Regressão

Angrist e Pischke, capítulo 2

É possível utilizar equações linear típicas dos modelos estimados por MQO para entender como a aleatorização resolve o problema de viés de seleção

Considerar a equação:

$$Y_i = \alpha + \beta D_i + \varepsilon_i \quad (1)$$

A expressão para a média condicional de (1) pode ser escrita como:

$$E[Y_i|D_i] = \alpha + \beta D_i + E[\varepsilon_i|D_i] \quad (2)$$

O que a aleatorização garante é que $E[\varepsilon_i|D_i]=0$ (3), ou seja, o atendimento de umas hipóteses fundamentais para que o modelo de MQO gere um estimador não viesado dos parâmetros da regressão

Devido aos resultados da expressão (3) é possível escrever:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = (\alpha + \beta) - \beta = \beta$$

Embora anteriormente as características observadas estejam incorporadas no termo ε na eq. 1, é possível explicitá-los no modelo sem produzir grandes alterações

Considerar então a eq. (1) modificada pelo introdução do vetor de características observadas X :

$$Y_i = X_i' \alpha + \beta D_i + \varepsilon_i \quad (1')$$

O experimento aleatório assegura, se bem executado, que:

$E[\varepsilon_i | X_i, D_i] = 0$ (3'), o que implica que

$$\begin{aligned} & E[Y_i | X_i, D_i = 1] - E[Y_i | X_i, D_i = 0] \\ &= (X_i' \alpha + \beta) - (X_i' \alpha) = \beta \end{aligned}$$

- Quando não é possível fazer a aleatorização é preciso procurar instrumentos alternativos para estabelecer uma relação causal
- Regressão: o mais básico desse instrumentos

*Make me a match,
run me a regression*

Exemplo: retorno econômico de estudar numa universidade privada de elite

2012-2013: aqueles que foram para uma universidade privada pagaram em média \$29,000 em mensalidades e taxas por ano, enquanto aqueles que foram para uma universidade pública em seu estado natal pagaram menos de \$9,000

Várias razões para a educação privada de elite ser melhor: as classes são menores, a infra-estrutura é melhor, o corpo docente é mais reconhecido e os alunos mais inteligentes

Questão: como \$20,000 por ano de estudo é uma quantidade grande de dinheiro, esta diferença vale a pena?

Harvard x U-Mass

Qual é o *pay-off* de ter um diploma de Harvard?

Comparação de salários dos alunos que frequentaram as duas universidades deve estar contaminada por viés de seleção

Necessário manter constantes as diferenças mais óbvias e importantes entre estudantes que vão para as universidades privadas e as universidades estaduais. Desta forma, espera-se (embora não seja possível prometer) manter as outras coisas iguais

Exemplo numérico com uma amostra pequena para ilustrar a idéia de *ceteris paribus*

Suponha que as únicas coisas que importam para explicar salários sejam os escores no SAT e que universidade o (a) aluno (a) frequentou

O indivíduo A e o indivíduo B têm ambos um escore de 1,400 no SAT (o máximo antes de 2005 era 1,600)

O indivíduo A foi para a U-Mass e o indivíduo B foi para Harvard. Porque assumiu-se que tudo que importa para explicar salários além da escolha da universidade é o escore no SAT , o indivíduo A e o indivíduo B são uma comparação *ceteris paribus*.

Complicação : o indivíduo A é uma mulher jovem e o indivíduo B é um homem jovem

O fato de o indivíduo B ganhar mais do que o indivíduo A pode ser o efeito de uma educação melhor em Harvard, ou pode simplesmente refletir o *gap* salarial homem-mulher resultante de outras coisas

O que queremos é separar o efeito Harvard puro destas outras coisas. Isto é fácil se estas outras coisas que importam é gênero. Basta substituir o indivíduo B por uma aluna mulher de Harvard que também tem um SAT de 1,400, digamos indivíduo C, e então comparar o indivíduo A e o indivíduo C

Como queremos conclusões gerais, devemos calcular o diferencial de salário médio entre alunos de Harvard e da U-Mass com o mesmo gênero e o mesmo escore no SAT

A média de todas as diferenças Harvard x U-Mass grupo-específicas é a estimativa do efeito causal de estudar em Harvard

Dizemos que o estimador controla por, ou mantém fixo, gênero e escores no SAT

- Infelizmente mais coisas explicam os salários que gênero, universidades frequentadas e escores no SAT. Uma vez que decisões de escolha de universidade não são aleatoriamente atribuídas, deve-se controlar todos os fatores que determinam tanto as decisões de que universidade frequentar quanto os salários futuros
- Controlar por uma variedade tão ampla de fatores parece assustador: as possibilidades são virtualmente infinitas e muitas características são difíceis de quantificar, mas esta questão discutiremos depois
- Começar com uma única variável de controle

Ingredientes chave de modelos de regressão

- 1) A variável dependente, neste caso, os salários do estudante i , também chamada variável de resultado Y_i
- 2) A variável de interesse, neste caso, a variável *dummy* que indica os estudantes que foram para uma universidade privada (P_i)
- 3) Um conjunto de variáveis de controle, no caso, uma única variável que é a variável resumo que identifica conjuntos de escolas para as quais os alunos aplicaram e foram admitidos

Entendendo um pouco da variável de controle (denominada por A)

Dale e Krueger (2002) : ao invés de identificar tudo que importa para a escolha da universidade e para os salários, trabalharam com uma medida resumo chave: as características das universidades para as quais os alunos aplicaram e foram aceitos

Organizaram os alunos em grupos definidos pelo conjunto de escolas para as quais eles aplicaram e foram admitidos. Dentro de cada grupo, os alunos têm provavelmente ambições de carreira similares, enquanto eles são também julgados como tendo habilidades semelhantes pelo *staff* de admissão das escolas para as quais eles aplicaram

Comparações dentro do grupo deveriam, portanto, ser consideravelmente mais comparações de coisas iguais do que comparações não controladas envolvendo todos os estudantes

A regressão de interesse pode ser escrita como:

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i \quad (1)$$

- Análise de regressão atribui valores aos parâmetros (α, β, γ) de tal forma a tornar $\hat{Y}_i$ tão próximo quanto possível de Y_i
- Isto é feito escolhendo-se valores que minimizam a soma dos quadrados dos resíduos

- Dizemos que estamos estimando os parâmetros da regressão quando executamos essa minimização para uma amostra particular
- É uma convenção reportar estimativas de regressão em quase toda investigação empírica de efeitos causais. Estimativas de regressão fornecem um *benchmark* para técnicas mais sofisticadas
- Sob algumas circunstâncias as estimativas de regressão são eficientes no sentido de que fornecem as estimativas estatisticamente mais precisas dos efeitos causais médios

Ceteris paribus?

Regressão é um jeito de fazer as outras coisas iguais, mas igualdade é gerada somente para as variáveis incluídas como controle do lado direito do modelo. Falha em incluir controles suficientes ou os controles corretos ainda nos deixa com um viés de seleção

Versão de regressão do viés de seleção é chamado viés de variável omitida (OVB – *omitted variable bias*)

Qual o viés de omitir a variável grupo ? A inclusão de A_i importa para as estimativas para o efeito de escolha privada na regressão acima?

A “regressão longa” inclui a variável *dummy*, A_i , que indica aqueles no grupo A.

O modelo de regressão que inclui A_i pode ser escrito como :

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l \quad (2)$$

Supor que estimamos uma regressão sem controle. Isto pode ser escrito como:

$$Y_i = \alpha^s + \beta^s P_i + e_i^s \quad (3)$$

A conexão formal entre coeficientes da regressão curta e longa tem dois componentes:

- 1) A relação entre a variável omitida (A_i) e a variável de interesse (P^i); em breve veremos como quantificar isto com uma regressão adicional.
- 2) A relação entre a variável omitida (A^i) e a variável de resultado (Y_i). Isto é dado pelo coeficiente da variável omitida na regressão longa, neste caso, o parâmetro γ na equação (2).

Fórmula do viés de variável omitida

Efeito de P_i no curto = Efeito de P_i no longo + ({Relação entre omitida e incluída}) x {Efeito da omitida no longo}

Para ser específico, quando a variável omitida é A_i e a variável de tratamento é P_i temos:

Efeito de P_i no curto = Efeito de P_i no longo + ({Relação entre A_i e P_i }) x {Efeito de A_i no longo}

O viés de variável omitida, definido como a diferença entre o coeficiente de P_i nos modelos de curto e longo prazo, é um simples rearranjo desta equação.

Viés de variável omitida = {Relação entre A_i e P_i } x {Efeito de A_i no longo}

É possível refinar a fórmula de viés de variável omitida usando o fato de que ambos os termos na fórmula são eles próprios coeficientes de regressão. O primeiro termo é o coeficiente de uma regressão da variável omitida A_i na *dummy* de escola privada, ou seja, é o coeficiente π_1 no modelo de regressão

$$A_i = \pi_0 + \pi_1 P_i + u_i$$

Onde u_i é o resíduo

Agora podemos escrever a fórmula do viés de variável omitida como:

$$\begin{aligned} \text{Viés de variável omitida} &= \text{Efeito de } P_i \text{ no curto} - \text{Efeito de } P_i \text{ no longo} = \beta^s - \beta^l \\ &= \pi_1 \times \gamma, \end{aligned}$$

Onde γ é o coeficiente de A_i na regressão longa

Exemplo

π_1 é 0,16667, $\gamma = 60,000$, refletindo os salários mais altos do grupo A.

Juntando tudo tem-se:

$$\begin{aligned}\text{Viés de variável omitida} &= \text{Curto} - \text{longo} \\ &= \beta^s - \beta^l \\ &= 20,000 - 10,000\end{aligned}$$

e Viés de variável omitida = {Regressão da omitida na incluída} x {Efeito da omitida no longo} = $\pi_1 \times \gamma = 0,1667 \times 60,000 = 10,000$

Por que a omissão da variável A muda tanto o efeito do ensino privado?
Por que $\beta^s = 20.000$ e $\beta^l = 10.000$, ou seja, o viés de omissão é tão grande (10.000)?

A renda média dos estudantes no grupo A é maior do que a dos estudantes no grupo B. Além disso, $\frac{2}{3}$ dos estudantes no grupo A frequentam uma universidade privada, enquanto no grupo B somente $\frac{1}{2}$. As diferenças nos salários entre os ex-alunos de universidades privadas e públicas resulta em parte do fato de que a maioria dos estudantes do grupo A têm uma renda mais alta de qualquer forma, independente de onde estudaram.

- Fórmula do viés de variável omitida é um resultado matemático que explica as diferenças entre os coeficientes de regressão em qualquer cenário curto versus longo, a despeito da interpretação causal dos parâmetros da regressão
- A nomenclatura curto e longo é puramente relativa. A regressão curta não precisa ser necessariamente curta, mas a regressão longa é sempre mais longa, uma vez que ela incluiu os mesmos regressores e pelo menos mais um

- Frequentemente, as variáveis adicionais que tornam a regressão longa são hipotéticas, ou seja, não estão disponíveis nos nossos dados
- A fórmula de viés de omissão de variáveis é um instrumento que nos permite considerar o impacto de controlar para variáveis que gostaríamos de ter

- Não podemos usar dados para verificar as consequências de variáveis omitidas que não observamos, mas é possível usar a fórmula de omissão de variáveis para fazer uma *educated guess* com relação às possíveis consequências de suas omissões
- Porque nós nunca podemos ter certeza se um dado conjunto de controles é suficiente para eliminar o viés de seleção, é importante perguntar quão sensíveis são os resultados à mudanças na lista de controles. Nossa confiança nas estimativas dos efeitos causais cresce quando os efeitos da variável de interesse são insensíveis – robustos – a se uma variável particular é adicionada ou dropada, desde que uns poucos controles centrais sejam sempre incluídos no modelo

Econometria I: Variáveis demais ou variáveis de menos?

- O que acontece se incluirmos na nossa especificação variáveis que não deveriam ser incluídas?
- Não há efeitos sobre as estimativas dos parâmetros e MQO permanece não viesado
- O que acontece se excluirmos da nossa especificação uma variável que deveria ser incluída?

Modelo populacional verdadeiro é

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

Supor que o interesse é β_1 , o efeito parcial de x_1 sobre y

Por ignorância ou indisponibilidade de dados estima-se

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

~ é utilizado para enfatizar que $\tilde{\beta}_1$ vem de um modelo subespecificado

Derivar o valor esperado de $\tilde{\beta}_1$

Derivar essa esperança não é difícil uma vez que $\tilde{\beta}_1$ é o estimador de inclinação de MQO de uma regressão simples

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1) y_i}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

Recall the true model, so that

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, so the numerator becomes

$$\begin{aligned} \sum (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) = \\ \beta_1 \sum (x_{i1} - \bar{x}_1)^2 + \beta_2 \sum (x_{i1} - \bar{x}_1)x_{i2} + \sum (x_{i1} - \bar{x}_1)u_i \end{aligned}$$

$$\tilde{\beta} = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1)x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)} + \frac{\sum (x_{i1} - \bar{x}_1)u_i}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

since $E(u_i) = 0$, taking expectations we have

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1)x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

Consider the regression of x_2 on x_1

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \text{ then } \tilde{\delta}_1 = \frac{\sum (x_{i1} - \bar{x}_1)x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

$$\text{so } E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$

Resumo da direção do viés

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Viés positivo	Viés negativo
$\beta_2 < 0$	Viés negativo	Viés positivo

- Dois casos em que o viés é zero
 $\beta_2 = 0$, ou seja, x_2 realmente não deveria estar no modelo
 x_1 e x_2 são não correlacionados na amostra

Variâncias na presença de variáveis omitidas

Modelo populacional verdadeiro

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

Considerar dois estimadores de β_1

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + u \quad (1)$$

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \quad (2)$$

- Quando $\beta_2 \neq 0$, a equação (2) exclui uma variável relevante do modelo e, como visto antes, isso induz um viés em $\tilde{\beta}_1$, a não ser que x_1 e x_2 sejam não correlacionados
- $\hat{\beta}_1$ é não viesado para β_1 , para qualquer valor de β_2 , incluindo $\beta_2 = 0$
- Se o viés for usado como único critério $\hat{\beta}_1$ é sempre preferível a $\tilde{\beta}_1$

$$\text{var}(\hat{\beta}_1) = \sigma^2 [SQT_1(1-R_i^2)]$$

$$\text{var}(\tilde{\beta}_1) = \sigma^2 / SQT_1$$

Em que SQT_1 é a variação total em x_1 e R_1^2 é o R quadrado da regressão de x_1 e x_2

- Quando $\beta_2 \neq 0$, $\tilde{\beta}_1$ é viesado, $\hat{\beta}_1$ é não viesado e $var(\tilde{\beta}_1) < var(\hat{\beta}_1)$
- Quando $\beta_2 = 0$, $\tilde{\beta}_1$ e $\hat{\beta}_1$ são ambos não viesados e $var(\tilde{\beta}_1) < var(\hat{\beta}_1)$

Artigo citado

Dale, S.B., Krueger, A.B. 2002. Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables. *Quarterly Journal of Economics*, vol. 117, pp. 1491-1527