Conceitos Básicos de Estatística, Regressão Linear, Testes de Hipótese e Arcabouço de Resultados Potenciais

Heitor Lima

Métodos Quantitativos para Avaliação de Políticas Públicas - MPP - 2023
Professora: Letícia Nunes

Insper Instituto de Ensino e Pesquisa

heitoraol@al.insper.edu.br

Informações Gerais

1. Monitorias: quartas-feiras, das 14:00hs às 16:00hs, sala Mario Haberfeld
   - Estes slides (e os próximos) são baseados no material de Fabiano Dal-Ri
2. Listas de exercícios: Individuais, software Stata, entrega via Blackboard
3. Trabalho final
4. Links úteis
   - Playlist (reduzida) do *Mastering 'Metrics* no YouTube: Link
   - Artigos e dados citados ao longo do *Mastering 'Metrics*: Link

Conceitos Básicos de Estatística

Conceitos Básicos de Estatística

Se $x_1, \ldots, x_n$ são todos os n valores (distintos ou não) da variável aleatória X, a esperança de X pode ser escrita como

$$\mathbb{E}[X] = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Para representar a variabilidade do conjunto de observações, definimos a variância de X como

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Definimos, ainda, o desvio-padrão como

$$dp(X) = \sqrt{Var(X)}$$

Conceitos Básicos de Estatística

Avaliações de políticas buscam, frequentemente, entender a associação entre variáveis quantitativas

Dados n pares de valores $(x_1, y_1), \ldots, (x_n, y_n)$, chamaremos de covariância entre as duas variáveis aleatórias X e Y a

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \end{aligned}$$

Dada esta definição, o coeficiente de correlação pode ser escrito como

$$Corr(X, Y) = \frac{Cov(X, Y)}{dp(X) \cdot dp(Y)}$$

População e Amostra

Definição: *População* é o conjunto de todos os elementos sob investigação. *Amostra* é qualquer subconjunto da população

Usaremos a amostra para *aproximar* os parâmetros desconhecidos de uma população:

$$\mathbb{E}[X] \to \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{e} \quad Var(X) \to S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

- $\bar{X}$ é a média amostral
- S^2 é a variância amostral não-viesada
  - Usa-se $n-1$ ao invés de n por convenção (Angrist e Pishcke (2015, p. 36))

População e Amostra

As aproximações (ou estimativas) irão depender fundamentalmente da amostra que tivermos

- $\bar{X}$ e S^2 também são variáveis aleatórias, i.e., variam para cada amostra

Algumas leis e propriedades estatísticas nos indicam para onde as estimativas estão apontando

Lei dos Grandes Números

"A média amostral pode ser aproximada da média populacional tanto quanto se queira, ao simplesmente aumentar o tamanho da amostra."

Regressão Linear Simples

Seguindo a notação do *Mastering 'Metrics*, vamos indexar as variáveis em i: X_i, Y_i, ...

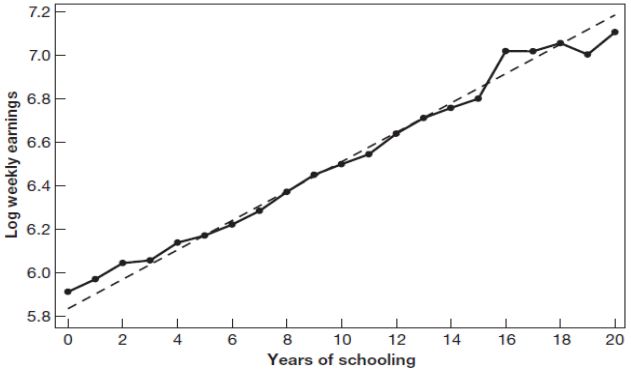
Além da esperança amostral, $\mathbb{E}[X_i]$, também estamos interessados na esperança condicional,

$$\mathbb{E}[Y_i|X_i = x]$$

- Lê-se: "esperança condicional de Y_i dado que X_i assume o valor específico x"
- Ideia: para cada valor x de X_i, há uma esperança diferente para Y_i

Regressão Linear Simples

FIGURE 2.1
The CEF and the regression line



Notes: This figure shows the conditional expectation function (CEF) of log weekly wages given years of education, and the line generated by regressing log weekly wages on years of education (plotted as a broken line).

Obtido de Angrist e Pischke (2015)

- Linha sólida: Função de esperança condicional de (log do) salário, dados os anos de educação
- Linha pontilhada: Reta de regressão do (log do) salário em anos de educação

Regressão Linear Simples

$$\mathbb{E}[Y_i|X_i] = \alpha + \beta X_i + \varepsilon_i$$

- α é o intercepto da reta
- β é o coeficiente da variável explicativa X_i (e também a inclinação da reta)
- ε_i é o resíduo

Regressão linear sempre assume que a FEC é linear; ainda que não seja, regressão linear é uma boa aproximação

Ao minimizar o quadrado da soma dos resíduos, $\{\mathbb{E}[Y_i - (\alpha + \beta X_i)]\}^2$, encontramos

$$\beta = \frac{Cov(Y_i, X_i)}{Var(X_i)} \quad \text{e} \quad \alpha = \mathbb{E}[Y_i] - \beta\,\mathbb{E}[X_i]$$

Não conhecemos os parâmetros populacionais: $\mathbb{E}[X_i]$, $Var(X)$, $Cov(Y_i, X_i)$

Por isso, ao usar amostras para estimá-los, estamos sujeitos às variações amostrais

Se a estimativa amostral de β, chamada de $\hat{\beta}$, for igual a b, não se pode garantir que $\beta = b$

É aqui que entram os conceitos de erro-padrão e teste de hipóteses

Erro-padrão

O erro-padrão é o desvio-padrão de uma estatística, e mede sua variabilidade devido à amostragem aleatória

- Desvio-padrão: medida de dispersão de uma variável aleatória
- Estatística: função de uma variável aleatória
- Erro-padrão: medida de dispersão de uma estatística

O erro-padrão de uma estimativa também precisa ser estimado. Nossos principais interesses:

- Estatística $\hat{\beta}$
- Seu erro-padrão estimado, $\hat{SE}(\hat{\beta})$

Teste de Hipóteses

Estamos interessados em um teste que verifica se β é igual a um determinado valor β_0

Chamamos tal hipótese de "hipótese nula":

$$H_0 : \beta = \beta_0$$

Dada a hipótese nula, calculamos a estatística t como

$$t = \frac{\hat{\beta} - \beta_0}{\hat{SE}(\hat{\beta})},$$

que segue uma distribuição t de Student com $n - 1$ graus de liberdade

- É por isso que alguns chamam o teste de hipótese de "teste t"

Teste de Hipóteses

Nível de significância α, que dá a probabilidade de se rejeitar H_0 quando ela é verdadeira

- Geralmente, considera-se $\alpha = 5\%$

Considerando H_0 ao nível de 5%, espera-se que a estatística t fique aproximadamente entre -2 e 2

- Se $|t| > 2$, dizemos que $\hat{\beta}$ é significativamente diferente de β_0

Importante: A hipótese nula mais comum é $H_0 : \beta = 0$

- Testar se o coeficiente é *estatisticamente significativo*
- O cálculo da estatística t torna-se um pouco diferente

Observações

- Focamos na regressão simples, mas a intuição é a semelhante para a regressão múltipla (duas ou mais variáveis explicativas)
  - Porém, algumas fórmulas mudam
- Geralmente, o erro-padrão é substituído pelo *erro-padrão robusto*, que leva em consideração a possibilidade de heterocedasticidade
- Se um coeficiente é estatisticamente não-significativo, não quer dizer necessariamente que não há relação entre as variáveis
  - Pode ser que não temos precisão estatística o suficiente para identificar o efeito

Arcabouço de Resultados Potenciais

Arcabouço de Resultados Potenciais

[...] people are either insured or not. We don't get to see them both ways, at least not at the same time in exactly the same circumstances.

Acts demolish their alternatives, that is the paradox.

We can't know what lies at the end of the road not taken.

The problem is other things equal, or lack thereof.

— **'Mastering Metrics**, ch. 1

Arcabouço de Resultados Potenciais

Usando a notação de Angrist e Pischke (2015), temos que

- D_i: indica se o indivíduo i foi tratado ($D_i = 1$) ou não ($D_i = 0$)

- Y_{0i}: indica o resultado potencial para o indivíduo i na ausência do tratamento
  - Independentemente de i ter recebido o tratamento ou não
- $[Y_{0i}|D_i = 0]$: indica o resultado potencial para i na ausência do tratamento, dado que i não recebeu o tratamento
  - Observado!
- $[Y_{0i}|D_i = 1]$: indica o resultado potencial para i na ausência do tratamento, dado que i recebeu o tratamento
  - Não observado!

Arcabouço de Resultados Potenciais

Usando a notação de Angrist e Pischke (2015), temos que

- D_i: indica se o indivíduo i foi tratado ($D_i = 1$) ou não ($D_i = 0$)

- Y_{1i}: indica o resultado potencial para o indivíduo i na presença do tratamento
  - Independentemente de i ter recebido o tratamento ou não
- $[Y_{1i}|D_i = 0]$: indica o resultado potencial para i na presença do tratamento, dado que i não recebeu o tratamento
  - Não observado!
- $[Y_{1i}|D_i = 1]$: indica o resultado potencial para i na presença do tratamento, dado que i recebeu o tratamento
  - Observado!

Arcabouço de Resultados Potenciais

Efeito Médio do Tratamento sobre os Tratados (ATT):

$$\begin{aligned} ATT &= \mathbb{E}[Y_{1i} - Y_{0i}] \\ &= \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1] \\ &= \mathbb{E}[Y_{1i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 1] \end{aligned}$$

Problema: Não observamos $\mathbb{E}[Y_{0i} | D_i = 1]$

- E se vacinados não tivessem sido vacinados?
- E se quem recebeu seguro-desemprego não tivesse recebido?
- E se quem usa cinto de segurança não usasse?

Arcabouço de Resultados Potenciais: Viés de Seleção

Uma possível solução seria usar $[Y_{0i}|D_i = 0]$, que é observado

Porém, geralmente, não se pode dizer que $[Y_{0i}|D_i = 1]$ é igual a $[Y_{0i}|D_i = 0]$

Viés de seleção:

- Quem decide se vacinar se cuida tanto quanto quem decide não se vacinar?
- Quem recebe seguro-desemprego possui a mesma rede de suporte social de alguém que não recebe?
- Quem usa cinto de segurança é mais ou menos cauteloso do que quem não usa?

Arcabouço de Resultados Potenciais

Experimento Aleatório Controlado (RCT): Faz com que $[Y_{0i}|D_i = 1]$ seja igual a $[Y_{0i}|D_i = 0]$

- Neste caso, podemos calcular o ATT diretamente como

$$\begin{aligned} ATT &= \mathbb{E}[Y_{1i} - Y_{0i}] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \end{aligned}$$

O curso abordará outros métodos além do RCT:

- Variáveis instrumentais
- Diferença-em-Diferenças
- Regressão com Descontinuidade
- Pareamento

Referências

Referências

Angrist, J., e J. Pischke (2015). *Mastering 'Metrics: The Path from Cause to Effect*. New Jersey, US: Princeton University Press.

Bussab, W., e P. Morettin (2017). *Estatística Básica*. 9 ed. São Paulo: Saraiva.

Wooldridge, J. (2013). *Introductory Econometrics: A Modern Approach*. 5th ed. Andover, UK: Cengage.

Notas de aula da Profª. Letícia Nunes.