ACH3657

Métodos Quantitativos para Avaliação de Políticas Públicas

Aula teórica 05

Valores esperados e variâncias dos estimadores de MQO

Alexandre Ribeiro Leichsenring
alexandre.leichsenring@usp.br

Organização

1. Valores esperados e variâncias dos estimadores de MQO
   - Inexistência de Viés em MQO

Valores esperados e variâncias dos estimadores de MQO

- Até agora, definimos modelo populacional:

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

- Afirmamos hipótese fundamental para utilidade da regressão:

$$\mathbf{E}(u|x) = 0$$

  para qualquer valor de x.

- Veremos agora propriedades estatísticas dos estimadores de MQO:
  - $\hat{\beta}_0$ e $\hat{\beta}_1$ como *estimadores* de β_0 e β_1
  - Propriedades da distribuição de $\hat{\beta}_0$ e $\hat{\beta}_1$ de diferentes amostras aleatórias da população

Inexistência de Viés em MQO

- Vamos estabelecer a "inexistência de viés" do modelo sob um conjunto simples de hipóteses.
- As hipóteses serão numeradas para referências futuras.

Hipótese RLS.1 (Linear nos parâmetros)

No modelo populacional, a variável dependente y está relacionada à variável independente x e ao erro (ou perturbação) u como

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

em que β_0 e β_1 são os parâmetros de intercepto e de inclinação populacionais, respectivamente.

Hipótese RLS.2 (Amostragem Aleatória)

Temos uma amostra aleatória de tamanho n

$$\{(x_i, y_i) : 1 = 1, 2, ..., n\}$$

proveniente do modelo populacional especificado em RLS.1.

- Nem todas as amostras de corte transversal podem ser vistas como resultados de amostras aleatórias, mas muitas podem ser assim entendidas.
  Podemos escrever (1), em termos da amostra aleatória como:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

  em que
  - u_i é o erro ou perturbação da observação i (por exemplo, pessoa i, empresa i, cidade i, etc.)
  - u_i contém os fatores não-observáveis da observação i que afetam y
  - O u_i não devem ser confundidos com os resíduos (do modelo estimado)

Hipótese RLS.3 (Variação amostral na variável independente)

Na amostra, as variáveis independentes $x_i, i = 1, 2, \ldots, n$ não são todas iguais a uma mesma constante.

- Isso exige alguma variação em x na população
- Das quatro hipóteses feitas, esta é a menos importante, pois ela essencialmente nunca falha em aplicações interessantes
- Se a hipótese RLS.3 não se sustentar, não podemos calcular os estimadores de Mínimos Quadrados, o que significa que a análise estatística é irrelevante

Hipótese RLS.4 (Média condicional zero)

Dado qualquer valor da variável explicativa x, a média condicional do termo de erro u é zero, ou seja:

$$\mathbf{E}(u|x) = 0$$

- Para uma amostra aleatória, essa hipótese implica que

$$\mathbf{E}(u_i|x_i) = 0, \text{ para todo } i = 1, 2, ..., n$$

- Restringe a relação entre u e x na população
- A hipótese de média condicional zero - juntamente com a hipótese de amostra aleatória - permite uma simplificação técnica conveniente

Teorema

Inexistência de viés nas estimativas por mínimos quadrados

Usando as hipóteses RLS.1 a RLS.4,

$$\mathbf{E}(\hat{\beta}_0) = \beta_0, \text{ e } \mathbf{E}(\hat{\beta}_1) = \beta_1$$

para quaisquer valores de β_0 e β_1.

➤ Em outras palavras, $\hat{\beta}_0$ é não-viesado para β_0, e $\hat{\beta}_1$ é não-viesado para β_1.

Exemplo: Desempenho em Matemática e o Programa de Merenda Escolar

Seja mate10 a percentagem de alunos do primeiro ano do ensino médio aprovados em um exame de matemática. Suponha que desejamos estimar o efeito do programa de merenda escolar financiado pelo governo sobre o desempenho dos alunos. Esperamos que o programa de merenda tenha um efeito *ceteris paribus* positivo sobre o desempenho: todos os outros fatores permanecendo iguais, se um estudante pobre torna-se beneficiário de um programa de merenda e passa a ter refeições regularmente, seu desempenho deveria melhorar. Seja prgalm a percentagem de estudantes participantes do programa de merenda escolar em cada escola. Portanto, o modelo de regressão simples é

$$mate10 = \beta_0 + \beta_1 \, prgalm + u, \tag{2}$$

em que u contém características da escola e do estudante que afetam o desempenho escolar total.

O modelo ajustado é:

$$\hat{mate}10 = 32,14 - 0,319\ prgalm,$$

- A equação prevê que se a participação dos estudantes no programa de merenda escolar aumenta em dez pontos percentuais, a percentagem de estudantes que passa no exame de matemática cai cerca de 3,2 pontos percentuais.
- Realmente devemos acreditar que a participação maior no programa de merenda escolar causa, de fato, um desempenho pior?
- Muito provavelmente não.
- Uma explicação melhor é que o termo erro u na Equação (2) esteja correlacionado com prgalm.
- De fato, u contém fatores como a taxa de pobreza das crianças que frequentam a escola, que afeta o desempenho dos estudantes e está altamente correlacionada com a qualificação no programa de merenda.
- Variáveis como qualidade e recursos da escola também estão contidas em u e, provavelmente, estão correlacionados com prgalm.
- É importante lembrar que a estimativa -0,319 é somente para essa amostra particular, mas seu sinal e magnitude nos fazem suspeitar de que u e x sejam correlacionadas, de modo que a regressão linear é viesada.

Exercício

Seja filhos o número de filhos de uma mulher e educ os anos de educação da mulher. Um modelo simples que relaciona a fertilidade a anos de educação é

$$filhos = \beta_0 + \beta_1 \, educ + u \tag{3}$$

em que u é um erro não-observável.

i) Que tipos de fatores estão contidos em u? É provável que eles estejam correlacionados com o nível de educação?

ii) Uma análise de regressão simples mostrará o efeito *ceteris paribus* da educação sobre a fertilidade? Explique.

Variâncias dos estimadores de Mínimos Quadrados

- Além de saber que a distribuição amostral de β_1, está centrada em torno de β_1, ($\hat{\beta}_1$, é não-viesado), é importante saber o quão distante, em média, podemos esperar que $\hat{\beta}_1$, esteja de β_1.
- Isso está relacionado com a variância σ^2 dos termos de erro u.

Hipótese RLS.5 (Homocedasticidade)

$$Var(u|x) = \sigma^2$$

- Essa hipótese afirma que a variância do termo não-observável u, condicionado a x, é constante.
- Ela é conhecida como a hipótese de homoscedasticidade ou de "variância constante".
- A hipótese de homoscedasticidade é completamente distinta da hipótese de média condicional zero, $\mathbf{E}(u|x) = 0$.
  - A hipótese RLS.4 compreende o valor esperado de u
  - a hipótese RLS.5 diz respeito à variância de u (ambos condicionados a x)

Exemplo: Heterocedasticidade em uma equação de salários

A fim de obter um estimador não-viesado do efeito *ceteris paribus* de educ sobre salárioh, devemos assumir que $\mathbf{E}(u|educ) = 0$, e isso implica

$$\mathbf{E}(salarioh|educ) = \beta_0 + \beta_1 educ$$

Se também usarmos a hipótese de homoscedasticidade, então

$$Var(u|educ) = \sigma^2$$

não depende do nível de educação formal, que é o mesmo que assumir que

$$Var(salrioh|educ) = \sigma^2$$

Assume-se que a variabilidade no salário horário em torno de sua média é constante através de todos os níveis de educação formal: isso pode não ser realista.

É provável que pessoas com maior nível educacional tenham uma variedade maior de interesses e de oportunidades de trabalho, o que poderia levar a uma variabilidade maior nos níveis de educação mais elevados.

Pessoas com níveis de escolaridade baixos têm poucas oportunidades e, freqüentemente, trabalham recebendo salário mínimo (isso tem o efeito de reduzir a variabilidade salarial nos níveis baixos de educação formal)

Se a hipótese RLS.5 se mantém ou não é uma questão empírica.

Var(salario | educ) crescendo com educação