


Econometria

Parte 4

Prof. Adalto Acir Althaus Junior oe

Sumário

- Causalidade
 - Endogeneidade
 - ✓ Viés de variável omitida
 - ✓ Viés de erro de medição
 - ✓ Viés de simultaneidade
 - Viés de seleção
- 

Motivação

- Como analistas e pesquisadores, estamos interessados em fazer declarações causais
 - ✓ Ex. #1 - qual é o efeito de uma mudança nos impostos das empresas sobre a escolha de alavancagem das empresas?
 - ✓ Ex. #2 - qual é o efeito de dar a um CEO mais participação acionária na empresa no desejo do CEO de assumir investimentos arriscados?
- Nós não gostamos de dizer apenas que as variáveis são "associadas" ou "correlacionadas" umas com as outras
- Na vida prática (mercado profissional), esse desejo de obter relações simples causais é ainda maior.

Motivação

- Lembre-se que se o nosso modelo linear é o seguinte ...

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- E, queremos inferir β_1 como o efeito causal de x_1 em y , mantendo todo o resto igual, então precisamos fazer as seguintes suposições ...

Hipóteses básicas

- Suposição #1: $E(u) = 0$
- Suposição #2: $E(u \mid x_1, \dots, x_k) = E(u)$
 - ✓ Em palavras, a média de u (ou seja, porção inexplicável de y) não depende do valor de x
- Isso é “independência média condicional” (CMI)
- De um modo geral, você precisa que o erro de estimativa não seja correlacionado com todos os x s

CMI vs Correlação

- CMI (que implica x e u não são correlacionados) é necessário para a imparcialidade [que é novamente uma propriedade de amostra finita]
- Mas, precisamos apenas assumir uma correlação zero entre x e u para consistência [que é uma propriedade de amostra grande]
- É por isso que normalmente apenas referimos se u e x estão ou não correlacionados no caso de avaliar se podemos fazer inferências causais

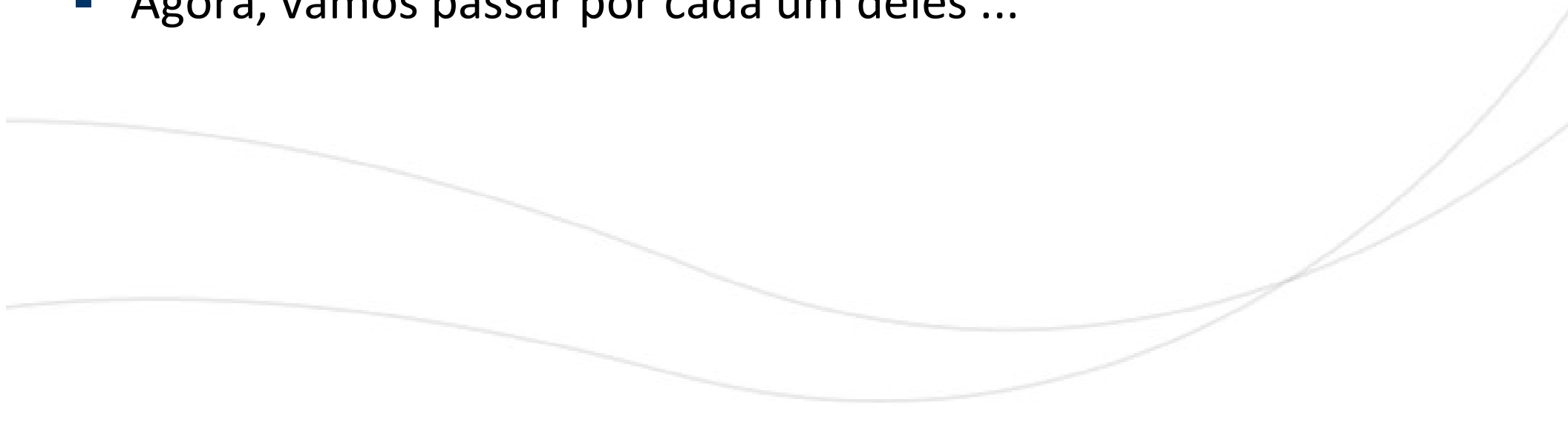
Cluster

- Erros padrão “clusterizados”
- Previne Multicolinearidade “localizada” nos dados
- Em estimações econométricas, um cluster significa que as observações dentro de um grupo “i” estão correlacionadas de alguma maneira desconhecida, induzindo correlação entre eles, mas que os grupos e não têm erros correlacionados.
- Ou seja, a correlação dos erros é apenas dentro do grupo
- Na presença de erros agrupados, as estimativas OLS ainda serão não visadas, mas os erros padrão podem estar bastante errados, levando a possíveis inferências incorretas em amostras finitas.
 - ✓ ..., vce(cluster variável grupo)

Erros Robustos

- Matriz Robusta de White – previne heterocedasticidade
- Como estimar $\text{Var}[\beta|X] = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$
- $\sigma^2X'\Omega X = \sigma^2 \sum_i \sum_j \omega_{ij} x_i x_j'$
- Formas de estimar:
 - ✓ The White estimator
 - ✓ Newey-West.

Violação da CMI

- Três maneiras principais como isso pode ser violado
 - Viés de variável omitida
 - Viés de erro de medição
 - Viés de simultaneidade
 - Agora, vamos passar por cada um deles ...
- 

Variável Omitida

- Provavelmente, é a preocupação mais comum que você vai ouvir dos pesquisadores, consultores e analistas de alto nível
- Ideia básica = o erro de estimativa, u , contém outra variável, por ex. z , que afeta y e está correlacionado com um x
- Por favor, **note que**:
 - ✓ A variável omitida é problemática **apenas** se correlacionada com algum x

Variável Omitida

- Você estima: $y = \beta_0 + \beta_1 x + u$
- Mas o modelo verdadeiro é: $y = \beta_0 + \beta_1 x + \beta_2 z + v$
- Então, $\beta_1 = \beta_1 + \delta_{xz}\beta_2$, onde δ_{xz} é o coeficiente que você obteria ao regredir a variável omitida, z , em x ; e

$$z = \delta_0 + \delta_{xz}x + w$$

$$\delta_{xz} = \frac{\text{cov}(x, z)}{\text{var}(x)}$$

Interpretando a variável Omitida

$$\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(x, z)}{\text{var}(x)} \beta_2$$

Effect of x on y

Regression of z on x

Effect of z on y

Bias

- Fácil notar que, o coeficiente estimado é não viesado apenas se:
 - ✓ $\text{cov}(x, z) = 0$ (ou seja, x e z não estão correlacionados)
- ou
- ✓ z não tem efeito em y , isto é, $\beta_2 = 0$

Direção do viés de variável Omitida

$$\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(x, z)}{\text{var}(x)} \beta_2$$

- Direção de viés dada por sinais de β_2 e $\text{cov}(x, z)$
 - ✓ Se você sabe que z tem efeito positivo em y (por exemplo, $\beta_2 > 0$) e x e z estão positivamente correlacionados ($\text{cov}(x, z) > 0$), então o viés será positivo
 - ✓ Magnitude do viés será dada por grandezas de β_2 e $\text{cov}(x, z) / \text{var}(x)$.

Direção do viés de variável Omitida

- Em geral não sabemos qual a magnitude do viés, mas podemos ter uma ideia da direção a partir dos sinais de β_2 e $cov(x, z)$.

	$Corr(x, z) > 0$	$Corr(x, z) < 0$
$\beta_2 > 0$	viés positivo	viés negativo
$\beta_2 < 0$	viés negativo	viés positivo

$$y = \beta_0 + \beta_1 x + \beta_2 z + v$$

viés $\beta_1 = \beta_1 + \delta_{xz}\beta_2$

não incluído na regressão

Exemplo

- Suponha que estimamos:

$$\ln(wage) = \beta_0 + \beta_1 educ + w$$

- Mas o verdadeiro modelo é:

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- Qual o viés em $\widehat{\beta}_1$?

$$\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(educ, ability)}{\text{var}(educ)} \beta_2$$

Exemplo

- Habilidade e salários provavelmente são correlacionados positivamente, então $\beta_2 > 0$
- Habilidade e educação provavelmente são correlacionados positivamente, então $cov(educ, ability) > 0$
- Assim, o viés é positivo!
 - ✓ $\widehat{\beta}_1$ é provavelmente muito grande!

Direção do viés de variável Omitida

- Caso GERAL:
- Quando o problema for além de apenas uma variável omitida, determinar o sinal (e a magnitude) do viés será muito mais difícil quando existem mais de uma....
 - ✓ Seja β vetor de coeficientes em k variáveis incluídas
 - ✓ Seja δ vetor de coeficiente em l variáveis excluídas
 - ✓ Seja X matriz de observações das variáveis incluídas
 - ✓ Seja Z matriz de observações das variáveis excluídas

$$\hat{\beta} = \beta + \frac{E[X'Z]}{E[X'X]} \delta$$

Direção do viés de variável Omitida

$$\hat{\beta} = \beta + \frac{E[X'Z]}{E[X'X]} \delta$$

Diagram illustrating the bias of omitted variables. The equation shows the estimated coefficient $\hat{\beta}$ as the sum of the true coefficient β and a bias term. An arrow points from the text "Vector of regression coefficients" to β . Another arrow points from the text "Vector of partial effects of excluded variables" to δ .

- A mesma ideia anterior, mas muito mais complicado no caso geral, com mais de uma variável omitida.

Eliminar viés de variável omitida

- Existem dois tipos de variável omitida
 - ✓ Observável e não incluída no modelo
 - ✓ Não observável

Eliminar viés de variável omitida

- Para lidar com o viés de variável omitida **observável**:
 - ✓ É simples, inclua tal variável no seu modelo!
- Um função mal especificada é um caso particular de variável omitida
 - ✓ Avalie a teoria subjacente, estude o assunto
 - ✓ Teste outras formas funcionais
 - ✓ Você precisa pensar bastante sobre como deveria ser a forma funcional correta!!!

Eliminar viés de variável omitida

- Para lidar com o viés de variável omitida **não observável**:
- Considere novamente a estimação

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- Problema: não podemos observar a habilidade... Nem medir
- Como resolver?
 - ✓ Solução possível: encontrar uma proxy que esteja correlacionada com a variável omitida, ou seja, para a habilidade nesse caso.

Eliminar viés de variável omitida

- Considere o seguinte caso

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- Vamos usar o QI como proxy para a habilidade. Que suposição precisamos fazer para utilizar isso? É plausível?

Resposta: Assumimos que $E(ability|educ, QI) = E(ability|QI)$, ou seja, a habilidade média não muda com a educação após a contabilizarmos o QI... Pode ser uma suposição questionável!

Erro de Medida

- **A estimativa terá erro de medida sempre que medirmos a variável de interesse de forma imprecisa**
 - ✓ Ex. #1: O taxa média de imposto é medida ruidosa da taxa de imposto marginal
- Este erro de medição pode causar viés e o viés pode ser bastante complicado

Erro de Medida vs Proxy

- Erro de medida é semelhante à variável proxy, mas muito diferente conceitualmente
 - ✓ Proxy é usado para algo que é totalmente inobservável ou mensurável (por exemplo, capacidade)
 - ✓ Com erro de medição, a variável que observamos é bem definida e pode ser quantificada... é apenas a nossa medida que contém erros

Erro de Medida na variável dep

- Geralmente não é um problema (em termos de viés); apenas faz com que nossos erros padrão sejam maiores.
 - ✓ Seja: $y^* = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$
- Mas nós medimos y^* com erro $e = y - y^*$
- Como nós observamos somente y , estimamos

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + (u + e)$$



nós sempre assumimos $E(e) = 0$; isso é inócuo pois caso seja falso, afeta apenas o viés da constante

Erro de Medida na variável dep

- Considerando $E(e|x) = 0$, as estimativas do OLS são consistentes e não viesadas
 - ✓ contanto que o erro de medição de y não esteja correlacionado com o x 's, estamos OK!
 - ✓ A única questão é que temos erros padrão maiores quando e e u não são correlacionados, (o que tipicamente se assume), já que $Var(u + e) > Var(u)$
- Quais são alguns exemplos comuns de Erro de Medida?

Erro de Medida na variável dep

- Alguns exemplos comuns
 - ✓ Alavancagem de mercado - geralmente usam o valor contábil da dívida porque o valor de mercado é difícil de observar
 - ✓ Valor da firma - mais uma vez, é difícil observar o valor de mercado da dívida, então usamos o valor contábil
 - ✓ Compensação de CEO - o valor das opções é aproximado usando Black-Scholes.
 - ✓ Variação do PIB – crescimento da economia
- É plausível assumirmos que e e x sejam não correlacionados?

Erro de Medida na variável dep

- Resposta = Talvez sim... talvez não
 - ✓ Ex. - A alavancagem da empresa é medida com erro; difícil observar o valor de mercado da dívida, então usamos o valor contábil
 - ✓ Mas, o erro de medição provavelmente será maior quando a empresa estiver em perigo... O valor de mercado da dívida cai; valor contábil não
 - ✓ Esse erro pode ser correlacionado com x se incluir itens como lucratividade (ou seja, maior para empresas de baixo lucro)
- Este tipo de Erro de Medida causará estimativas inconsistentes

Erro de Medida na variável Indep

- Vamos supor que o modelo seja $y = \beta_0 + \beta_1 x^* + u$
- Mas, observamos x^* com erro, $e = x - x^*$
 - ✓ Assumimos que $E(y|x^*, x) = E(y|x^*)$ [ou seja, x não afeta y depois de controlar por x^* ; isso é padrão e incontroverso porque é uma consequência de se afirmar que escrevemos o modelo correto]
- Quais são alguns exemplos?

Erro de Medida na variável Indep

- Existem muitos exemplos!
 - ✓ Q de tobin mede oportunidades de investimento com erro
 - ✓ Índices de inflação, empregabilidade, ...
 - ✓ O índice de nível de governança da Bovespa, etc., são apenas medidas ruidosas do nível relativamente abstrato de “governança” da firma.
- Esses erros de mensuração causarão viés?

Erro de Medida na variável Indep

- **Resposta!** depende crucialmente do que assumimos sobre o erro de medição, e
- A literatura enfoca dois pressupostos extremos
 - ✓ # 1 - erro de medição, e , não está correlacionado com a medida observada, x
 - ✓ #2 - Erro de medição, e , não está correlacionado com a medida não observada, x^*

Erro de Medida na variável Indep

- **Pressuposto 1:** e , não está correlacionado com x
- Substituindo x^* pelo que realmente observamos, $x^* = x - e$, dentro do verdadeiro modelo, temos $y = \beta_0 + \beta_1 x + u - \beta_1 e$
- Existe algum viés?
 - ✓ Resposta = Não. x não está correlacionada com e por suposição, e x não está correlacionada com u por suposições anteriores
- O que acontece com nossos erros padrão?
 - ✓ Resposta = Eles ficam maiores; A variância de erro é agora
$$\sigma_u^2 + \beta_1^2 \sigma_e^2$$

Erro de Medida na variável Indep

- Nós ainda estamos estimando $y = \beta_0 + \beta_1 x + u - \beta_1 e$, mas agora x está correlacionado com e
 - ✓ e correlacionado com x^* garante que é correlacionado com x ;
 $\text{cov}(x, e) = E(xe) = E(x^*e) + E(e^2) = \sigma_e^2$
 - ✓ Ou seja, se uma variável independente for correlacionada com o erro ... vamos obter estimativas viesadas!
- É o que as pessoas chamam de suposição clássica de Erro nas Variáveis (CEV)

Erro de Medida na variável Indep

- Deixando a matemática de lado, é possível mostrar que as estimativas são:

$$plim(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right) \leftarrow \text{Está sempre entre 0 e 1}$$

- A estimativa é sempre na direção do zero; ou seja, é um viés de atenuação
 - ✓ E, se a variação de erro, σ_e^2 , for pequena, o viés de atenuação não será tão problemático

Erro de Medida não tão ruim?

- Na configuração atual, o erro de medição não parece tão ruim...
 - ✓ Se erro não correlacionado com o observado x , não viés
 - ✓ Se o erro não estiver correlacionado com o não observado x^* , obtemos um viés de atenuação ... então pelo menos o sinal em nosso coeficiente de interesse ainda estará correto
- Por que isso é enganoso?

Erro de Medida muito RUIM!!

- A verdade é que o erro de medição é provavelmente correlacionado um pouco com o observado x e não observado x^*
 - ✓ Se for verdade, então é muito difícil descobrir qual é o sinal e a magnitude do viés
 - ✓ Além disso, mesmo no caso do CEV, se houver mais de uma variável independente, o viés fica terrivelmente complicado...
 - ✓ Se mais de um x conter erro de medida
 - TODOS os β serão enviesados se a variável com erro de medida estiver correlacionada com qualquer outra x [que presumivelmente é verdadeira se foi incluída!]
 - O sinal e a magnitude dos vieses dependerão de todas as correlações entre xxs ; ou seja, uma grande bagunça

Viés de Simultaneidade

- Isso ocorrerá sempre que qualquer uma das variáveis supostamente independentes (ou seja, os x 's) puder ser afetada por alterações na variável y ; Por exemplo.

$$y = \beta_0 + \beta_1 x + u$$

$$x = \delta_0 + \delta_1 y + v$$

- ✓ mudanças em x afetam y e mudanças em y afetam x ; este é o caso mais simples de causalidade reversa
- ✓ Uma estimativa de $y = \beta_0 + \beta_1 x + u$ será tendenciosa...

Viés de Simultaneidade

Exemplo anterior é o caso de causalidade reversa; a variável de interesse também é afetada por y

Mas, se y afetar qualquer x , haverá um viés; Por exemplo.


$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$x_2 = \alpha_0 + \alpha_1 y + w$$

É possível mostrar que x_2 está correlacionado com u ; e haverá um viés em todos os coeficientes

É por isso que as pessoas usam os x 's defasados em muitos estudos de finanças e economia.

Viés de Simultaneidade

- Na minha opinião, o exemplo anterior é o que significa ter um problema de “endogeneidade” ou uma variável “endógena”
 - Mas, como eu mencionei anteriormente, há muitos problemas com o uso da palavra “endogeneidade” em economia e finanças. Então, seria melhor dizer “viés de simultaneidade”.
- 

Viés de Simultaneidade – Bad Controls

- Outro exemplo de viés de simultaneidade é quando um x é afetado por outro x ; por exemplo.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$x_2 = \gamma_0 + \gamma_1 x_1 + v$$

- Angrist-Pischke chama isso de controle ruim (Bad Control)
- Você não conseguirá obter uma estimativa consistente de β_1 ou β_2 ao estimar a equação.

Viés de Simultaneidade – Bad Controls

- Veja porque as estimativas que você obtém não são de β_1 ou β_2 , considere que o problema ocorre apenas com x_2

$$y = \beta_0 + \beta_1 x_1 + \beta_2 (\gamma_0 + \gamma_1 x_1 + v) + u$$

$$y = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) x_1 + (u + \beta_2 v)$$



Coeficiente estimado para x_1 vai conterá mais do que apenas o efeito de x_1 em y

- Isso é problemático especialmente se o x_2 , nesse caso, não tem uma relação econômica com y .

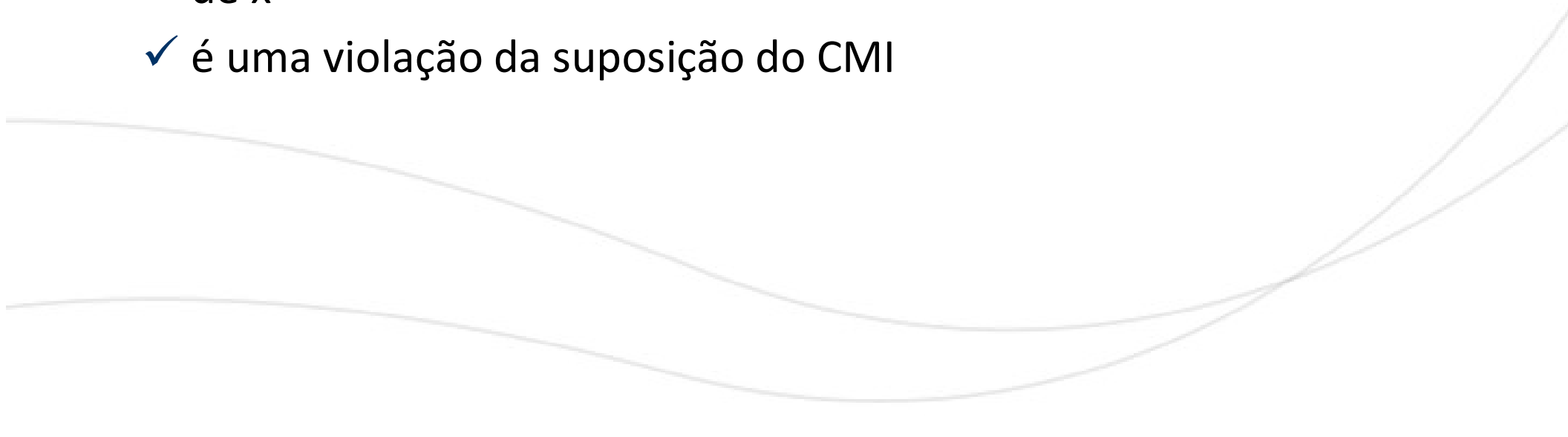
Viés de Simultaneidade – Bad Controls

- Se você tem um x que é verdadeiramente exógeno (ou seja, aleatório) [como você pode ter em experimentos naturais], não coloque controles que também são afetados por x !
 - ✓ Somente adicione controles não afetados por x , ou apenas regrida seu y nos vários x 's e em x sozinho!
- Adicionar controles é um ponto muitas vezes controverso...

Viés de Simultaneidade

- Resumindo
- Se o seu x também puder ser afetado pelo y (ou seja, causalidade reversa), você não poderá fazer inferências causais usando o OLS
 - ✓ Variáveis instrumentais ou experimentos naturais serão úteis com este problema
- Não é possível obter estimativas causais com o OLS se os controles forem afetados pelo y ou outro x 's

Viés de Seleção

- O que é Viés de Seleção?
 - É mais fácil pensar nisso apenas como um problema de variável omitida, em que a variável omitida é o contrafactual não observado dos elementos da sua amostra
 - ✓ Especificamente o erro, u , contém algum contrafactual não observado que está correlacionado com certos valores observados de x
 - ✓ é uma violação da suposição do CMI
- 

Viés de Seleção - Exemplo

- Média de um indicador de saúde dos visitantes do hospital = 3,21
- Média de um indicador de saúde dos não visitantes = 3,93
- Podemos concluir que ir ao hospital (ou seja, o x) deixa você menos saudável?
 - ✓ Resposta = Não. As pessoas que vão ao hospital são inerentemente menos saudáveis [isto é o viés de seleção]
 - ✓ Outra maneira de dizer isso: falhamos em controlar quais condições de saúde estariam ausentes da visita, e esse contrafactual não observado está correlacionado com a hospitalização ou não [ou seja, variável omitida]


Viés de Sobrevivência

- O que é Viés de Sobrevivência?
- É mais fácil pensar nisso apenas como um problema particular de viés de seleção, em que a variável omitida são elementos contrafactuais não observados que deveriam fazer parte da sua amostra

Viés de Sobrevivência - Exemplo

- Ex. # 1 - As empresas que fazem IPO e são adicionadas aos conjuntos de dados que cobrem empresas abertas podem ser diferentes das empresas que não fazem IPO
 - ✓ Ex. # 1 - por exemplo a abertura de capital pode não causar um alto crescimento; é só que as empresas que vão a público vão crescer mais rápido de qualquer maneira
- Ex. # 2 - As empresas afetadas negativamente por algum evento podem, subsequente, serem retiradas da base de dados devido a problemas ou falências
 - ✓ Ex. # 2 - Pode não encontrar efeito adverso do evento (ou pode subestimar o efeito) se algumas empresas afetadas forem à falência e forem retiradas

Viés de Sobrevivência

- Mais uma vez, não há soluções fáceis; mas, se você preocupado com o viés de sobrevivência ...
 - Base de dados limitadas
 - Amostra limitada
- 

Viés de Sobrevivência - Exemplo

- Tenha cuidado com as inferências para evitar afirmações que carecem de validade externa
- Argumente que sua amostra é representativa se abrange um grupo economicamente importante
- Colete manualmente seus próprios dados se a teoria que você está interessado em testar requer isto!
 - ✓ Isso pode realmente resultar em um ótimo trabalho e está se tornando cada vez mais importante