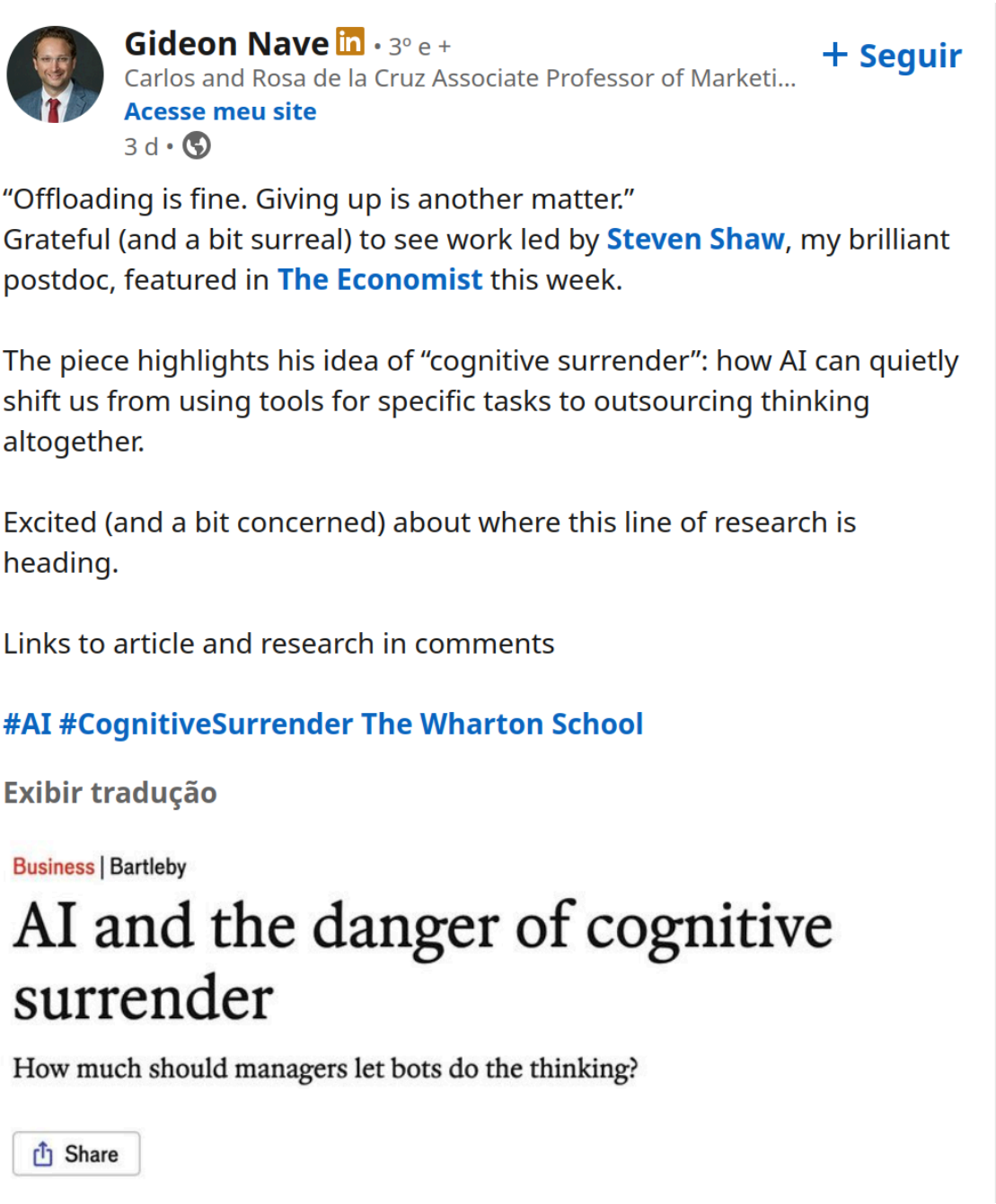
Variável Instrumental e MQ2E

IBM0288 - 2026.2

Prof. Raphael Gouvea

1 Para reflexão



Gideon Nave • 3º e + — + Seguir
Carlos and Rosa de la Cruz Associate Professor of Marketi...
Acesse meu site
3 d •

"Offloading is fine. Giving up is another matter."
Grateful (and a bit surreal) to see work led by **Steven Shaw**, my brilliant postdoc, featured in **The Economist** this week.

The piece highlights his idea of "cognitive surrender": how AI can quietly shift us from using tools for specific tasks to outsourcing thinking altogether.

Excited (and a bit concerned) about where this line of research is heading.

Links to article and research in comments

#AI #CognitiveSurrender The Wharton School

Exibir tradução

Business | Bartleby

AI and the danger of cognitive surrender

How much should managers let bots do the thinking?

Share

2 Aulas passadas: vieses

Viés de seleção: ocorre quando grupos de tratamento e controle não são comparáveis devido a atribuição não aleatória dos participantes em cada grupo, o que causa diferenças sistemáticas entre eles.

- Solução: Se tivermos dados experimentais, aleatorização resolve o problema do viés. Com dados observacionais, incluir controles para fazer comparações *ceteris paribus*.

Viés de variável omitida (OVB): quando uma variável não incluída no modelo gera correlação entre o resíduo e algum regressor do modelo.

- Solução: incluir variável omitida ou trabalhar com estrutura da painel para controlar por efeitos fixos de entidade e tempo.
- Observação: painel não permite eliminar todo tipo OVB. Quais não elimina?

3 Outras fontes de viés

Viés de causalidade simultânea: ocorre quando além da causalidade ir dos regressores para a variável dependente (X causa Y), há também causalidade reversa da variável dependente para um ou mais regressores (Y causa X).

- Exemplo notas e tamanho de turmas: Se políticas públicas priorizam contratar mais professores em escolas com notas baixas (subsídio), surge causalidade simultânea: o tamanho da turma afeta as notas e as notas afetam afetam o tamanho das turmas!

Viés de erros nas variáveis: ocorre quando a variável independente X é reportada imprecisamente.

- Exemplo: lançamentos errados em bases administrativas; erros de respostas em pesquisas amostrais; questões ambíguas; respostas imprecisas intencionais.

4 Variável Instrumental (VI): por quê?

Vieses

Ao se trabalhar com dados observacionais em regressões MQO, a presença de viés de variável omitida, viés de causalidade simultânea ou viés de erros de medida será um problema de validade interna que inviabiliza a obtenção de estimativas causais. Todos três problemas resultam em $E(u \mid X) \neq 0$!

A regressão com variáveis instrumentais pode eliminar o viés quando $E(u \mid X) \neq 0$ (independentemente do tipo)!

5 Um Regressor e um Instrumento

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- O objetivo é estimar o efeito causal β_1. Porém, X é correlacionada com o termo de erro, e não podemos resolver o problema simplesmente incluindo variáveis de controle.
- A regressão com variáveis instrumentais (VI) divide X em duas partes: uma que pode ser correlacionada com u, e outra que não é. Ao isolar a parte não correlacionada com u, é possível estimar β_1.
- Isso é feito usando uma variável instrumental, Z_i, que é correlacionada com X_i mas não correlacionada com u_i.

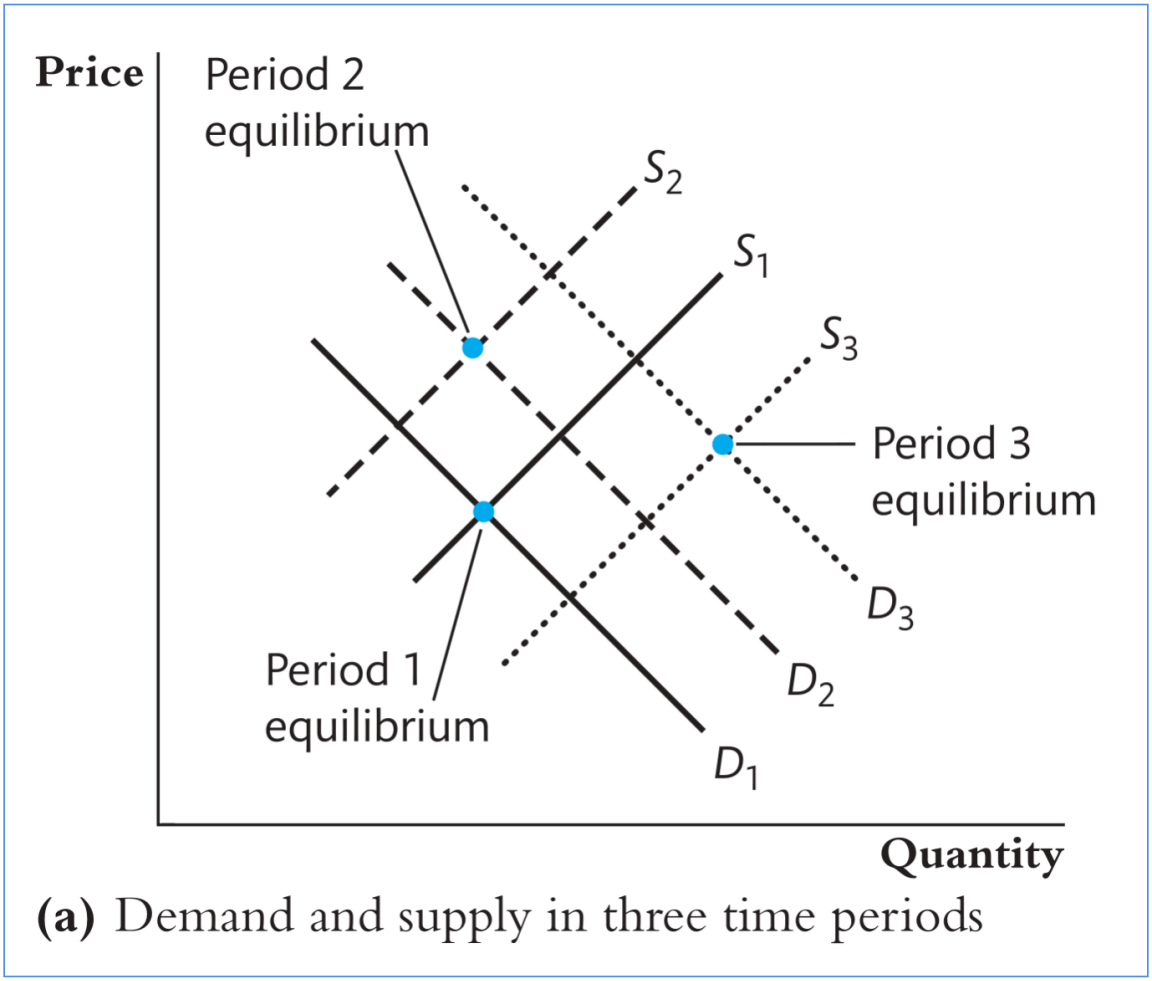
6 O Problema de Philip Wright

- Philip Wright se preocupava com um tema central de sua época: como definir uma **tarifa de importação** . Para entender o efeito econômico de uma tarifa, era crucial obter **estimativas** das curvas de demanda e de oferta dos bens.
- Considere o problema de estimar a **elasticidade da demanda por manteiga** a partir da equação de demanda:

$$\ln(Q_i^{\text{butter}}) = \beta_0 + \beta_1 \ln(P_i^{\text{butter}}) + u_i$$

  onde Q_i^{butter} é a i-ésima observação do **consumo de manteiga**, P_i^{butter} é o seu **preço**, e u_i agrega outros fatores que afetam a demanda (renda, gostos, etc.).
- O coeficiente β_1 mensura a **elasticidade** de Q em relação a P.
- Devido às **interações entre oferta e demanda**, o regressor $\ln(P_i^{\text{butter}})$ tende a ser **correlacionado** com o erro u_i (endogeneidade).

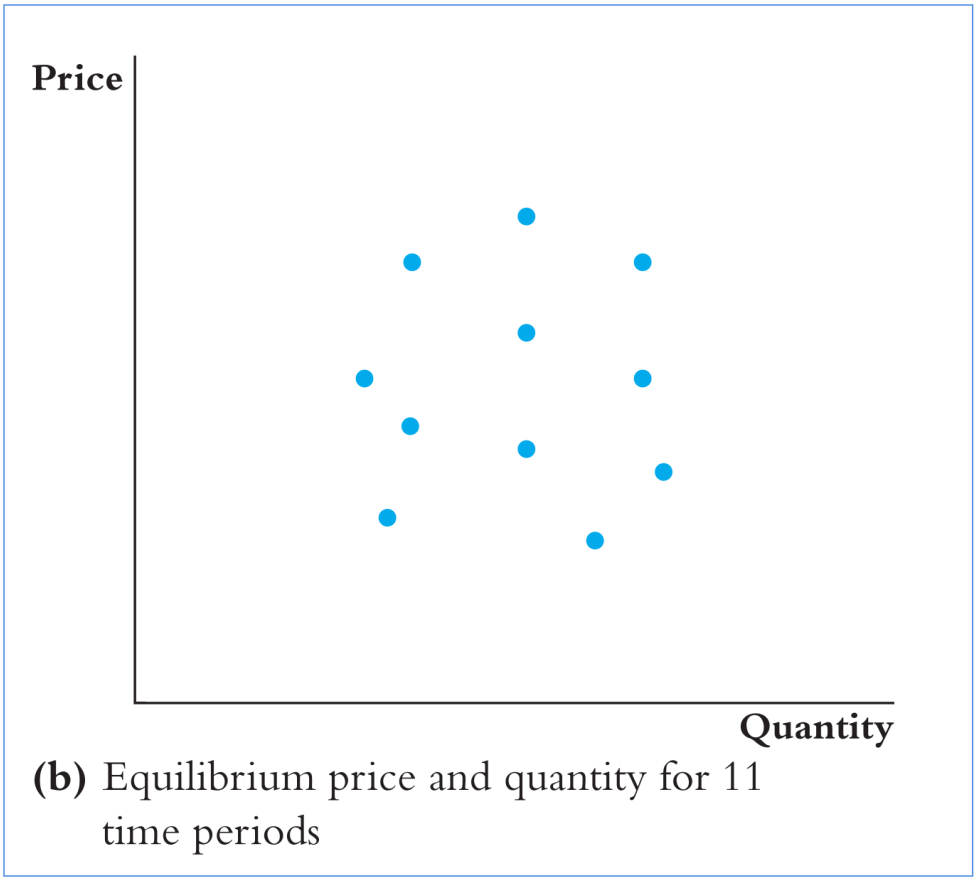
7 Identificação: oferta e demanda



(a) Demand and supply in three time periods

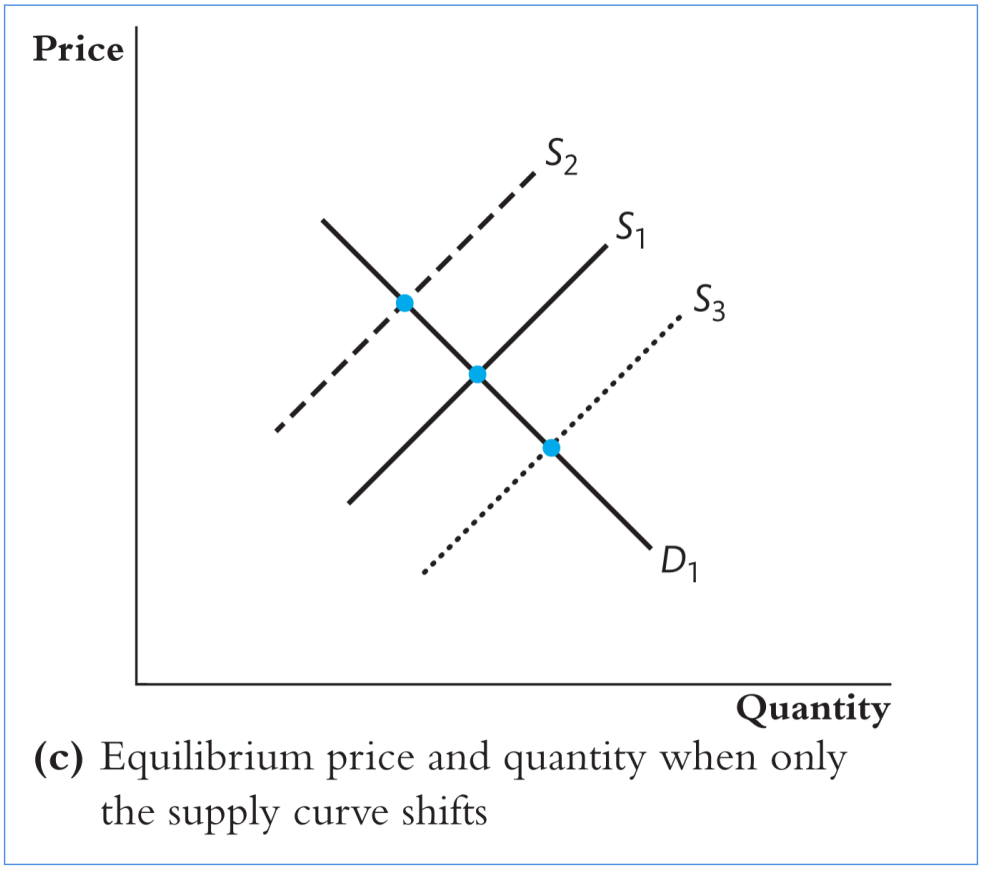
- No ano 1, as curvas são D_1 (demanda) e S_1 (oferta). O preço e a quantidade de equilíbrio do período são dados por sua **interseção**.
- No ano 2, a demanda sobe $D_1 \rightarrow D_2$ e a oferta cai $S_1 \rightarrow S_2$.O novo equilíbrio é determinado pela nova interseção.
- No ano 3, a demanda sobe novamente D_3 e a oferta sobe S_3. Um novo preço e quantidade de equilíbrio são determinados.
- Como os pontos são gerados por **mudanças simultâneas** em demanda **e** oferta, **não** é possível identificar isoladamente nenhuma das duas curvas a partir deles.

8 Identificação: oferta e demanda



(b) Equilibrium price and quantity for 11 time periods

Considere o gráfico de dispersão de (P, Q) de equilíbrio: ajustar uma reta **não** estima a curva de demanda **nem** a de oferta!



Para identificar a **demanda**, precisaríamos **fixar a oferta**; e vice-versa, para identificar a **oferta**, precisaríamos fixar a demanda.

9 O Problema de Philip Wright: (VI)

- Wright mostrou que precisamos de uma **terceira variável** — a **variável instrumental (VI)** — que **desloque a oferta** sem deslocar a demanda.
- A VI deve ser **correlacionada com o preço** (**relevância**).
  - A VI desloca a curva de oferta ⇒ altera o preço de equilíbrio.
- A VI deve ser **não correlacionada com o erro** da demanda (**exogeneidade**).
  - A VI **não desloca a demanda** (não afeta diretamente a utilidade/gosto).
- Wright considerou o **clima** (chuvas) como instrumento:
  - **Chuvas abaixo da média** em região leiteira prejudicam a pastagem e **reduzem a produção** de manteiga a um dado preço ⇒ deslocam a **oferta para a esquerda** e **elevam o preço** de equilíbrio.
- **Relevância do instrumento:** a chuva na região leiteira afeta **diretamente a oferta** de manteiga e, portanto, o **preço**.
- **Exogeneidade do instrumento:** a chuva na região leiteira **não afeta diretamente a demanda** por manteiga.

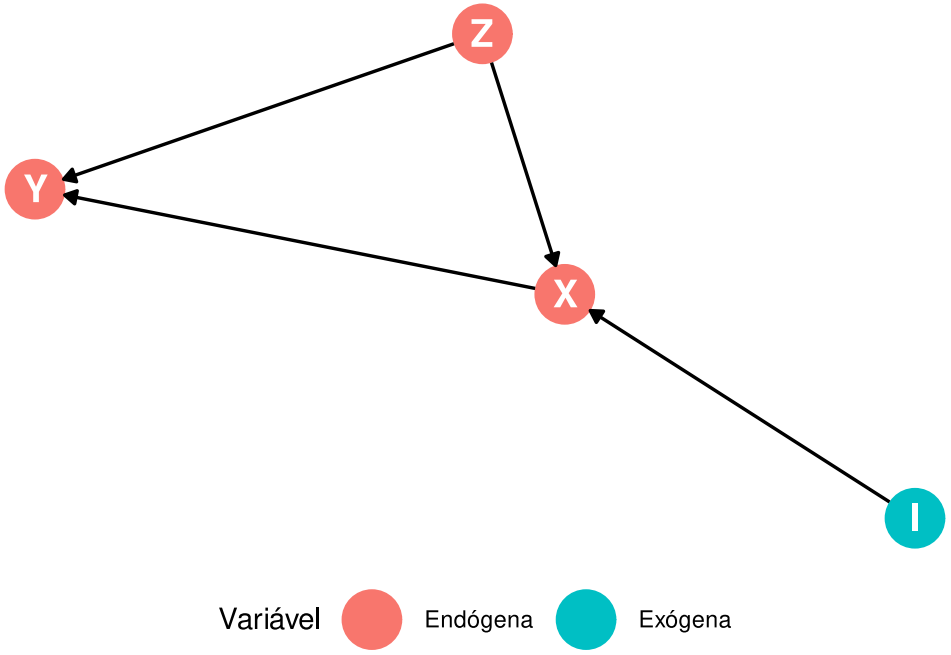
10 Terminologia

- **Variável endógena**: variável que é correlacionada com u
- **Variável exógena**: variável que **não** é correlacionada com u
- **Variável Instrumental**: X é endógena e existe um instrumento Z que é exógeno.

Ibmec

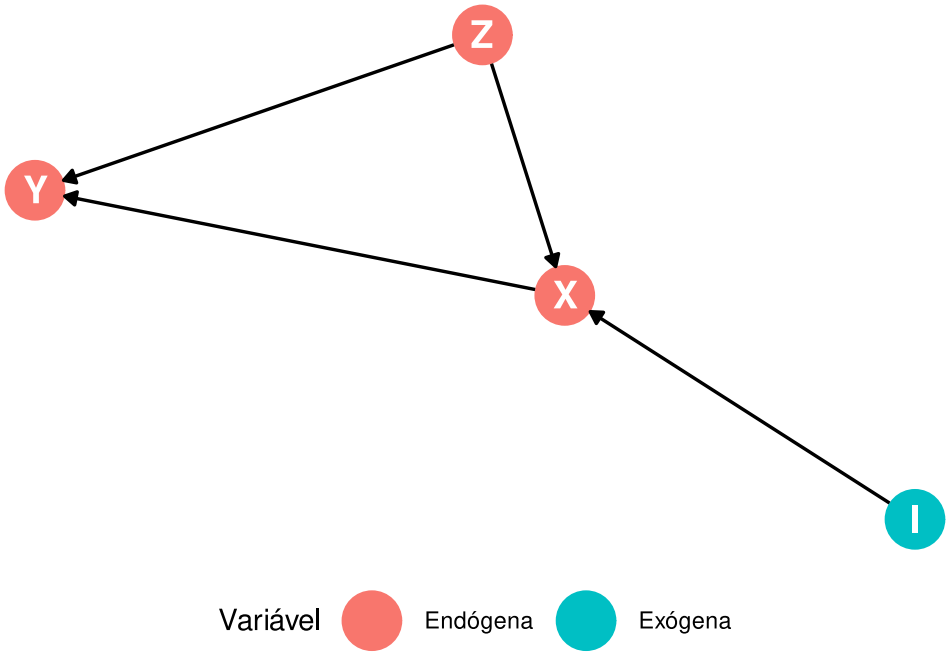
> **Nota terminológica**
>
> "Endógena" significa literalmente "determinada dentro do sistema." Se X é determinada conjuntamente com Y, então a regressão de Y em X está sujeita ao viés de causalidade simultânea. Mas essa definição é muito restrita, pois a regressão VI também pode tratar o viés de variável omitida e de erros nas variáveis. Por isso, utilizamos a definição mais ampla acima.

11 Entendendo as VIs



- **X** e **Y** são correlacionadas.
- Existe uma variável de **Z** que satisfaz as condições do **viés de variável omitida**:
  1. Causa Y (está presente no termo de erro u);
  2. É correlacionada com X.
- Relação causal entre X e Y:
  1. $X \rightarrow Y$ (causal, via "porta da frente");
  2. $X \leftarrow Z \rightarrow Y$ (não causal, via "porta dos fundos").

12 Entendendo as VIs



- Considere a variável **I**, que causa **X**, mas **não** causa diretamente **Y**.
- A variável **I** não possui caminhos de "porta dos fundos" entre ela e Y.
  A única forma de chegar a Y a partir de I é **por meio de X**: $I \rightarrow X \rightarrow Y$
- A variável **I** é um **bom instrumento** para X se satisfizer duas condições:
  1. **Condição de relevância:** I explica estatisticamente e de forma significativa X;
  2. **Condição de exclusão:** I é não correlacionada com u, portanto **não afeta diretamente Y**.

13 Regressão com VI

- **Regressão VI:** método para obter um estimador **consistente** dos coeficientes causais desconhecidos quando o regressor X é **correlacionado** com o erro u.
- **Intuição da IV:** separar a variação de X em duas partes:
  1. uma parte **correlacionada** com u;
  2. outra parte **não correlacionada** com u. Se conseguirmos isolar a parte não correlacionada, podemos eliminar a componente de X que gera viés no estimador MQO. Essa informação é obtida por meio de **variáveis instrumentais**.
- Seja β_1 o **efeito causal** de X sobre Y: $Y_i = \beta_0 + \beta_1 X_i + u_i$ onde u_i representa fatores omitidos que também determinam Y_i.
- Se X_i e u_i forem correlacionados, o estimador MQO será **inconsistente**.
- A estimação por variáveis instrumentais usa uma variável Z para **isolar a parte de X** que é **não correlacionada** com u.

14 Validade do Instrumento

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Para que uma variável instrumental Z seja válida, ela deve satisfazer duas condições:

1. **Relevância do instrumento:** $\text{corr}(Z_i, X_i) \neq 0$
2. **Exogeneidade do instrumento:** $\text{corr}(Z_i, u_i) = 0$

Relevância: a variação do instrumento está relacionada à variação do regressor explicativo.

Exogeneidade: a parte da variação explicada pelo instrumento que é **exógena**, ou seja, não correlacionada com fatores não observados.

15 MQ2E: primeiro estágio

Explicação #1: Mínimos Quadrados em Dois Estágios (MQ2E)

Como o nome sugere, o MQ2E tem dois estágios, ou seja, duas regressões:

1. Isole a parte de X não correlacionada com u, regredindo X em Z por MQO:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- Como Z_i não é correlacionada com u_i, o termo $\pi_0 + \pi_1 Z_i$ também não é correlacionado com u_i. Não conhecemos π_0 ou π_1, mas podemos estimá-los por MQO.

Obtenha os valores ajustados do primeiro estágio:

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

16 MQ2E: segundo estágio

2. No segundo estágio, regrida Y_i em $\hat{X}_i$ por MQO:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- Como $\hat{X}_i$ é não correlacionada com u_i (por ser função apenas de Z_i), o estimador de MQO de β_1 na regressão (2) é consistente.

Portanto, em amostras grandes, β_1 pode ser estimado por MQO usando a regressão (2).

17 MQ2E: Resumo

Suponha que Z_i satisfaça as duas condições para um instrumento válido:

- **Relevância:** $\text{corr}(Z_i, X_i) \neq 0$
- **Exogeneidade:** $\text{corr}(Z_i, u_i) = 0$

Mínimos Quadrados em Dois Estágios:

Estágio 1: Regrida X_i em Z_i por MQO $\longrightarrow$ obtenha $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$

Estágio 2: Regrida Y_i em $\hat{X}_i$ por MQO $\longrightarrow$ obtenha $\hat{\beta}_1^{MQ2E}$

Os coeficientes do segundo estágio são os estimadores de MQ2E, mas os **erros-padrão estão errados** (usar comando único no R).

18 MQ2E: derivação algébrica

Explicação #2: Derivação algébrica direta

$Y_i = \beta_0 + \beta_1 X_i + u_i$

Assim,

$$\begin{aligned}
\text{cov}(Y_i, Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\
&= \text{cov}(\beta_0, Z_i) + \text{cov}(\beta_1 X_i, Z_i) + \text{cov}(u_i, Z_i) \\
&= 0 + \beta_1 \ \text{cov}(X_i, Z_i) + 0 \\
&= \beta_1 \ \text{cov}(X_i, Z_i)
\end{aligned}$$

onde $\text{cov}(u_i, Z_i) = 0$ pela exogeneidade do instrumento! Logo:

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

Ibmec

19 MQ2E: covariâncias amostrais

O estimador VI substitui essas covariâncias populacionais por covariâncias amostrais:

$$\hat{\beta}_1^{MQ2E} = \frac{s_{YZ}}{s_{XZ}}$$

onde $s_{YZ} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})$ e $s_{XZ} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})$ são as covariâncias amostrais.

Este é o estimador de MQ2E apenas com uma derivação diferente!

> Derivação por forma reduzida
>
> Existe também uma terceira forma de derivar que consiste em resolver a forma reduzida que relaciona Y com Z e X com Z: $X_i = \pi_0 + \pi_1 Z_i + v_i$ e $Y_i = \gamma_0 + \gamma_1 Z_i + w_i$, onde w_i é um termo de erro. Como Z é exógena, Z é não correlacionada com v_i e w_i. Não será demonstrado por questão de tempo.

20 Distribuição Amostral e Inferência

- Para um único regressor X e um único instrumento Z, o estimador MQ2E possui uma forma simples.
- O estimador MQ2E de β_1 é o **quociente** entre a covariância amostral entre Z e Y e a covariância amostral entre Z e X: $\hat{\beta}_1^{\text{MQ2E}} = \frac{s_{ZY}}{s_{ZX}}$
- Em amostras grandes, $\hat{\beta}_1^{\text{MQ2E}}$ é **consistente** e **assintoticamente normal**: $\hat{\beta}_1^{\text{MQ2E}} \xrightarrow{p} \beta_1$ $\hat{\beta}_1^{\text{MQ2E}} \sim \mathcal{N}(\beta_1, \sigma^2_{\hat{\beta}_1^{\text{MQ2E}}})$ $\sigma^2_{\hat{\beta}_1^{\text{MQ2E}}} = \frac{1}{n}\frac{\text{Var}[(Z_i - \mu_Z)u_i]}{[\text{Cov}(Z_i, X_i)]^2}$
- Como $\hat{\beta}_1^{\text{MQ2E}}$ é **normalmente distribuído** em grandes amostras, testes de hipótese sobre β_1 podem ser realizados utilizando a **estatística *t***.

21 A Demanda por Cigarros

- Usamos o **método MQ2E** (*Two-Stage Least Squares*) para estimar a **elasticidade-preço da demanda por cigarros**, utilizando dados anuais dos **48 estados contíguos dos EUA**, de **1985 a 1995**.
- **Variável dependente (Regressanda):** $Q^{\text{cigarettes}}$ Número de maços de cigarros vendidos **per capita** em cada estado.
- **Regressor:** $P^{\text{cigarettes}}$ Preço médio **real** por maço de cigarros (incluindo todos os impostos).
- **Instrumento:** SalesTax Parcela do imposto sobre cigarros correspondente ao **imposto sobre vendas** geral, em dólares reais por maço.
- **Relevância do instrumento:** um imposto sobre vendas mais alto **aumenta** o preço final pago pelo consumidor.
- **Exogeneidade do instrumento:** o imposto sobre vendas afeta a demanda por cigarros **apenas indiretamente**, via preço, pois varia entre estados **por razões políticas**, não por fatores ligados à demanda.

22 A Demanda por Cigarros – MQ2E

Primeira etapa: estimação do preço em função do imposto sobre vendas

$$\widehat{\ln(P^{\text{cigarettes}})} = 4,62\,(0,03) + 0,031\,(0,005)\,\text{SalesTax}$$

- $\bar{R}^2 = 0,47$
- **Impostos mais altos** estão associados a **preços mais altos**.
- A variação do imposto explica **47%** da variância dos preços entre estados.

. . .

Segunda etapa: estimação da demanda de cigarros utilizando o preço previsto

$$\widehat{\ln(Q^{\text{cigarettes}})} = 9,72\,(1,53) - 1,08\,(0,32)\,\widehat{\ln(P^{\text{cigarettes}})}$$

- Um **aumento de 1% no preço** reduz o consumo em **1,08%**.
- Isso sugere que a **demanda é elástica**, embora possam existir **variáveis omitidas** (como **renda**) que também influenciam o consumo.

23 Notas e Tamanho da Turma

- As regressões de nota/tamanho de turma da Califórnia ainda podem ter viés de variável omitida (ex.: envolvimento dos pais).

- Em princípio, esse viés pode ser eliminado por regressão IV (MQ2E).
- A regressão IV requer um instrumento válido, isto é, um instrumento que seja:
  - **relevante:** $\text{corr}(Z_i, STR_i) \neq 0$
  - **exógeno:** $\text{corr}(Z_i, u_i) = 0$

24 Notas e Tamanho da Turma: VI

Instrumento hipotético:alguns distritos, atingidos aleatoriamente por um terremoto, "dobram" as salas de aula

- $Z_i = Terremoto_i = 1$ se atingido pelo terremoto, $= 0$ caso contrário
- As duas condições para um instrumento válido são atendidas?

. . .

- O terremoto torna a situação como se os distritos estivessem num experimento de atribuição aleatória. Assim, a variação em STR decorrente do terremoto é exógena.
- O primeiro estágio do MQ2E regride STR em Terremoto, isolando a parte de tamanhos das turmas que é exógena (a parte "como se" fosse aleatoriamente atribuída).

25 O Modelo Geral de VI

- Até agora consideramos a regressão VI com um único regressor endógeno (X) e um único instrumento (Z).
- Precisamos estender isso para:
  - múltiplos regressores endógenos $(X_1, ..., X_k)$
  - múltiplas variáveis exógenas incluídas $(W_1, ..., W_r)$ ou variáveis de controle
  - múltiplos instrumentos $(Z_1, ..., Z_m)$: com mais instrumentos (relevantes) pode-se produzir uma variância menor no MQO2E. Por quê?
- **Nova terminologia:** identificação e sobreidentificação

26 Identificação

- Em geral, diz-se que um parâmetro é **identificado** se diferentes valores do parâmetro produzem diferentes distribuições dos dados.
- Na regressão VI, se os coeficientes são identificados depende da relação entre o número de instrumentos (m) e o número de regressores endógenos (k).
- Intuitivamente, se há menos instrumentos do que regressores endógenos, não podemos estimar $\beta_1, ..., \beta_k$.
  - Por exemplo, suponha $k = 1$ mas $m = 0$ (sem instrumentos)!

27 Condições para identificação

Os coeficientes $\beta_1, ..., \beta_k$ são:

- **Exatamente identificados** se $m = k$.
  - Há instrumentos suficientes para estimar $\beta_1, ..., \beta_k$.
- **Sobreidentificados** se $m > k$.

- Há mais instrumentos do que o necessário. Nesse caso, é possível *"testar"* a validade dos instrumentos (testes de sobreidentificação).

- **Subidentificados** se $m < k$.
  - Há instrumentos insuficientes para estimar $\beta_1, ..., \beta_k$. É preciso obter mais instrumentos!

28 Resumo da Terminologia

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

- Y_i é a variável dependente
- $X_{1i}, ..., X_{ki}$ são os **regressores endógenos** (potencialmente correlacionados com u_i)
- $W_{1i}, ..., W_{ri}$ são as **variáveis exógenas incluídas** (não correlacionadas com u_i) ou variáveis de controle
- $\beta_0, \beta_1, ..., \beta_{k+r}$ são os coeficientes de regressão desconhecidos
- $Z_{1i}, ..., Z_{mi}$ são os m **instrumentos** (variáveis exógenas excluídas)
- Os coeficientes são sobreidentificados se $m > k$; exatamente identificados se $m = k$; subidentificados se $m < k$.

29 MQ2E com um regressor endógeno

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

Com m instrumentos: $Z_{1i}, ..., Z_{mi}$

Primeiro estágio: Regrida X_1 em todos os regressores exógenos: regrida X_1 em $W_1, ..., W_r, Z_1, ..., Z_m$, e um intercepto, por MQO $\longrightarrow$ obtenha $\hat{X}_{1i}$

Segundo estágio: Regrida Y_i em $\hat{X}_{1i}, W_{1i}, ..., W_{ri}$, e um intercepto, por MQO.

Os coeficientes desse segundo estágio são os estimadores de MQ2E, mas **os desvios-padrão precisam ser corrigidos**. Para obter os desvios corretos, faça isso em um único passo R.

30 Exemplo: demanda por cigarros

Suponha que queremos estimar a elasticidade da demanda de cigarros e quea renda seja exógena (plausível — por quê?):

$$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln Inc_i + u_i$$

Temos dois instrumentos:

- $Z_{1i} =$ imposto geral de vendas$_i$
- $Z_{2i} =$ imposto específico sobre cigarros$_i$

Variável exógena incluída: $\ln(Inc_i)$ (W)

β_1 é sobre-, sub- ou exatamente identificado?

31 Hipóteses para inferência causal

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

1. $E(u_i \mid W_{1i}, ..., W_{ri}) = 0$ — "os regressores exógenos são exógenos"
2. $(Y_i, X_{1i}, ..., X_{ki}, W_{1i}, ..., W_{ri}, Z_{1i}, ..., Z_{mi})$ são i.i.d.
3. Os X's, W's, Z's e Y têm momentos de 4ª ordem não nulos e finitos.
4. Os instrumentos $(Z_{1i}, ..., Z_{mi})$ são válidos.

Sob as hipóteses 1–4, o MQ2E e sua estatística t são normalmente distribuídos em amostras grandes.

O requisito crítico é que os instrumentos sejam válidos.

32 Qual o papel das variáveis W?

- Em muitos casos, o propósito de incluir as $W's$ é controlar por fatores omitidos para que Z não seja correlacionada com u.
- Nesse caso, as $W's$ não precisam ser exógenas: elas precisam ser variáveis de controle eficazes!
  - Foco agora é em produzir um instrumento exógeno e não aferir causalidade dos Ws
- Tecnicamente, a condição para as $W's$ serem variáveis de controle eficazes é que a média condicional de u_i não dependa de Z_i, dado W_i:

$$E(u_i \mid W_i, Z_i) = E(u_i \mid W_i)$$

> **i** Papel das variáveis de controle em VI
>
> A ideia central para inclusão de controles em regressões com variáveis instrumentais é que, em muitas aplicações, controles são essenciais para garantir que os instrumentos sejam plausivelmente exógenos, ou sej, não correlacionados com o resíduo da regressão.

33 Exemplo: efeito de estudar sobre notas

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Y = GPA do primeiro semestre
- X = horas médias de estudo por dia
- Z = 1 se o colega de quarto trouxe videogame, = 0 caso contrário
- Colegas de quarto foram designados aleatoriamente

Vocês conseguem pensar em um motivo pelo qual Z pode ser correlacionada com u mesmo sendo designada aleatoriamente?

O que mais entra no termo de erro? quais são os outros determinantes das notas, além do tempo de estudo?

34 Exemplo: efeito de estudar sobre notas

Por que Z poderia ser correlacionada com u? Uma possibilidade: **o sexo do estudante**:

- Colegas de quarto são designados aleatoriamente: exceto sempre homens com homens e mulheres com mulheres.
- Mulheres tiram notas melhores do que homens, mantendo constantes as horas de estudo.
- Homens têm mais probabilidade de trazer um videogame do que mulheres.

Então $\text{corr}(Z_i, u_i) < 0$:

- homens têm mais probabilidade de ter um colega de quarto [homem] que traz videogame
- homens também tendem a ter notas mais baixas, mantendo constante a quantidade de estudo.

Como $\text{corr}(Z_i, u_i) < 0$, o instrumento não seria válido. A solução é controlar pelo sexo para que $\text{corr}(Z_i, u_i) = 0$!

35 Exemplo: efeito de estudar sobre notas

Essa lógica leva a incluir W = sexo do estudante como variável de controle na regressão VI:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

A condição de independência condicional da média para um instrumento exógeno é: $E(u_i \mid Z_i, W_i) = E(u_i \mid W_i)$.

Em palavras: entre os homens (condicional em W = masculino), os colegas de quarto são designados aleatoriamente, portanto se o seu colega traz um videogame é aleatório. O mesmo vale para as mulheres.

O instrumento não é exógeno se W não está incluída na regressão! Mas quando W é incluída, a condição de independência condicional é atendida e o instrumento é válido.

Ibmec

36 Verificando a validade dos instrumentos

Lembre-se dos dois requisitos para instrumentos válidos:

1. **Relevância** (caso especial de um X)
   - Pelo menos um instrumento deve entrar na regressão do primeiro estágio.
2. **Exogeneidade**
   - Todos os instrumentos devem ser não correlacionados com o termo de erro: $\text{corr}(Z_{1i}, u_i) = 0, ..., \text{corr}(Z_{mi}, u_i) = 0$
   - O que acontece se um desses requisitos não é satisfeito? Como verificar? O que fazer?
   - Se você tem múltiplos instrumentos, quais deve usar?

37 Verificando a Hipótese #1

Focamos em um único regressor endógeno incluído:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

Regressão do primeiro estágio:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+k} W_{ki} + v_i$$

- Os instrumentos são **relevantes** se pelo menos um dos $\pi_1, ..., \pi_m$ é não nulo.
- Os instrumentos são **fracos** se todos os $\pi_1, ..., \pi_m$ são zero ou próximos de zero.
- Instrumentos fracos explicam muito pouca variação em X, além daquela explicada pelas W's.

38 Quais são as consequências de instrumentos fracos?

Se os instrumentos são fracos, a distribuição amostral do MQ2E e sua estatística t **não são normais**, mesmo com n grande.

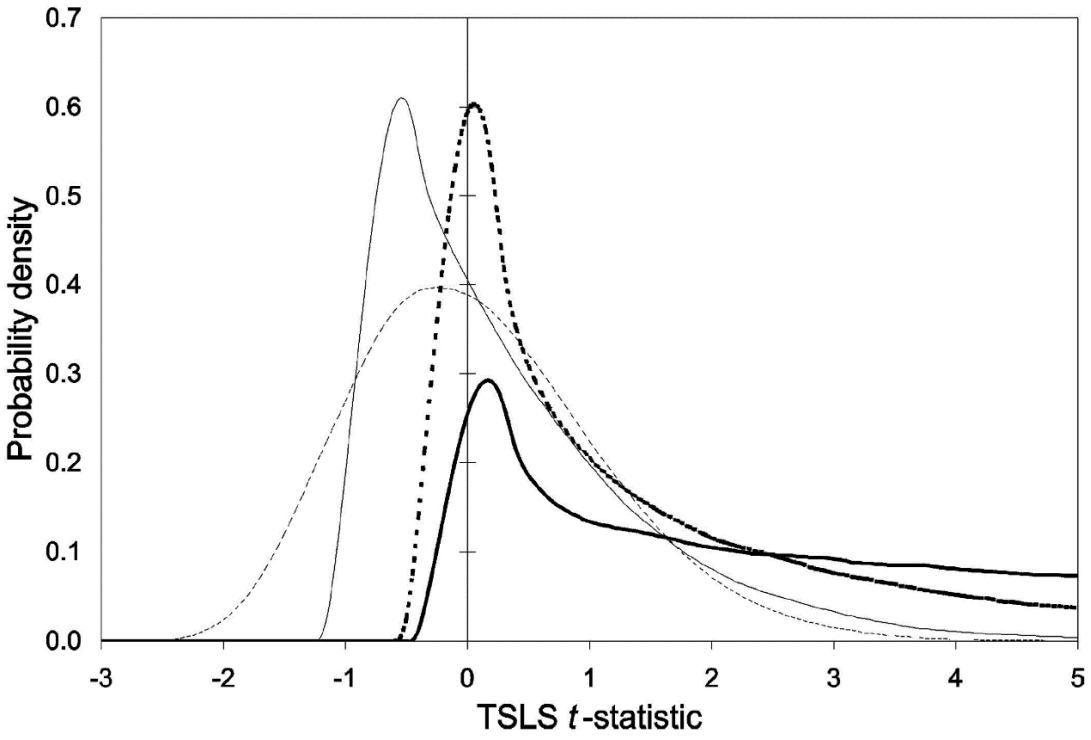
Considere o caso mais simples de um X, um Z e sem variáveis de controle:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad X_i = \pi_0 + \pi_1 Z_i + v_i$$

Se $\text{cov}(X, Z)$ é zero ou pequena, então s_{XZ} será pequena. Com instrumentos fracos, o denominador está próximo de zero:

$$\hat{\beta}_1^{MQ2E} = \frac{s_{YZ}}{s_{XZ}} \Rightarrow \text{denominador} \approx 0 \Rightarrow \text{distribuição altamente não normal}$$

39 Simulação distribuição amostral



- Linha sólida = instrumentos irrelevantes
- Linha tracejada clara = instrumentos fortes

40 Por que a aproximação para normal falha?

Suponha que em uma amostra você calcule $s_{XZ} = 0,00001...$

$$\hat{\beta}_1^{MQ2E} = \frac{s_{YZ}}{s_{XZ}} = \frac{s_{YZ}}{0,00001}$$

Pequenas variações amostrais em s_{YZ} produzem enormes variações em $\hat{\beta}_1^{MQ2E}$!

Se os instrumentos são fracos, os métodos usuais de inferência são não confiáveis

41 Medindo a força dos instrumentos

- **Regressão do primeiro estágio (um X):** Regrida X em $Z_1, ..., Z_m, W_1, ..., W_k$.
- Instrumentos totalmente irrelevantes $\leftrightarrow$ todos os coeficientes de $Z_1, ..., Z_m$ são zero.
- A **estatística F do primeiro estágio** testa a hipótese de que $Z_1, ..., Z_m$ não entram na regressão do primeiro estágio.
- Instrumentos fracos implicam uma estatística F do primeiro estágio pequena.

42 Verificando instrumentos fracos com um X

- Calcule a estatística F do primeiro estágio.

- **Regra prática:** Se a estatística F do primeiro estágio for menor que 10, então o conjunto de instrumentos é fraco.
- Se for o caso, o estimador de MQ2E será viesado, e as inferências estatísticas (erros-padrão, testes de hipótese, intervalos de confiança) podem ser enganosas.

43 Por que 10%?

Por que comparar a estatística F do primeiro estágio com 10?

- Simplesmente rejeitar a hipótese nula de que os coeficientes dos Z's são zero não é suficiente: é necessário um conteúdo preditivo substancial para que a aproximação normal seja boa.
- Comparar a F do primeiro estágio com 10 testa se o viés do MQ2E, relativo ao MQO, é menor que 10%. Se F < 10, o viés relativo supera 10%, ou seja, o MQ2E pode ter viés substancial (ver Apêndice 12.5 de SW).

44 O que fazer se há instrumentos fracos

- Obtenha melhores instrumentos (muitas vezes mais fácil falar do que fazer!).
- Se você tem muitos instrumentos, alguns provavelmente são mais fracos que outros — é uma boa ideia descartar os mais fracos:
  - descartar um instrumento irrelevante aumenta a F do primeiro estágio).
- Se você tem poucos instrumentos, e todos são fracos, então é necessário fazer uma análise VI diferente do MQ2E:
  - Separe o problema de estimação de β_1 e a construção de intervalos de confiança.
  - Isso parece estranho, mas se o MQ2E não é normalmente distribuído, faz sentido (certo?).

45 Verificando a Hipótese #2

- **Exogeneidade dos instrumentos:** todos os instrumentos são não correlacionados com o termo de erro: $\text{corr}(Z_{1i}, u_i) = 0, ..., \text{corr}(Z_{mi}, u_i) = 0$
- Se há mais instrumentos do que regressores endógenos, é possível testar — parcialmente — a exogeneidade dos instrumentos.

46 Testando as Restrições de Sobreidentificação

Considere o caso mais simples:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Suponha que existam dois instrumentos válidos: Z_{1i}, Z_{2i}. Então seria possível calcular duas estimativas de MQ2E separadas.

Intuitivamente, se essas 2 estimativas de MQ2E forem muito diferentes entre si, algo deve estar errado: um ou ambos os instrumentos devem ser inválidos.

O **Teste J de restrições de sobreidentificação** faz essa comparação de maneira estatisticamente precisa.

Isso só pode ser feito se $\#Z > \#X$ (sobreidentificado).

47 O Teste J

Suponha $= m > k$ (sobreidentificado):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

Passos:

1. Calcule os resíduos de MQ2E: $\hat{u}_i = Y_i - \hat{\beta}_0^{MQ2E} - \hat{\beta}_1^{MQ2E} X_{1i} - \cdots$

- Use X_i ao invés de $\hat{X}_i$ para obter $_i

2. Regrida $\hat{u}_i$ em $Z_{1i}, ..., Z_{mi}, W_{1i}, ..., W_{ri}$
3. Compute a estatística F testando a hipótese de que os coeficientes de $Z_{1i}, ..., Z_{mi}$ são todos zero
4. A **estatística J** é $J = mF$

48 O Teste J

$J = mF$, onde F = a estatística F testando os coeficientes de $Z_{1i}, ..., Z_{mi}$ numa regressão dos resíduos de MQ2E em $Z_{1i}, ..., Z_{mi}, W_{1i}, ..., W_{ri}$.

Distribuição da estatística J:

- Sob a hipótese nula de que todos os instrumentos são exógenos, J tem distribuição qui-quadrado com $m - k$ graus de liberdade.
- Se $m = k$, $J = 0$ (isso faz sentido?)
- Se alguns instrumentos são exógenos e outros são endógenos, a estatística J será grande e a hipótese nula de que todos os instrumentos são exógenos será rejeitada.

49 Verificando a validade dos Zs: Resumo

Este resumo considera o caso de um único X. Os dois requisitos para instrumentos válidos são:

1. Relevância

- Pelo menos um instrumento deve entrar na regressão do primeiro estágio.
- Se os instrumentos são fracos, o estimador de MQ2E é viesado e a estatística t tem distribuição não normal.
- Para verificar instrumentos fracos com um único regressor endógeno incluído, verifique a F do primeiro estágio:
  - Se F > 10, os instrumentos são fortes — use MQ2E.
  - Se F < 10, instrumentos fracos — tome alguma providência.

50 Verificando a validade dos Zs: Resumo

2. Exogeneidade

- Todos os instrumentos devem ser não correlacionados com o termo de erro: $\text{corr}(Z_{1i}, u_i) = 0, ..., \text{corr}(Z_{mi}, u_i) = 0$
- Podemos testar parcialmente a exogeneidade: se $m > 1$, podemos testar a hipótese nula de que todos os instrumentos são exógenos, contra a alternativa de que até $m - 1$ são endógenos (correlacionados com u).
- O teste é o **Teste J**, construído usando os resíduos de MQ2E.
- Se o Teste J rejeitar, então pelo menos alguns de seus instrumentos são endógenos — portanto, você deve tomar uma decisão difícil e descartar alguns (ou todos) os seus instrumentos.