PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Aprofundamento em regressão multivariada

Aula 6
19 de outubro de 2022

Ana Paula Karruz

Agenda para esta aula e a próxima

1. **Multicolinearidade**
2. Heteroscedasticidade
3. Viés de variável omitida
4. Variável irrelevante
5. Variável dummy
6. Interação

As propriedades mais bacanas de MQO

Veja, propriedade é diferente de premissa

*Se as premissas de MQO forem atendidas, MQO é **BLUE**!*

- ***Best** (variância mínima, i.e., máxima precisão)*
- ***Linear***
- ***Unbiased** (livre de vieses)*
- ***Estimator***

- Se as premissas de MQO estiverem atendidas, MQO é “melhor” em relação às alternativas, quais sejam: adaptações de MQO (e.g., Mínimos Quadrados Generalizados) e estimadores de Máxima Verossimilhança (um algoritmo iterativo que permite não linearidades nos βs)

- As propriedades BLUE do MQO são provadas pelo teorema de Gauss-Markov

Para que o estimador MQO seja BLUE, é preciso atender às chamadas "premissas clássicas"

Premissas de MQO

- Não há correlação entre cada uma das variáveis explicativas e o termo de erro – esta é a premissa de **exogeneidade: correlação (X_j, ε) = 0**
- As variáveis explicativas não são uma função linear das outras (i.e., **não há multicolinearidade**)
- O termo de erro tem variância constante (i.e., é **não há heteroscedasticidade**)
- Os erros não são correlacionados entre si (i.e., **não existe autocorrelação** serial ou espacial)
- O modelo de regressão é **linear nos parâmetros** (i.e., coeficientes são adicionados e com expoente = 1)
- **ε**, o termo de erro aleatório, tem **média populacional = 0**
- O **erro** tem distribuição **normal**: ε ~ N(0, σ)

O que eventualmente ficou de fora não é correlacionado com as variáveis incluídas – se for, teremos o problema do viés de variável omitida

Já falamos sobre isso, e falaremos mais!

Premissa requerida para teste de hipótese e intervalo de confiança, não para estimação de MQO

Para que o estimador MQO seja BLUE, é preciso atender às chamadas "premissas clássicas"

Premissas de MQO

- Não há correlação entre cada uma das variáveis explicativas e o termo de erro – esta é a premissa de **exogeneidade: correlação (X_j, ε) = 0**
- As variáveis explicativas não são uma função linear das outras (i.e., **não há multicolinearidade**)
- O termo de erro tem variância constante (i.e., é **não há heteroscedasticidade**)
- Os erros não são correlacionados entre si (i.e., **não existe autocorrelação** serial ou espacial)
- O modelo de regressão é **linear nos parâmetros** (i.e., coeficientes são adicionados e com expoente = 1)
- **ε**, o termo de erro aleatório, tem **média populacional = 0**
- O **erro** tem distribuição **normal**: ε ~ N(0, σ)

O que eventualmente ficou de fora não é correlacionado com as variáveis incluídas – se for, teremos o problema do viés de variável omitida

Já falamos sobre isso, e falaremos mais!

Premissa requerida para teste de hipótese e intervalo de confiança, não para estimação de MQO

Temos multicolinearidade quando há associação linear entre variáveis independentes

- É **improvável** que encontremos variáveis explicativas **ortogonais** (i.e., que apresentem correlação = 0); portanto, conviveremos com algum grau de multicolinearidade

- **Não há um limiar** para definir o que é um **nível "aceitável"** ou **"preocupante"** de multicolinearidade

- Mas é possível medir **quanto** a correlação existente **infla a var($\beta_j$hat)**, em comparação com um **cenário de zero correlação** entre variáveis explicativas; fazemos isso com base no **variance inflation factor (VIF)**

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

VIF é parte da fórmula da variância do coeficiente de inclinação estimado na regressão múltipla

$$var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{n * var(X_j) * (1 - R_j^2)}$$

$$var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{n * var(X_j)} * \frac{1}{(1 - R_j^2)}$$

Fórmula da var(B_1hat) na regressão simples

VIF

VIF no R

```
> if (! "haven" %in% installed.packages()) install.packages("haven", dep = T) # for reading .dta files
> if (! "car" %in% installed.packages()) install.packages("car", dep = T) # for vif e lht (F test)
> library(haven)
> library(car)

> dados <- read_dta('auto.dta')
> dados = as.data.frame(dados)

> reg_main = lm(price ~ mpg + trunk + foreign, data = dados)
> summary(reg_main)

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

Residuals:
    Min      1Q  Median      3Q     Max
-3461.1 -1704.2  -873.2  1014.3 10287.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10033.08    2256.68   4.446 3.21e-05 ***
mpg          -261.99      64.91  -4.036 0.000137 ***
trunk          83.65      86.50   0.967 0.336871
foreign  1887.46     711.42   2.653 0.009861 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2532 on 70 degrees of freedom
Multiple R-squared:  0.2933,        Adjusted R-squared:  0.263
F-statistic: 9.683 on 3 and 70 DF,  p-value: 1.996e-05

> vif(reg_main)
       mpg      trunk foreign
  1.605833   1.558684   1.220334
```

Como interpretar os VIFs?

Comumente, a interpretação de VIFs baseia-se em regras de bolso. O'Brien (2007, 684-685) recomenda que sejam considerados outros fatores capazes de influenciar a variância das estimativas:

Regras de bolso *para valores de VIF têm aparecido na literatura:* ***regra do 4, regra do 10, etc****. Segundo essas regras, quando o VIF excede esses valores, a multicolinearidade é considerada muito alta e* ***pairam dúvidas sobre os resultados da análise de regressão****.*
[...]
Demonstramos que as regras práticas associadas ao VIF (e à tolerância*) precisam ser interpretadas no contexto de outros fatores [e.g., tamanho da amostra, variância da variável independente] que influenciam a estabilidade das estimativas do coeficiente de regressão em questão*. *Esses efeitos podem facilmente reduzir a variância dos coeficientes de regressão muito mais do que o VIF infla essas estimativas, mesmo quando o VIF é 10, 20, 40 ou mais. É importante ressaltar que* ***a preocupação com os efeitos da inflação de variância é diferente em situações em que rejeitamos a hipótese nula [...] em relação às situações em que a hipótese nula não é rejeitada*** *[...]. No primeiro caso, encontramos um resultado estatisticamente significativo, [...] mesmo com inflação da variância. No segundo caso, podemos ter sido prejudicados pelo aumento da variância associada ao coeficiente de regressão.*

Referência

O'Brien, R. M., 2007. A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), pp. 673-690. https://doi.org/10.1007/s11135-006-9018-6

* Tolerância = 1/VIF.

Multicolinearidade não causa viés, e não exige correção da var($\beta_j$hat)

É muito importante entender o que a multicolinearidade faz. ***Não causa viés. Nem mesmo faz com que os erros padrão de $\beta_1$hat*** *[ou, genericamente falando, de $\beta_j$hat]* ***sejam incorretos****. Simplesmente faz com que os erros padrão sejam maiores do que seriam se não houvesse multicolinearidade. Em outras palavras, [em caso de multicolinearidade] o MQO [...produz] estimativas não enviesadas e com a incerteza [i.e., o erro padrão] adequadamente calculada. [A consequência da multicolinearidade é que] quando as variáveis [explicativas] estão fortemente relacionadas entre si, teremos mais incerteza – as distribuições de $\beta_1$hat [ou, genericamente falando, de $\beta_j$hat] serão mais dispersas, o que significa que será mais difícil aprender com os dados.*

Bailey (2016: 228)

Multicolinearidade não parece ser um grande problema. Ainda assim, devemos fazer algo sobre ela? Depende.

- Se a var($\beta_j$hat) for pequena, será **possível distinguir o efeito de diferentes variáveis explicativas** – neste caso, **deixe estar**; exemplos:
  - Table 5.2 (Bailey, 2016: 201): corr(adult height, adolescente height) = 0,86
  - Table 5.3 (Bailey, 2016: 216): corr(years of school, test score) = 0,81
- Se a **multicolinearidade for substancial**, não sendo possível distinguir o efeito de diferentes variáveis explicativas, **conduza o test F de múltiplas restrições** e apresente seus resultados
  - O teste F de múltiplas restrições indicará **se**, **tomadas em conjunto**, as **variáveis colineares parecem importar para a explicação de Y**, ainda que não possamos apurar os efeitos individuais dessas variáveis

Não caia na tentação de descartar (drop) uma das variáveis colineares: se você tinha uma boa razão teórica para ter essas variáveis na equação, faltará uma boa razão teórica para remover uma delas.

Vide APÊNDICE:
Testes F

Dose letal

*Uma dose letal de multicolinearidade é chamada de **multicolinearidade perfeita**, que ocorre quando **uma variável independente é completamente explicada por outras variáveis independentes**. Se isso acontecer, $R_j^2 = 1$ e var($\beta_j$hat) não pode ser calculada, pois tem $(1 - R_j^2)$ no denominador (no sentido de que o denominador se torna zero, o que causa indefinição). Nesse caso, o software estatístico se recusará a estimar o modelo ou excluirá [will drop] automaticamente variáveis independentes até que não haja multicolinearidade perfeita. Um exemplo bobo de multicolinearidade perfeita é quando alguém inclui a mesma variável duas vezes em um modelo.*

Bailey (2016: 230)

Ou quando incluímos dummies para todas as categorias possíveis (e.g., uma dummy para brasileiro e outra para estrangeiro no mesmo modelo)

Agenda para esta aula e a próxima

1. Multicolinearidade
2. **Heteroscedasticidade**
3. Viés de variável omitida
4. Variável irrelevante
5. Variável dummy
6. Interação

Para que o estimador MQO seja BLUE, é preciso atender às chamadas "premissas clássicas"

Premissas de MQO

O que eventualmente ficou de fora não é correlacionado com as variáveis incluídas – se for, teremos o problema do viés de variável omitida

Já falamos sobre isso, e falaremos mais!

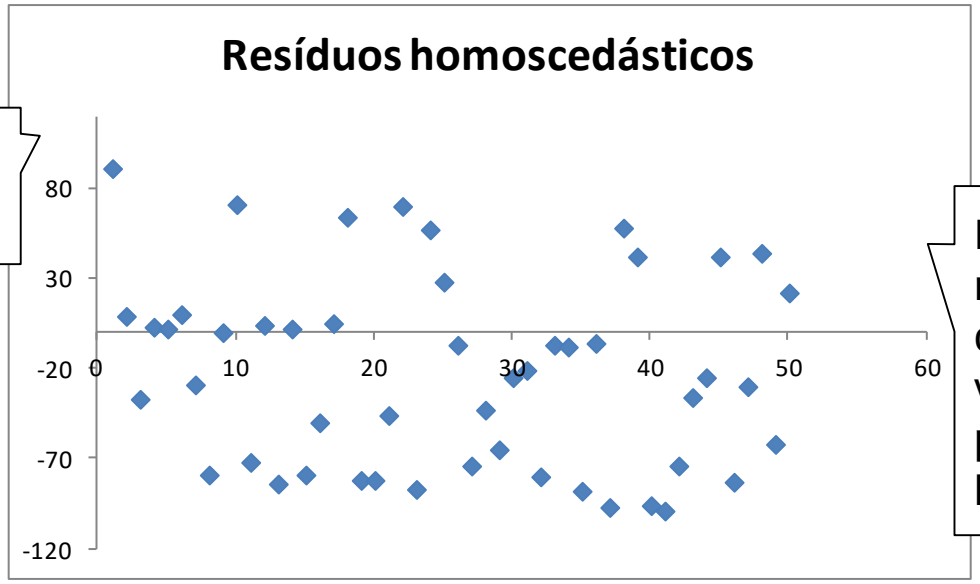
- Não há correlação entre cada uma das variáveis explicativas e o termo de erro – esta é a premissa de **exogeneidade: correlação (X_j, ε) = 0**
- As variáveis explicativas não são uma função linear das outras (i.e., **não há multicolinearidade**)
- O termo de erro tem variância constante (i.e., é **não há heteroscedasticidade**)
- Os erros não são correlacionados entre si (i.e., **não existe autocorrelação** serial ou espacial)
- O modelo de regressão é **linear nos parâmetros** (i.e., coeficientes são adicionados e com expoente = 1)

Premissa requerida para teste de hipótese e intervalo de confiança, não para estimação de MQO

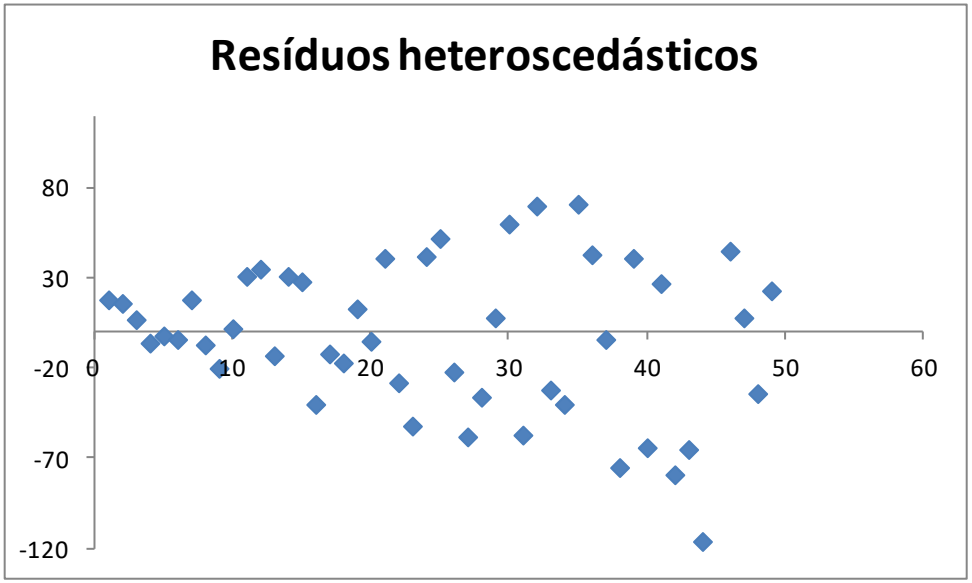
- **ε**, o termo de erro aleatório, tem **média populacional = 0**
- O **erro** tem distribuição **normal**: ε ~ N(0, σ)

Ilustração: Homoscedasticidade vs. Heteroscedasticidade

Eixo vertical representa os resíduos



Eixo horizontal representa variável X ou Y (qualquer variável do modelo pode ser fonte de heteroscedasticidade)



Heteroscedasticidade

Pode ocorrer em regressões simples ou múltiplas

A heteroscedasticidade é geralmente causada por variáveis que assumem valores altos, as quais se associam com erros altos. Exemplos: Pessoas de alta renda podem ter maior variância em seu consumo; escolas grandes podem ter maior variância na nota dos estudantes em testes de desempenho

Implicação

- Tende a subestimar erro padrão, portanto:
  - Tende a aumentar artificialmente a precisão das estimativas (e a estreitar os intervalos de confiança)
  - Pode induzir-nos a "encontrar" significância estatística onde ela não existe

Heteroscedasticidade não causa viés, porém requer que cálculo do EP(βhat) seja ajustado

Detecção

- Examine os resíduos plotados contra a variável suspeita de causar heteroscedasticidade – esta abordagem só funciona em casos óbvios/ dramáticos
- Use um teste estatístico em que a H_0 = não existe heteroscedasticidade (e.g., teste Breusch-Pagan, Park, White)

Mitigação

- Use erros padrão robustos (heteroscedasticity-consistent standard errors)
- Transforme as variáveis que possam estar causando heteroscedasticidade (e.g., logaritmo, per capita, razão)

Calculando erros padrão robustos à heteroscedasticidade (exemplo com dados womenlabor.csv)

- Estime a regressão (como de costume); por exemplo:
  `spec1 <- lm(my_formula, data = womenlabor)`

- Calcule erros padrão robustos à heteroscedasticidade
  `coeftest(spec1, vcov = vcovHC(spec1, "HC1"))`

`vcov` matriz de variância-covariância do estimador de MQO; em sua forma matricial, OLS calcula os erros padrão a partir dessa matriz, cuja diagonal principal contém var($\beta_j$hat)

$$\begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \vdots & \vdots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \text{cov}(\hat{\beta}_k, \hat{\beta}_2) & \dots & \text{var}(\hat{\beta}_k) \end{bmatrix}$$

`vcovHC` Função que calcula a vcov para erros padrão robustos (heteroscedasticity-consistent standard errors)

`type` Tipo de ajuste aplicado; vide documentação da função vcovHC no pacote `sandwich`; neste exemplo, usamos `type = "HC1"`

- Considere os erros padrão robustos à heteroscedasticidade para os testes de significância de β_0 e $\beta_j$hat

Calculando erros padrão robustos à heteroscedasticidade (exemplo com dados womenlabor.csv)

```
> colnames(womenlabor)
 [1] "wlfp"    "yf"      "ym"      "educ"    "ue"      "mr"
"dr"      "urb"     "wh"      "d90"     "stateid"
>
> indpvars = c(names(womenlabor))
> indpvars = indpvars[-c(1, 10:11)]
> indpvars
[1] "yf"   "ym"   "educ" "ue"   "mr"   "dr"   "urb"  "wh"
> indpvars = (paste(indpvars, collapse = ' + '))
> indpvars
[1] "yf + ym + educ + ue + mr + dr + urb + wh"
>
> my_formula = c(paste("wlfp ~ ", indpvars))
> class(my_formula)
[1] "character"
> my_formula = as.formula(my_formula)
> class(my_formula)
[1] "formula"
> my_formula
wlfp ~ yf + ym + educ + ue + mr + dr + urb + wh
>
> spec1 <- lm(my_formula, data = womenlabor)
```

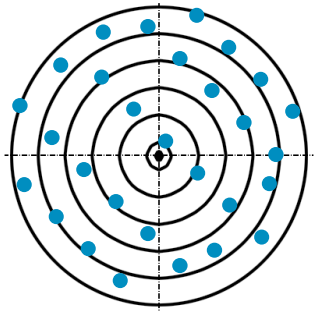
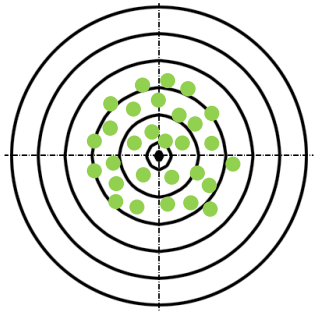
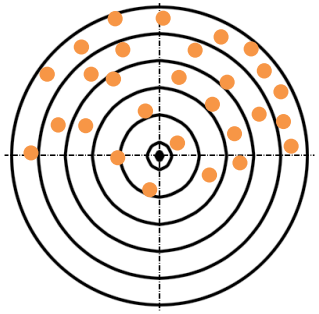
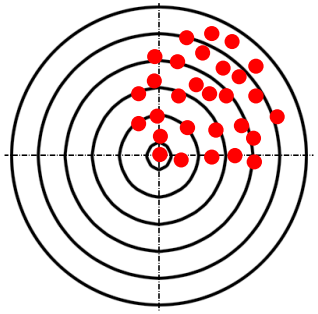
```
> summary(spec1)
Call:
lm(formula = my_formula, data = womenlabor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5874 -1.7729 -0.2061  1.7253  8.5357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.7767655  8.4815645   6.930 5.88e-10 ***
yf           0.0004277  0.0004080   1.048 0.297292
ym          -0.0003552  0.0003104  -1.144 0.255502
educ         0.3973625  0.0570042   6.971 4.87e-10 ***
ue          -0.9404541  0.2536938  -3.707 0.000360 ***
mr          -0.3109530  0.1329970  -2.338 0.021577 *
dr           0.1776552  0.1739730   1.021 0.309883
urb         -0.0108626  0.0243890  -0.445 0.657096
wh          -0.1158179  0.0311566  -3.717 0.000348 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.835 on 91 degrees of freedom
Multiple R-squared:  0.7575,    Adjusted R-squared:  0.7362
F-statistic: 35.53 on 8 and 91 DF,  p-value: < 2.2e-16

> # Calculando erros padrao robustos
> coeftest(spec1, vcov = vcovHC(spec1, "HC1")) # HC1 = default
type of robust standar error in Stata 16

t test of coefficients:

             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 58.77676548 10.80428819  5.4401 4.455e-07 ***
yf           0.00042774  0.00050230  0.8516  0.396694
ym          -0.00035524  0.00034710 -1.0234  0.308809
educ         0.39736253  0.06486707  6.1258 2.266e-08 ***
ue          -0.94045409  0.34945774 -2.6912  0.008473 **
mr          -0.31095296  0.16266547 -1.9116  0.059073 .
dr           0.17765515  0.23533805  0.7549  0.452262
urb         -0.01086261  0.02775466 -0.3914  0.696432
wh          -0.11581789  0.03610629 -3.2077  0.001848 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cálculo de erros padrão robustos à heteroscedasticidade não altera coeficientes estimados (betas)

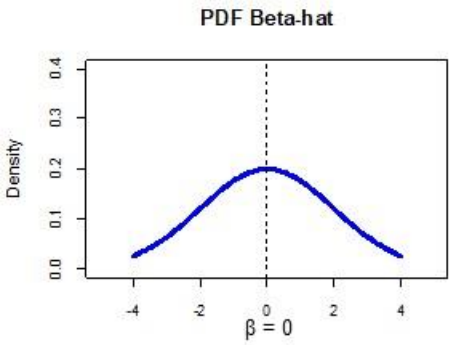
Agenda para esta aula e a próxima

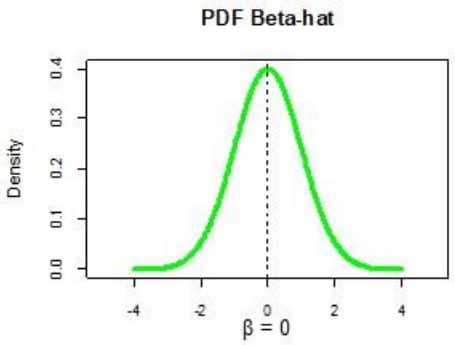
1. Multicolinearidade
2. Heteroscedasticidade
3. **Viés de variável omitida**
4. Variável irrelevante
5. Variável dummy
6. Interação

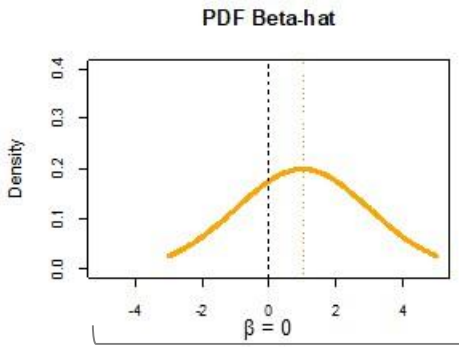
Acurácia, precisão e a distribuição teórica de βhats

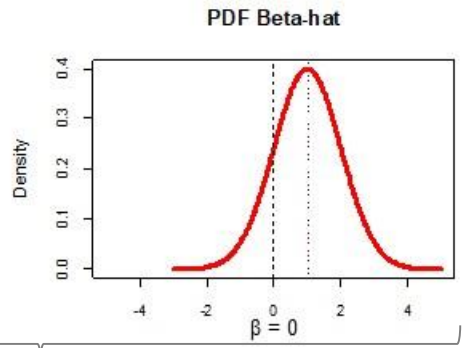





Acurácia (ausência de viés)	✓	✓		
Precisão (baixa dispersão)		✓		✓









Estes cenários mostram viés positivo (centro da distribuição de βhats está à direita do verdadeiro β). Viés negativo é igualmente prejudicial à acurácia do modelo

Dois desafios à análise estatística inferencial: aleatoriedade e endogeneidade

Fontes de incerteza quanto ao efeito estimado de X sobre Y

Sampling randomness: amostras de diferentes tamanhos geram coeficientes estimados diferentes; amostras diferentes de um mesmo tamanho também geram coeficientes estimados diferentes; na estatística frequentista, coeficiente populacional é fixo)

Modeled randomness: aleatoriedade e complexidade na formação de Y redundam em variáveis omitidas; nota: aqui não estamos falando de variáveis omitidas correlacionadas com X

Aleatoriedade (compromete a **precisão**)

$\hat{\beta}_j$

(qualquer coeficiente de inclinação estimado)

Como a regressão múltipla reduz risco de viés (de variável omitida)?

Variáveis omitidas correlacionadas com X: existência dessas variáveis implica espuriedade

Endogeneidade (compromete a **acurácia**)

E se as pessoas altas comerem mais donuts? A altura está no termo de erro como um fator que contribui para o peso, e se as pessoas altas comem mais donuts, podemos atribuir erroneamente aos donuts o efeito da altura.

Bailey (2016: 14)

Endogeneidade e viés de variável omitida

- O modelo deveria ser:

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 m_i + \varepsilon_i$$

- Onde Y = Rendimento, T = Exposição ao tratamento (participação voluntária em programa de qualificação profissional), M = Motivação do trabalhador

- Todavia, o modelo estimado foi:

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i$$

- A exposição ao tratamento é provavelmente determinada por fatores que também causam Y (e.g., M), mas que estão omitidos na regressão; no modelo que omite M, T é considerada uma **variável endógena**, porque ela é correlacionada com o termo de erro

- Modelo não consegue separar efeito de T do efeito de M; o estimador de β_1 produz uma combinação desses efeitos

- Como consequência, o estimador do efeito de T (i.e., o estimador de β_1) está carregando o efeito de T mas também de M; é como se $\beta_1$hat "absorvesse" parte do efeito de M. Assim, a E($\beta_1$hat) se afasta do verdadeiro β_1, o que caracteriza o viés

Efeito sistemático das variáveis omitidas é capturado pelos coeficientes estimados

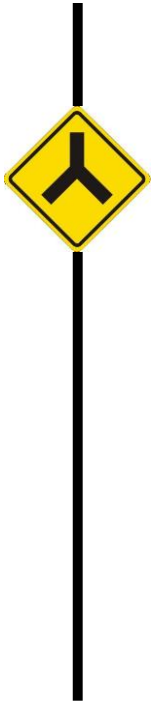
Numa conversa mais estruturada....

- Na **equação populacional "completa"**, **ε** carrega apenas o efeito não sistemático de variáveis omitidas; tudo que afeta sistematicamente o Y aparece como variável explicativa

- Na **equação populacional "simplificada"**, aquela que realmente estimamos, nem todos os fatores formadores de Y estão presentes como variáveis explicativas; **ε** carrega o efeito médio das variáveis omitidas
  - **Premissa:** Variáveis omitidas correlacionam-se com Y, porém não com X, i.e., **corr(ε, X) = 0**
  - **Premissa: E(ε) = 0**, portanto ε não precisa ser "estimado", e não aparece na equação ajustada

- **Para onde vai, então, o efeito sistemático médio das variáveis omitidas?**

Variáveis omitidas **não correlacionadas com X** têm seu efeito médio atribuído a **$\beta_0$hat**, afetando o nível de Yhat (intercepto estimado), mas não o efeito de X sobre Y (inclinação da reta ajustada, ou seja, efeito estimado de X em Y)



Variáveis omitidas **correlacionadas com X violam premissa de MQO,** têm seu efeito parcialmente absorvido pelo estimador do **βhat** de inclinação respectivo, **e o enviesam**

- O modelo não conseguirá distinguir o efeito de X sobre Y do efeito de Z sobre Y, e **atribuirá erroneamente a X parte do efeito de Z**
- Dada essa **afinidade (correlação)** entre X e Z, é como se o βhat **"atraísse"** parte do efeito médio das variáveis omitidas (que, na ausência de correlação entre X e Z, teria sido totalmente incorporado pelo estimador de $\beta_0$hat)

Se corr(ε, X) ≠ 0, então X é considerado uma variável endógena

Um variável independente é exógena se sua variação não estiver relacionada a fatores embutidos no ε

Exogeneidade é o oposto de endogeneidade

"exo" = externo; variável está fora do modelo no sentido de que não se correlacionada com outros fatores que influenciam Y

Exogeneidade: **corr(X,ε) = 0**

"endo" = interno; variável está dentro do modelo no sentido de que se correlaciona com outros fatores que influenciam Y

Endogeneidade: **corr(X,ε) ≠ 0**

Lembrete
Ordem das variáveis não altera correlação: corr(X,ε) = corr(ε,X)

Estatisticamente falando, destacamos esse grande desafio ao dizer que a variável donut é endógena. ***Uma variável independente é endógena se as mudanças nela estiverem relacionadas a fatores no termo de erro. […] A endogeneidade está em toda parte; é endêmica.***

Bailey (2016: 14-15)

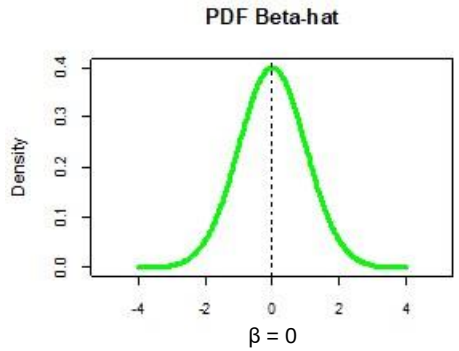
Se corr(ε, X) ≠ 0, então X é considerado uma variável endógena

Endogeneidade faz MQO produzir um estimador enviesado do verdadeiro β de inclinação

Exemplo de viés: Distribuição de βhat de inclinação não está centrada no verdadeiro β

Ausência de viés = acurácia (i.e., estimador não tenderá a super ou subestimar β)

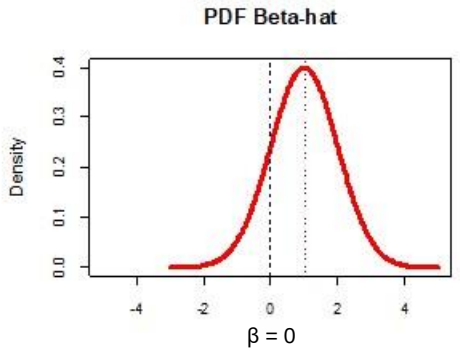
- E(βhat) = β
- Distribuição de βhat está centrada no verdadeiro valor de β



- Good news: Na maioria dos casos, estimador não enviesado produzirá "bom" βhat
- Bad news: Estimador não enviesado pode produzir βhat bem distante de β

Viés = "inacurácia" (i.e., estimador tenderá a super ou subestimar β)

- E(βhat) **≠** β
- Distribuição de βhat **não** está centrada no verdadeiro valor de β



- Modelo não consegue separar efeito de X do efeito de Z e produz βhat que é uma combinação desses efeitos. Exemplo:

$$Violent\ crime_t = \beta_0 + \beta_1 Ice\ cream\ sales_t + \epsilon_t$$

Uma regressão simples de crimes violentos e venda de sorvetes provavelmente captará uma associação entre essas variáveis; todavia, essa associação é espúria

Endogeneidade e viés de variável omitida

- O modelo deveria ser:

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 m_i + \varepsilon_i$$

- Onde Y = Rendimento, T = Exposição ao tratamento (participação voluntária em programa de qualificação profissional), M = Motivação do trabalhador

- Todavia, o modelo estimado foi:

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i$$

- A exposição ao tratamento é provavelmente determinada por fatores que também causam Y (e.g., M), mas que estão omitidos na regressão; no modelo que omite M, T é considerada uma **variável endógena**, porque ela é correlacionada com o termo de erro

- Modelo não consegue separar efeito de T do efeito de M; o estimador de β_1 produz uma combinação desses efeitos

- Como consequência, o estimador do efeito de T (i.e., o estimador de β_1) está carregando o efeito de T mas também de M; é como se $\beta_1$hat "absorvesse" parte do efeito de M. Assim, a E($\beta_1$hat) se afasta do verdadeiro β_1, o que caracteriza o viés

Determinando a direção do viés

- **O problema é:** $E(\hat{\beta}_1) \neq \beta_1$

- **Qual a direção do viés?**
  - Positivo ("para cima"): $E(\hat{\beta}_1) > \beta_1$
  - Negativo ("para baixo"): $E(\hat{\beta}_1) < \beta_1$

- **Como determinar a direção do viés? Através de uma multiplicação de sinais**

Sinal do viés = Sinal do β_{om} * Sinal de *f*(correlação$_{in,\ om}$)

  - **B_{om}** é o efeito (não observável) da variável omitida sobre Y
  - ***f*(correlação$_{in,\ om}$)** é uma função da correlação entre a variável incluída e a omitida (T e M, no exemplo)

Endogeneidade e viés de variável omitida

- O modelo deveria ser:

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 m_i + \varepsilon_i$$

- Onde Y = Rendimento, T = Exposição ao tratamento (participação voluntária em programa de qualificação profissional), M = Motivação do trabalhador

- Todavia, o modelo estimado foi:

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i$$

- **Como determinar a direção do viés? Através de uma multiplicação de sinais**

Sinal do viés = Sinal do β_{om} * Sinal de *f*(correlação$_{in,\ om}$)

  - **B_{om}** é o efeito (não observável) da variável omitida (M) sobre Y
  - ***f*(correlação$_{in,\ om}$)** é uma função da correlação entre a variável incluída (T, para cujo coeficiente estamos analisando a direção do viés) e a variável omitida (M)

Sinal do viés = positivo * positivo
Sinal do viés = positivo

Estes sinais nos são desconhecidos; consideramos nossas hipóteses sobre eles

Conclusão: diante da omissão de M, estimador de β_1 tende a inflacionar o efeito do tratamento

Exemplo de remoção de viés: Educação ajuda o crescimento econômico?

Table 5.3: Economic Growth and Education Using Multiple Measures of Education

	Without math/science test scores
Avg. years of school	0.44*
	(0.10)
	[t = 4.22]
Math/science test scores	
GDP in 1960	-0.39*
	(0.08)
	[t = 5.19]
Constant	1.59*
	(0.54)
	[t = 2.93]
N	50
$\hat{\sigma}$	1.13
R^2	0.36

Standard errors in parentheses, * indicates significance at $p < 0.05$

Fonte: Bailey (2016: 216, 218).

Exemplo de remoção de viés: Educação ajuda o crescimento econômico?

Table 5.3: Economic Growth and Education Using Multiple Measures of Education

	Without math/science test scores	With math/science test scores
Avg. years of school	0.44*	0.02
	(0.10)	(0.08)
	[t = 4.22]	[t = 0.28]
Math/science test scores		1.97*
		(0.24)
		[t = 8.28]
GDP in 1960	-0.39*	-0.30*
	(0.08)	(0.05)
	[t = 5.19]	[t = 6.02]
Constant	1.59*	-4.76*
	(0.54)	(0.84)
	[t = 2.93]	[t = 5.66]
N	50	50
$\hat{\sigma}$	1.13	0.72
R^2	0.36	0.74

Standard errors in parentheses, * indicates significance at $p < 0.05$

Fonte: Bailey (2016: 216, 218).

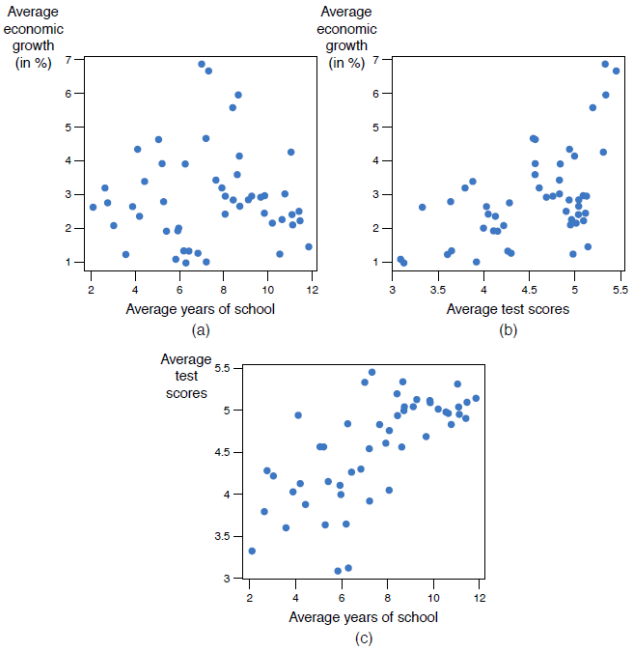


FIGURE 5.4: Economic Growth, Years of School, and Test Scores

Not all schooling is of equal quality

Exemplo de remoção de viés: Educação ajuda o crescimento econômico?

*Observe a **história muito diferente** que temos **nas duas colunas**. **Na primeira**, **anos de escolaridade são suficientes para o crescimento econômico**. **Na segunda** especificação, **a qualidade da educação** medida com os resultados dos testes de matemática e ciências **é mais importante**. A segunda especificação é melhor porque mostra que uma variável teoricamente sensata importa muito. **A exclusão dessa variável**, como faz a primeira especificação, expõe a análise a viés de variável omitida. Em suma, esses resultados sugerem que a educação é sobre qualidade, não quantidade.*

….

Esses resultados não encerram a conversa sobre educação e crescimento econômico, mas avançam alguns passos.

Bailey (2016: 219)

*Como a escala da variável de pontuação do teste não é imediatamente óbvia, precisamos trabalhar um pouco para **interpretar a significância substantiva da estimativa do coeficiente**. Com base na estatística descritiva (não relatada), o **desvio padrão** da variável pontuação do teste é de 0,61. Os resultados, portanto, implicam que o aumento das pontuações médias dos testes por um desvio padrão está associado a um aumento de **0,61 * 1,97 = 1,20** ponto percentual na taxa média de crescimento anual […] ao longo desses quarenta anos. Esse aumento é grande quando estamos falando de um crescimento composto ao longo de quarenta anos.*

Bailey (2016: 219)

Grosseiramente falando, o desvio padrão de uma variável corresponde ao seu desvio médio em relação à média

Agenda para esta aula e a próxima

1. Multicolinearidade
2. Heteroscedasticidade
3. Viés de variável omitida
4. **Variável irrelevante**
5. Variável dummy
6. Interação

Fórmula de var(βhat) de inclinação elucida o que acontece quando incluímos variáveis irrelevantes

São irrelevantes as variáveis que não pertencem à regressão populacional

- O modelo deveria ser:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

- Seu modelo é:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

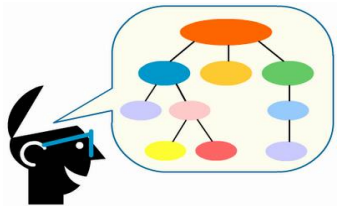
$$var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{n * var(X_j) * (1 - R_j^2)}$$

Implicações

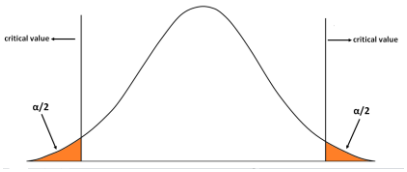
- **Acurácia:** A inclusão ou exclusão da variável irrelevante não causa viés de variável omitida, já que $\beta_2 = 0$ (vide fórmula do viés de variável omitida)

- **Precisão:** A consequência da adição de uma variável irrelevante é incerta:
  - Se a variável irrelevante se correlacionar substantivamente com Y, ainda que fortuitamente, diminuirá a variância da regressão (σ^2hat), reduzindo também a variância dos βhat de inclinação
  - Ao mesmo tempo, se a variável irrelevante estiver substantivamente correlacionada com as demais variáveis explicativas, R_j^2 será elevado e inflacionará a variância do βhat dessas variáveis

Evitando variáveis irrelevantes



- Comece pela **teoria**: existe uma razão teórica forte para incluir a variável?



- Execute o **teste t**: A variável é estatisticamente significante?

$\overline{R}^2$

- Observe o grau de ajuste do modelo: a inclusão da variável eleva o **R^2 ajustado**? Importante comparar especificações que **difiram apenas pela inclusão de uma variável** (potencialmente irrelevante)

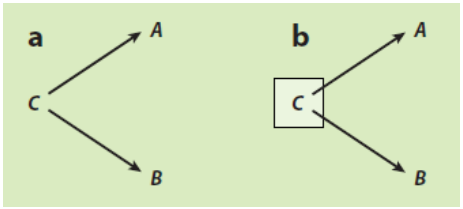
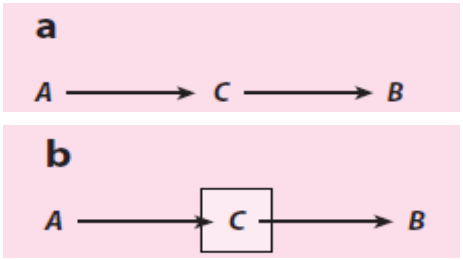
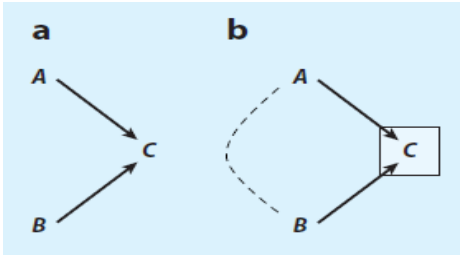
βββββββ

- Analise os **demais βs**: a inclusão da variável muda as magnitudes e, principalmente, os sinais dos demais coeficientes?

Se você respondeu **"Não" a todas essas questões**, a variável em tela pode ser considerada **irrelevante**

ATENÇÃO: Além do VVO, existem outros tipos de vieses. Em certos casos, adicionar controles causa viés*

NÃO EXAUSTIVO**

Tipo de viés	Exemplo abstrato		Exemplo real
Confounding bias VVO	a) A associação observada entre A e B não é causal; ela existe porque essas variáveis compartilham uma causa comum (C) b) Uma vez condicionada em C, a associação entre A e B desaparece **Recomendação:** Condicione em C		Hanushek e Woessmann (2009), sobre efeito da escolaridade no crescimento, se ignorada a qualidade da educação.
Overcontrol bias	a) A associação entre A e B é causal; C é variável intermediária nesse caminho causal b) Uma vez condicionada em C (i.e., mantendo-se C constante), a associação observada entre A e B desaparece **Recomendação:** Não condicione em C		Gratz (2019), sobre controlar pelo alcance educacional ao se estimar a relação entre origem social e rendimento do trabalho.
Endogenous selection bias	a) Não há associação causal entre A e B; C é uma variável de resultado dita "collider" (determinada por mais de uma variável) b) Ao condicionar-se em C, observa-se uma associação entre A e B, porém essa associação não é causal **Recomendação:** Não condicione em C		Lin, Schaeffer e Seltzer (1999), sobre efeito da renda no child support, com amostra de pais que responderam a survey.

Aprofun-damento

* Este slide é fortemente baseado em Elwert e Winship (2014).

** Um exemplo de viés não tratado aqui é o attenuation bias, causado por erro de mensuração em X.

What is measured with error?* If Y, OLS is OK. If X, then we will face attenuation bias

Erro de mensuração em Y: MQO é ok!

MQO funcionará bem se o erro de mensuração estiver apenas na variável dependente. Nesse caso, o ***erro de mensuração é simplesmente parte do termo de erro geral****. Quanto maior o erro [de mensuração], maior a variância do termo de erro.*

Bailey (2016: 220)

Se a variância do erro (i.e., a variância da regressão) crescer, então crescerá também o erro padrão dos βhat de inclinação

Erro de mensuração em X: Viés de atenuação (attenuation bias)

O truque aqui é pensar neste exemplo ***como um problema de variável omitida onde v_i [o erro de mensuração] é a variável omitida****. Não observamos o erro de mensuração diretamente, certo? Se pudéssemos observá-lo, ajustaríamos nossa medida de X_1. Então, o que fazemos é tratar o erro de mensuração como uma variável não observada que, por definição, devemos omitir e ver como essa forma particular de viés de variável omitida afeta o modelo.*

[…]

Bailey está falando sobre $\beta_1$hat, mas o mesmo é verdadeiro para outros coeficientes de inclinação

Referimo-nos a este exemplo particular de viés de variável omitida como viés de atenuação porque quando omitimos o termo de erro de medição do modelo, nossa estimativa $\beta_1$hat se desvia do valor verdadeiro por um fator multiplicativo entre zero e um*. Isso significa que $\beta_1$hat tenderá a estar mais próximo de zero do que deveria estar quando X_1 for medido com erro. Se o verdadeiro valor de β_1 for algum número positivo, tendemos a ver valores de $\beta_1$hat que são menores do que deveriam ser. Se o verdadeiro valor de β_1 for negativo, tendemos a ver valores de $\beta_1$hat maiores (significando mais próximos de zero) do que deveriam ser.*

Bailey (2016: 222-3)

Aprofun-damento

* Assume que o erro de mensuração (v_i) tem média zero e não está correlacionado com o valor verdadeiro (ou seja, a distância entre os valores reais e relatados não é sistematicamente diferente para diferentes níveis de X).

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Aprofundamento em regressão multivariada

Aula 6
19 de outubro de 2022

Ana Paula Karruz