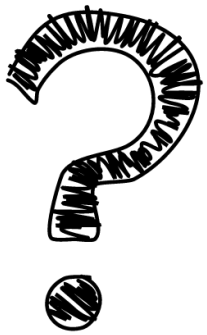
Aula 7
26 de outubro de 2022

Ana Paula Karruz

Agenda para esta aula

1. **Variável dummy**
2. Interação

Para implementação em R, vide Exercício 3



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: sexo, opinião (e.g., sim ou não, concorda ou não concorda) e diferentes categorizações de idade e escolaridade

Definição de variável dummy

- Uma variável dummy assume apenas dois valores: 0 ou 1

- 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 (Yhat) 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

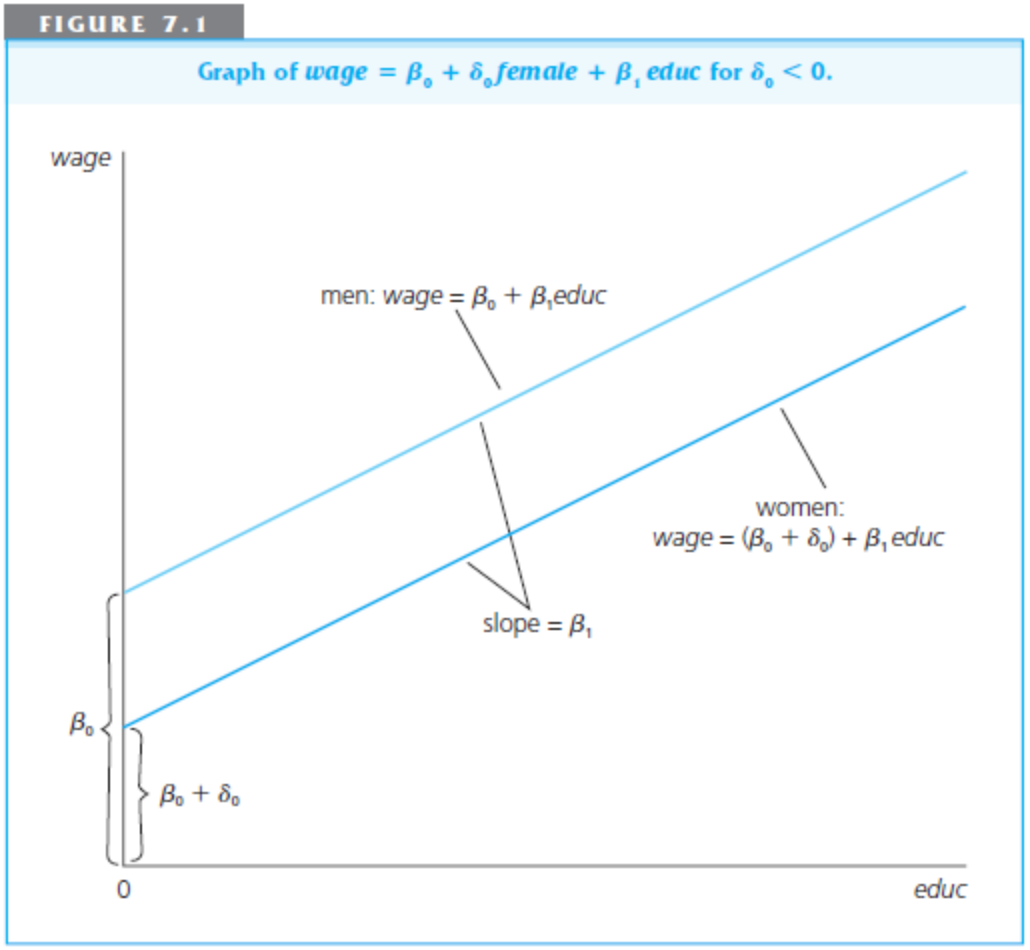
Uso de dummy com uma única variável independente binária

- 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

$$\mathbf{wage = \beta_0 + \beta_1 female + \beta_2 educ + \varepsilon}$$

- Aqui, β_1 é a diferença média no salário entre mulheres e homens, para um mesmo grau de escolaridade
  - Se $\beta_1 < 0$, então as mulheres ganham em média menos que os homens, para o mesmo nível dos outros fatores (aqui, apenas escolaridade)
  - Se $\beta_1 > 0$, então as mulheres ganham em média mais que os homens, para o mesmo nível dos outros fatores (aqui, apenas escolaridade)
  - Se $\beta_1 = 0$, 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**, provocando duas linhas paralelas (uma para cada sexo)

Deslocamento de intercepto: uma linha para cada sexo



Fonte: Wooldridge (2009: 228).

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

Precisamos de um grupo de referência

$$\mathbf{wage = \beta_0 + \beta_1 female + \beta_2 educ + \varepsilon}$$

- Na equação acima, o intercepto para homens é β_0 e o intercepto para mulheres é $\beta_0 + \beta_1$
- 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 female + 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

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***	-0.83***	-1.12***	-1.02***	-1.03***	-0.83***
	(0.13)	(0.08)	(0.11)	(0.07)	(0.03)	(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 covariávies

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***	-0.866***	-1.681***	-1.407***	-1.974***	-1.420***
	(0.144)	(0.151)	(0.100)	(0.111)	(0.135)	(0.136)
Educational Attainment (years)		0.072***		0.157***		0.191***
		(0.018)		(0.021)		(0.021)
Test Score (standard deviations)		0.021		0.137**		0.410***
		(0.045)		(0.059)		(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.

APÊNDICE:
Testes t de diferença de médias

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 assumidos por essas variáveis, 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

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

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

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 $X_1 = 0$

Agenda para esta aula

1. Variável dummy
2. **Interação**

Para implementação em R, vide Exercício 3

Interação com dummy

- Nossa discussão anterior sobre variáveis dummy tratou de casos em que a variável categórica afetava a dependente no intercepto (i.e., no nível), mas não na inclinação. Endereçávamos questões do tipo "Há uma diferença sistemática entre salários (Y) de mulheres e homens, para um mesmo nível de escolaridade (X)?"

$$\mathbf{wage = \beta_0 + \beta_1 female + \beta_2 educ + \varepsilon}$$

Qual a representação gráfica desta equação?

- Adicionalmente, pode ser relevante saber se a relação entre a variável dependente (Y) e uma variável independente (X) difere para subgrupos da população. Isso corresponde a questões do tipo "Mulheres e homens obtêm o mesmo retorno do investimento em educação?". Para endereçar esta questão é preciso interagir variáveis (i.e., multiplicá-las), como segue:

$$\mathbf{wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 (female*educ) + \varepsilon}$$

- Na equação com interação, o sexo afeta tanto o intercepto como a inclinação
  Salário estimado:

Male: $\widehat{wage} = \widehat{\beta_0} + \widehat{\beta_2} educ$
Female: $\widehat{wage} = \widehat{\beta_0} + \widehat{\beta_1} + (\widehat{\beta_2} + \widehat{\beta_3}) educ$

Efeito estimado do aumento de uma unidade em educ:

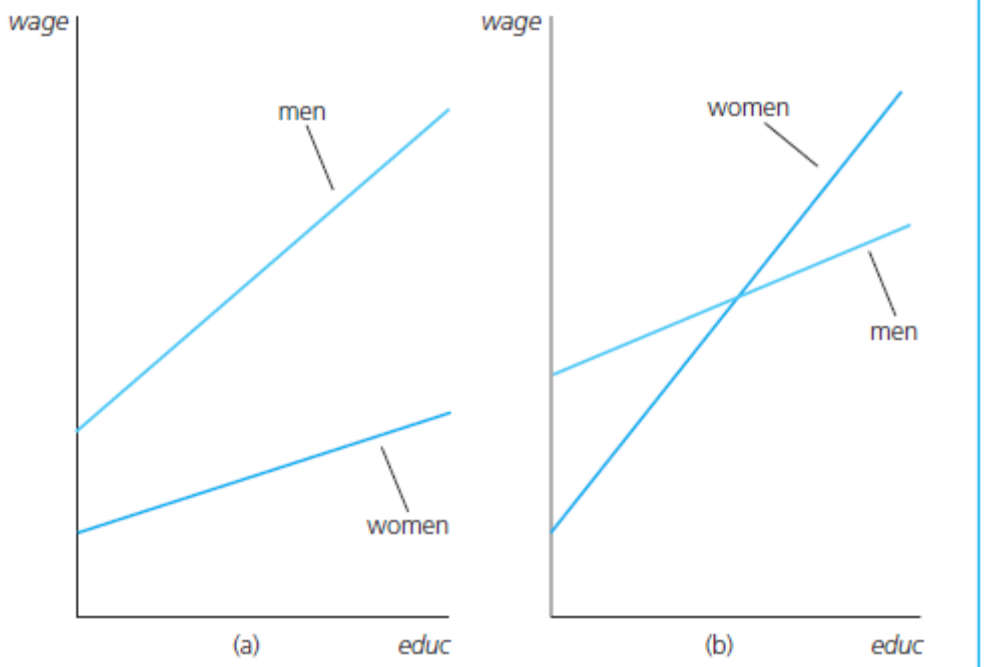
Male: $\widehat{\beta_2}$
Female: $\widehat{\beta_2} + \widehat{\beta_3}$

B_3hat, o coeficiente estimado do termo de interação, é nossa estimativa do efeito diferencial da educação no salário de mulheres em relação a homens

Interação com dummy: Representações gráficas

FIGURE 7.2

Graphs of equation (7.16): (a) $\delta_0 < 0, \delta_1 < 0$; (b) $\delta_0 < 0, \delta_1 > 0$.



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

Fonte: Wooldridge (2009: 240).

Exemplo: Interação com dummy

Quais coeficientes são estatisticamente diferentes de zero? O que isso implica para as conclusões desta análise?

wage = β_0 + $\beta_1$female + $\beta_2$educ + β_3(female*educ) + ε

femeduc = female*educ

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.539476	.0642229	8.40	0.000	.4133089	.6656432
female	-1.198523	1.32504	-0.90	0.366	-3.801589	1.404543
femeduc	-.085999	.1036388	-0.83	0.407	-.2895994	.1176014
_cons	.2004963	.8435616	0.24	0.812	-1.456696	1.857689

$$\widehat{wage} = 0,20 - 1,20female + 0,54educ - 0,09(educ * \text{female})$$

- Mulheres sem qualquer educação formal ganham, em média, 1,20 menos que os homens
- Um ano adicional de educação formal associa-se com um aumento de 0,54 no salário de homens
- Um ano adicional de educação formal associa-se com um aumento de 0,45 (0,54 – 0,09) no salário de mulheres
- O efeito diferencial de um ano adicional de educação formal para mulheres em relação a homens é de -0,09

Interação entre duas dummies

$$wage = \beta_0 + \beta_1 fem + \beta_2 mar + \beta_3 (fem \cdot mar) + \varepsilon$$

$$\hat{wage} = 5.2 - .56\, fem + 2.8 mar - 2.9(fem \cdot mar)$$

Interpretação:

Group	Values	Predicted wage
Unmarried men	fem=0, mar=0	$ 5.2
Unmarried women	fem=1, mar=0	$ 4.64 (5.2-.56)
Married men	fem=0, mar=1	$ 8.00 (5.2 +2.8)
Married women	fem=1, mar=1	$ 4.54 (5.2-.56+2.8-2.9)

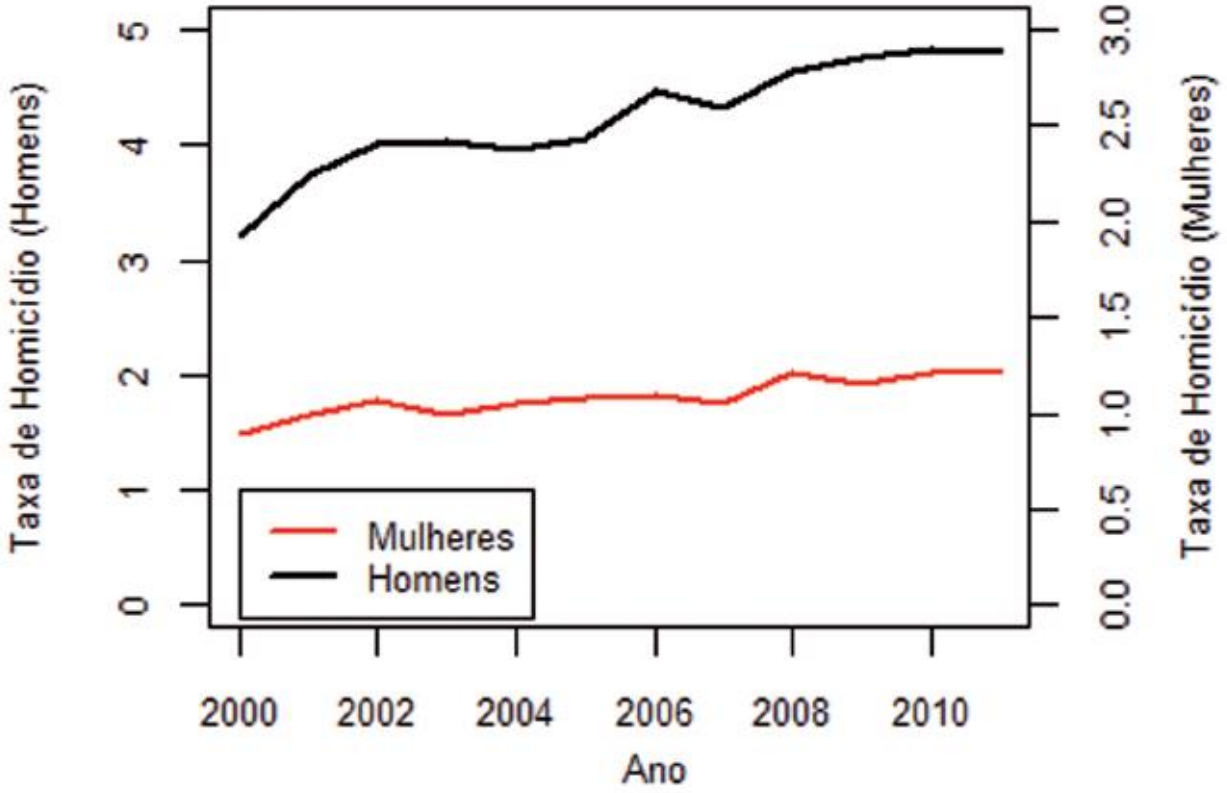
- Coeficiente de *fem*: Estima-se que o salário médio de mulheres solteiras seja 0,56 inferior ao de homens solteiros

- Coeficiente de *mar*: Estima-se que o salário médio de homens casados seja 2,8 superior ao de homens solteiros

- Coeficiente de *fem*mar*: Estima-se que o impacto médio do casamento no salário seja 2,90 inferior para mulheres em relação a homens

Exemplo: Interação entre duas dummies (modelo de diferença em diferenças)

GRÁFICO 2
Taxa de homicídios ocorridos em residência – Brasil (2000-2011)
(Por 100 mil habitantes)



Fonte: SIM.
Elaboração: Diest/Ipea.
Obs.: imagem cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais disponibilizados pelos autores para publicação (nota do Editorial).

2048
TEXTO PARA DISCUSSÃO
Brasília, março de 2015

AVALIANDO A EFETIVIDADE DA LEI MARIA DA PENHA[1]

Daniel Cerqueira[2]
Mariana Vieira Martins Matos[3]
Ana Paula Antunes Martins[4]
Jony Pinto Junior[5]

https://www.ipea.gov.br/portal/images/stories/PDFs/TDs/td_2048k.pdf

Exemplo: Interação entre duas dummies (modelo de diferença em diferenças)

TABELA 3

Resumo da estimação dos modelos de diferenças para a variável *logaritmo da taxa de homicídios em residência*

Variável dependente: *ln*(taxa de homicídios em residência)				
	(5)	(6)	(7)	(8)
Sexo (β_1)	-1,179	-1,181	-1,182	-1,183
	(<0,001)***	(<0,001)***	(<0,001)***	(<0,001)***
Vigência da lei (β_2)	0,339	0,340	NA	NA
	(<0,001)***	(<0,001)***	-	-
Sexo*lei (β_3)	-0,099	-0,097	-0,096	-0,096
	(<0,001)***	(<0,001)***	(<0,001)***	(<0,001)***
ln(Taxa_alcool)	-	-	-	0,135
	-	-	-	(<0,001)***
ln(Taxa_armas)	-	-	-	0,091
	-	-	-	(<0,001)***
Efeito fixo de microrregião	Não	Sim	Sim	Sim
Efeito fixo de tempo	Não	Não	Sim	Sim
Prob. > *F*	(<0,001)***	(<0,001)***	(<0,001)***	(<0,001)***
R – ajustado	0,453	0,695	0,705	0,710
Número de observações	13.392	13.392	13.392	13.358

Fonte: SIM.
Elaboração: Diest/Ipea.
Obs.: * <0,05; ** <0,01; ***<0,001; NA – não definido por causa de singularidade. A taxa de armas é uma *proxy* para a prevalência de armas de fogo nas microrregiões construída a partir da proporção de suicídios por armas de fogo em relação ao total de suicídios. A taxa de álcool é uma *proxy* para consumo de bebida alcoólica nas microrregiões, construída pela soma de óbitos ocasionados pelo envenenamento por bebidas alcoólicas, relativizados pela população residente na localidade. Para contabilizar apenas os homicídios que ocorreram em residências, utilizamos o terceiro dígito da CID-10.

AVALIANDO A EFETIVIDADE DA LEI MARIA DA PENHA[1]

Daniel Cerqueira[2]
Mariana Vieira Martins Matos[3]
Ana Paula Antunes Martins[4]
Jony Pinto Junior[5]

https://www.ipea.gov.br/portal/images/stories/PDFs/TDs/td_2048k.pdf

Como interpretar interações entre variáveis contínuas?

UCLA Institute for Digital Research & Education

https://stats.idre.ucla.edu/stata/faq/how-can-i-explain-a-continuous-by-continuous-interaction-stata-12/

https://stats.idre.ucla.edu/r/faq/how-can-i-explain-a-continuous-by-continuous-interaction/

A abordagem que demonstraremos é calcular inclinações simples, ou seja, as inclinações da variável dependente na variável independente quando a variável moderadora é mantida constante em diferentes combinações de valores de muito baixo a muito alto.

APÊNDICE:
Testes t de diferença de médias

Teste t de diferença de médias apura se dois grupos diferem em relação a um atributo (variável)

Testes t de diferença de médias podem assumir duas configurações:

- **Teste pareado:** compara a média da variável de interesse entre duas observações das mesmas unidades (e.g., pressão arterial antes e depois do uso de um medicamento)

- **Teste não pareado:** compara a média da variável de interesse entre duas amostras compostas por unidades diferentes (e.g., pressão arterial de Atleticanos e Cruzeirenses)

  - Assunção de variâncias desiguais

  - Assunção de variâncias iguais

Teste t de diferença de médias: PAREADO

H_0: $\mu_1 = \mu_2$

$$t = \frac{m}{s/\sqrt{n}}$$

onde,

- m = média das diferenças
- s = desvio padrão das diferenças
- n = tamanho da amostra

Podemos computar o p-valor correspondente ao valor absoluto da estatística t (|t|) para os graus de liberdade (degrees of freedom = df)

- df = n – 1

Teste t de diferença de médias: PAREADO

```
> # Executa teste t de diferenca de medias:
> # t.test(Y[Dum==1], Y[Dum==0], paired = FALSE, var.equal = FALSE) # esta eh a configuracao default
do teste
> l = t.test(womenlabor$wlfp[womenlabor$d90==1],
+         womenlabor$wlfp[womenlabor$d90==0],
+         paired = TRUE)
> l

              Paired t-test

data:  womenlabor$wlfp[womenlabor$d90 == 1] and womenlabor$wlfp[womenlabor$d90 == 0]
t = 35.478, df = 49, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.801606 7.618394
sample estimates:
mean of the differences
                   7.21

> l$stderr
[1] 0.203224

> # Executa teste t pareado "manualmente":
> summary(dados$diff)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.800   6.250   7.100   7.210   7.675  10.000
> var(dados$diff)
[1] 2.065
>
> m = mean(dados$diff)
> v = var(dados$diff)
> s = v^0.5
> r = 50^0.5
> d = s/r
> d
[1] 0.203224
> t_stat = m/d
> t_stat
[1] 35.47809
```

Equivalências entre MQO e teste t de diferença de médias: PAREADO

Um modelo **MQO nulo (i.e., sem variáveis explicativas)** tendo como variável dependente as **diferenças pareadas** produz erro padrão e estatística t do intercepto **equivalentes** ao erro padrão ($\frac{s}{\sqrt{n}}$) e estatística t produzidos por um **teste t pareado**.

Equivalências entre MQO e teste t de diferença de médias: PAREADO

```
> # Carrega dados sobre participacao das mulheres no
mercado de trabalho para comparar test t pareado e modelo
nulo
> #####
> womenlabor = read.table("womenlabor.csv", header=T,
dec=".", sep = ";")
> colnames(womenlabor)
 [1] "wlfp"    "yf"      "ym"      "educ"    "ue"      "mr"
"dr"      "urb"     "wh"      "d90"     "stateid"
> dim(womenlabor)
[1] 100  11

> # Calcula as diferenças pareadas (i.e., dentro de cada
estado)
> dados = womenlabor[,c("wlfp", "stateid", "d90")]
> dados = reshape(dados, idvar = "stateid", timevar =
"d90", direction = "wide")
> colnames(dados)
[1] "stateid" "wlfp.0"  "wlfp.1"
> dados$diff = dados$wlfp.1 - dados$wlfp.0

> reg_diff = lm(diff ~ 1, data = dados)
> summary(reg_diff)

Call:
lm(formula = diff ~ 1, data = dados)

Residuals:
   Min     1Q Median     3Q    Max
-4.410 -0.960 -0.110  0.465  2.790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.2100     0.2032   35.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 1.437 on 49 degrees of freedom
```

```
> # Executa teste t de diferenca de medias:
t.test(Y[Dum==1], Y[Dum==0], paired = FALSE, var.equal =
FALSE)
> l = t.test(womenlabor$wlfp[womenlabor$d90==1],
+         womenlabor$wlfp[womenlabor$d90==0],
+         paired = TRUE)
> l
             Paired t-test

data:  womenlabor$wlfp[womenlabor$d90 == 1] and
womenlabor$wlfp[womenlabor$d90 == 0]
t = 35.478, df = 49, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 6.801606 7.618394
sample estimates:
mean of the differences
                    7.21
> l$stderr
[1] 0.203224
```

Teste t de diferença de médias: NÃO PAREADO

$H_0: \mu_1 = \mu_2$

Variâncias desiguais: Welch test

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Conservative p-values can be obtained using the *t(k)* distribution, with *k* equal to either the smaller n_1-*1* and n_2-*1* or the calculated degrees of freedom.

Variâncias iguais

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is used in the *pooled two-sample t statistic*

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has a $t(n_1 + n_2 - 2)$ distribution.

Teste t de diferença de médias: NÃO PAREADO

$H_0: \mu_1 = \mu_2$

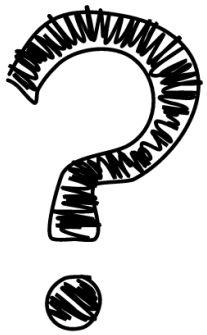
```
> # Carrega dados sobre carros
> dados <- as.data.frame(read_dta('auto.dta'))
> colnames(dados)
 [1] "make"         "price"        "mpg"          "rep78"
"headroom"     "trunk"
 [7] "weight"       "length"       "turn"
"displacement" "gear_ratio"   "foreign"
> dim(dados)
[1] 74 12
> # (i) test t de diferenca de medias (nao pareado): sem
assuncao de variancias iguais (teste Welch)
> # Executa teste t de diferenca de medias: variancias
desiguais
> # t.test(Y[Dum==1], Y[Dum==0], paired = FALSE, var.equal
= FALSE) # esta eh a configuracao default do teste
> l_desigual = t.test(dados$price[dados$foreign==1],
           dados$price[dados$foreign==0],
           paired = FALSE,
           var.equal = FALSE)
> l_desigual
Welch Two Sample t-test

data:  dados$price[dados$foreign == 1] and
dados$price[dados$foreign == 0]
t = 0.44296, df = 46.447, p-value = 0.6599
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -1106.339  1730.856
sample estimates:
mean of x mean of y
 6384.682  6072.423
> l_desigual$stderr
[1] 704.9376
> d = 6384.682 - 6072.423
> d
[1] 312.259
> d/704.9376
[1] 0.4429598
```

```
> # Executa teste t de diferenca de medias (nao pareado)
com assuncao de variancias iguais
> # t.test(Y[Dum==1], Y[Dum==0], paired = FALSE, var.equal
= FALSE) # esta eh a configuracao default do teste
> l_igual = t.test(dados$price[dados$foreign==1],
                   dados$price[dados$foreign==0],
                   paired = FALSE,
                   var.equal = TRUE)
> l_igual
             Two Sample t-test

data:  dados$price[dados$foreign == 1] and
dados$price[dados$foreign == 0]
t = 0.41389, df = 72, p-value = 0.6802
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -1191.708  1816.225
sample estimates:
mean of x mean of y
 6384.682  6072.423

> l_igual$stderr
[1] 754.4488
>
> d = 6384.682 - 6072.423
> d
[1] 312.259
> d/754.4488
[1] 0.4138902
```

Equivalências entre MQO e teste t de diferença de médias: NÃO PAREADO

Um modelo **MQO** produz erro padrão e estatística t da dummy correspondente ao grupo **equivalentes** ao erro padrão e estatística t produzidos por um **teste t não pareado** em que se assume **variância constante entre grupos**. Um modelo MQO com erros padrão robustos à heteroscedasticidade produz erro padrão e estatística t da dummy correspondente ao grupo **equivalentes ao teste de diferença de médias em se assume variância diferente entre grupos.**

Bailey (2016: 262)



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

R. Usando dummies.

Equivalências entre MQO e teste t de diferença de médias: NÃO PAREADO

```
> l_desigual = t.test(dados$price[dados$foreign==1],
           dados$price[dados$foreign==0],
           paired = FALSE,
           var.equal = FALSE)
> l_desigual
Welch Two Sample t-test

data:  dados$price[dados$foreign == 1] and
dados$price[dados$foreign == 0]
t = 0.44296, df = 46.447, p-value = 0.6599
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1106.339  1730.856
sample estimates:
mean of x mean of y
 6384.682  6072.423
> l_desigual$stderr
[1] 704.9376
> d = 6384.682 - 6072.423
> d
[1] 312.259

> l_igual = t.test(dados$price[dados$foreign==1],
                   dados$price[dados$foreign==0],
                   paired = FALSE,
                   var.equal = TRUE)
> l_igual
                Two Sample t-test

data:  dados$price[dados$foreign == 1] and
dados$price[dados$foreign == 0]
t = 0.41389, df = 72, p-value = 0.6802
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1191.708  1816.225
sample estimates:
mean of x mean of y
 6384.682  6072.423

> l_igual$stderr
[1] 754.4488
>
> d = 6384.682 - 6072.423
> d
[1] 312.259
```

```
> reg_regular = lm(price ~ foreign, data = dados)
> summary(reg_regular)

Call:
lm(formula = price ~ foreign, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-2781.4 -1885.6 -1160.4   259.8  9833.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6072.4      411.4  14.762   <2e-16 ***
foreign        312.3      754.4   0.414     0.68
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2966 on 72 degrees of freedom
Multiple R-squared:  0.002374,Adjusted R-squared:  -0.01148
F-statistic: 0.1713 on 1 and 72 DF,  p-value: 0.6802


> coeftest(reg_regular, vcov = vcovHC(reg_regular, type = "HC2"))

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6072.42     429.49  14.139   <2e-16 ***
foreign       312.26     704.94   0.443   0.6591
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[FIM DO] APÊNDICE:
Testes t de diferença de médias

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Variável dummy e interação

Aula 7
26 de outubro de 2022

Ana Paula Karruz