



DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

**Viés de variável omitida.
Variáveis irrelevantes.**

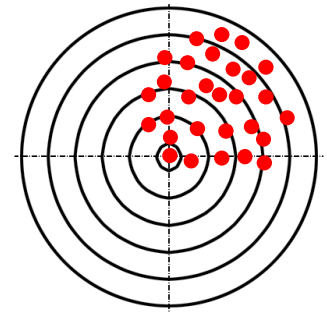
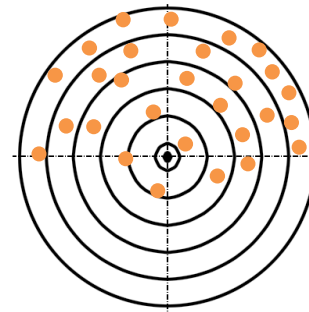
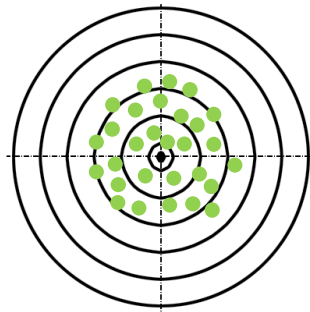
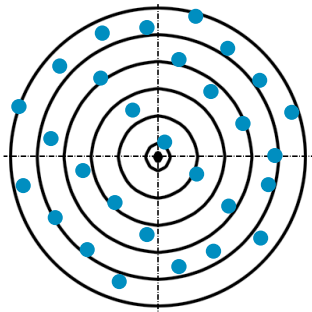
Aula 20
08 de junho de 2022

Ana Paula Karruz

Agenda

1. **Viés de variável omitida**
2. Variáveis irrelevantes

Acurácia, precisão e a distribuição teórica de β hats



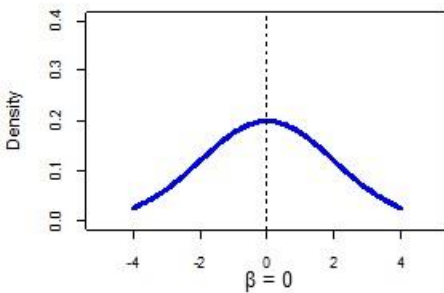
Acurácia
(ausência
de viés)



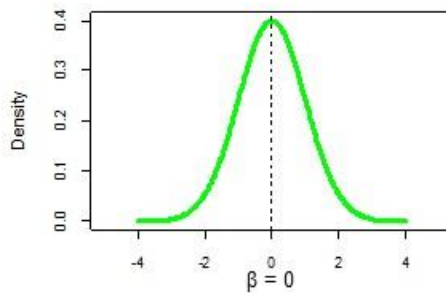
Precisão
(baixa
dispersão)



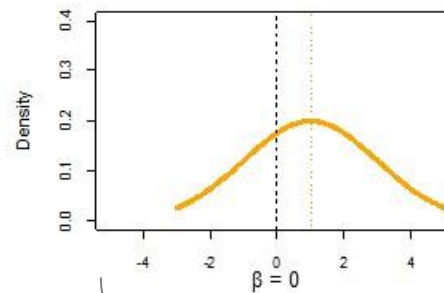
PDF Beta-hat



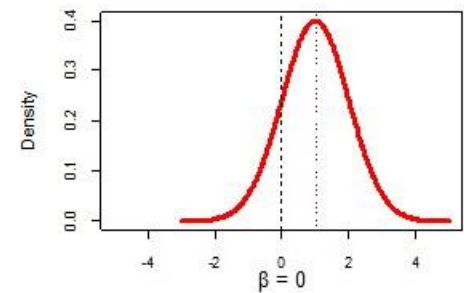
PDF Beta-hat



PDF Beta-hat



PDF Beta-hat



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

Dois desafios à análise estatística: 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

Variáveis omitidas correlacionadas com X: existência dessas variáveis implica espuriedade

Aleatoriedade
(compromete a
precisão)

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

Endogeneidade
(compromete a
acurácia)

$\hat{\beta}_1$

(ou qualquer outro
coeficiente de
inclinação estimado)

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 $\hat{\beta}_1$ “absorvesse” parte do efeito de M . Assim, a $E(\hat{\beta}_1)$ 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....

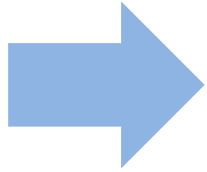
- Na **equação populacional “completa”**, ε carrega apenas o efeito não sistemático de variáveis omitidas; tudo que afeta sistematicamente o Y aparece como variável explicativa
- Na **equação populacional “simplificada”**, aquela que realmente estimamos, nem todos os fatores formadores de Y estão presentes como variáveis explicativas; ε carrega o efeito médio das variáveis omitidas
 - **Premissa:** Variáveis omitidas correlacionam-se com Y, porém não com X, i.e., **$\text{corr}(\varepsilon, X) = 0$**
 - **Premissa:** **$E(\varepsilon) = 0$** , 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 $\beta_0\text{hat}$, afetando o nível de \hat{Y} (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 $\beta_0\text{hat}$)



Se $\text{corr}(\varepsilon, X) \neq 0$, então X é considerado uma variável endógena

Um variável independente é exógena se sua variação não é 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 correlaciona com outros fatores que influenciam Y

Exogeneidade: $\text{corr}(X, \varepsilon) = 0$



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

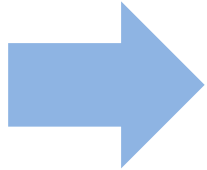
Endogeneidade: $\text{corr}(X, \varepsilon) \neq 0$

Lembrete

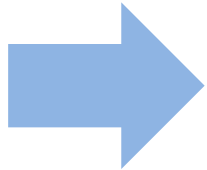
Ordem das variáveis não altera correlação: $\text{corr}(X, \varepsilon) = \text{corr}(\varepsilon, X)$

*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)₉



Se $\text{corr}(\varepsilon, X) \neq 0$, então X é considerado uma variável endógena

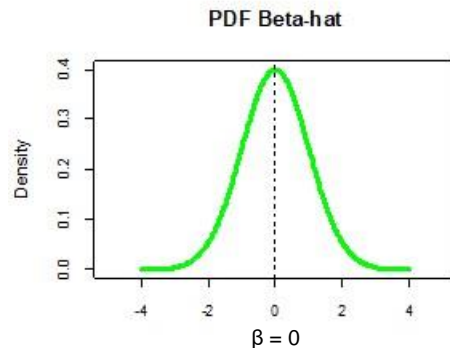


Endogeneidade faz MQO produzir um estimador enviesado do verdadeiro β de inclinação

Exemplo de viés: Distribuição de $\hat{\beta}$ de inclinação não está centrada no verdadeiro β

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

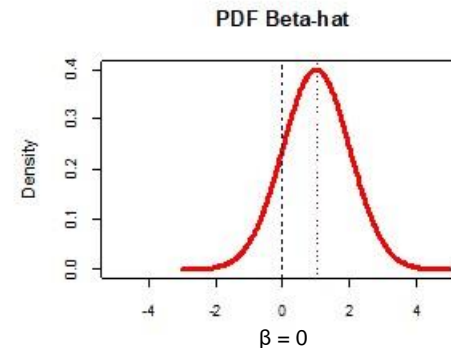
- $E(\hat{\beta}) = \beta$
- Distribuição de $\hat{\beta}$ está centrada no verdadeiro valor de β



- Good news: Na maioria dos casos, estimador não enviesado produzirá “bom” $\hat{\beta}$
- Bad news: Estimador não enviesado pode produzir $\hat{\beta}$ bem distante de β

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

- $E(\hat{\beta}) \neq \beta$
- Distribuição de $\hat{\beta}$ **não** está centrada no verdadeiro valor de β



- Modelo não consegue separar efeito de X do efeito de Z e produz $\hat{\beta}$ 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 $\hat{\beta}_1$ “absorvesse” parte do efeito de M . Assim, a $E(\hat{\beta}_1)$ 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

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

- β_{om} é o efeito (não observável) da variável omitida sobre Y
- $f(\text{correlação}_{in, om})$ 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**

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

- **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* (0.10) [t = 4.22]	0.02 (0.08) [t = 0.28]
Math/science test scores		1.97* (0.24) [t = 8.28]
GDP in 1960	-0.39* (0.08) [t = 5.19]	-0.30* (0.05) [t = 6.02]
Constant	1.59* (0.54) [t = 2.93]	-4.76* (0.84) [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 $p < 0.05$

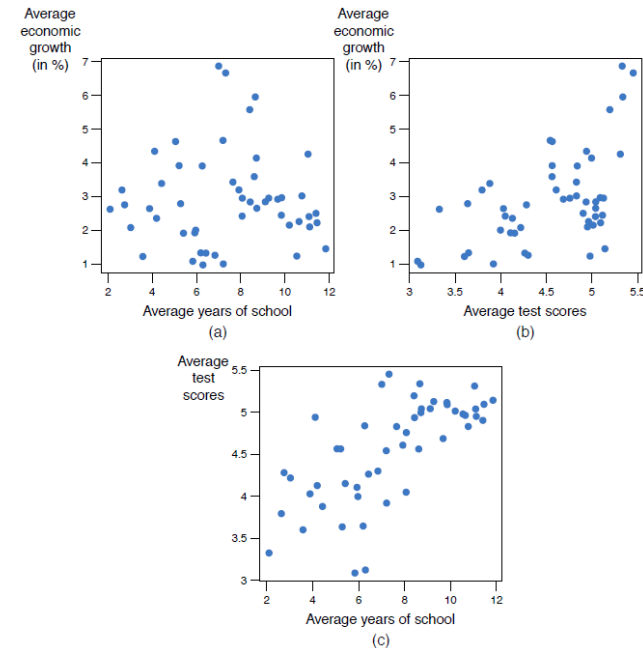


FIGURE 5.4: Economic Growth, Years of School, and Test Scores

Fonte: Bailey (2016: 216, 218).

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** pontos percentuais 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

1. Viés de variável omitida
2. **Variáveis irrelevantes**

Fórmula de $\text{var}(\hat{\beta})$ de inclinação elucidada o que acontece quando incluímos variáveis irrelevantes

São irrelevantes as variáveis que não pertencem à regressão populacional

Implicações

- O modelo deveria ser:

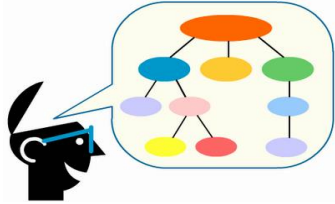
$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- Seu modelo é:

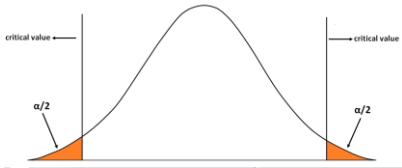
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- **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 ($\sigma^2_{\hat{\beta}}$), reduzindo também a variância dos $\hat{\beta}$ 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 $\hat{\beta}$ dessas variáveis

Evitando variáveis irrelevantes



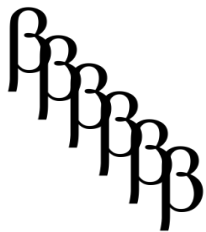
- Comece pela **teoria**: existe uma razão teórica forte para incluir a variável?



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

$$\overline{R}^2$$

- 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)



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

Aprofundamento

* 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 β_{hat} 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, ajustariamos 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 β_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 β_1 hat se desvia do valor verdadeiro por um fator multiplicativo entre zero e um. Isso significa que β_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 β_1 hat que são menores do que deveriam ser. Se o verdadeiro valor de β_1 for negativo, tendemos a ver valores de β_1 hat maiores (significando mais próximos de zero) do que deveriam ser.

Bailey (2016: 222-3)



DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

**Viés de variável omitida.
Variáveis irrelevantes.**

Aula 20
08 de junho de 2022

Ana Paula Karruz