Econometria I

O estimador de Mínimos Quadrados Ordinários

Luis A. F. Alvarez

10 de abril de 2025

Ambiente

- Pesquisador observa n pares (Y_i, X_i), $i = 1 \ldots, n$, para os quais supõe um modelo linear da forma:

$$Y_i = X_i'\beta + \epsilon_i \ \ldots i = 1, \ldots n\,, \tag{1}$$

onde $\beta \in \mathbb{R}^k$ é um parâmetro desconhecido, e ϵ_i, $i = 1, \ldots, n$ são variáveis aleatórias não observadas.
  - Recorde-se, da última aula, que cabe ao pesquisador postular o modelo e a interpretação dos coeficientes.
  - No caso mais comum, $(Y_i, X_i) \overset{d}{=} (Y, X)$ para $i = 1, \ldots, n$, onde a distribuição de (Y, X) representa a distribuição das variáveis numa população de interesse.
    - Por exemplo, podemos ter que $\{(Y_i, X_i)\}_{i=1}^n$ é uma amostra aleatória de uma população com distribuição $\mathbb{P}_{Y,X}$, para a qual postulamos um modelo linear.
    - Mas também podemos ter que as observações entre pares apresentem dependência entre si, embora com leis $\mathbb{P}_{(Y_i,X_i)}$ comuns a todo i.
  - De modo mais geral, no entanto, pode ser que os (Y_i, X_i) não possuam a mesma distribuição conjunta, mas haja uma relação comum e estável ao longo de i.

Notação matricial

- No que segue, definimos as seguintes matrizes aleatórias:

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Com base na notação acima introduzida, podemos reescrever (1) em notação matricial como:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \tag{2}$$

Estimador de mínimos quadrados ordinários

- O estimador de mínimos quadrados ordinários de β, denotado por $\hat{b}$, consiste em estimar β minimizando a distância, na norma Euclidiana, entre $\boldsymbol{y}$ e uma combinação linear das colunas de $\boldsymbol{X}$, i.e.

$$\hat{b} \in \operatorname{argmin}_{b \in \mathbb{R}^k} \|\boldsymbol{y} - \boldsymbol{X} b\|_2^2 = \operatorname{argmin}_{b \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 \,,$$

  - Em outras palavras, encontramos o coeficiente b que maximiza a contribuição dos X_i à explicação de Y_i, tal qual medida pela média da distância ao quadrado entre os Y_i e $X_i' b$.
- Estimador é obtido como solução a uma otimização de uma função convexa, sem restrições $\implies$ condição de primeira ordem caracteriza o conjunto de soluções.
- Condições de primeira ordem podem ser escritas como:

$$\mathbf{0}_{k \times 1} = \sum_{i=1}^{n} X_i (Y_i - X_i' b) = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X} b)$$

Condição de posto e unicidade do mínimo

Sob a condição

Hipótese (H1-POSTO)

A matriz $\boldsymbol{X}$ apresenta posto k.

Temos que $\boldsymbol{X}'\boldsymbol{X}$ é invertível (por quê), de modo que existe uma única solução ao problema de otimização, dada por:

$$\hat{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y}) .$$

- Condição de posto requer que nenhuma das colunas seja escrita como combinação linear das demais.
  - Se a primeira entrada dos X_i corresponde a um intercepto (i.e. $X_{i,1} = 1$ para todo $i = 1, \ldots, n$), nenhuma das colunas pode ser escrita como função afim das demais
- Observe que, como $\text{rank}(\boldsymbol{X}) \leq \min\{n, k\}$, condição implica que $n \geq k$.

Um caso simples

- Considere, para fixar as ideias, o caso em que $X_i = \begin{bmatrix} 1 & T_i \end{bmatrix}'$.
- Nesse caso, a matriz $\boldsymbol{X}'\boldsymbol{X}$ é dada por:

$$\begin{bmatrix} n & \sum_{i=1}^{n} T_i \\ \sum_{i=1}^{n} T_i & \sum_{i=1}^{n} T_i^2 \end{bmatrix} \tag{3}$$

  de modo que a condição de posto é equivalente a $\widehat{V(T)} = \frac{1}{n}\sum_{i=1}^{n} T_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} T_i\right)^2 > 0$.
- Se condição de posto é satisfeita, estimador de MQO é dado por:

$$\hat{b}_2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(T_i - \bar{T})}{\sum_{i=1}^{n}(T_i - \bar{T})^2} = \frac{\widehat{\text{cov}(T, Y)}}{\widehat{V(T)}}$$

MQO: Propriedades Algébricas

Matriz de projeção

- Definimos a matriz de projeção de $\boldsymbol{X}$ como:

$$P = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

  - Observe que a matriz de projeção é tal que $X\hat{b} = P\boldsymbol{y}$.
  - $P\boldsymbol{z}$ é a projeção (em termos de minimização da distância Euclidiana) de $\boldsymbol{z} \in \mathbb{R}^n$ no espaço gerado pelas colunas de $\boldsymbol{X}$.
- Matriz de projeção tem as seguintes propriedades:
  - Simétrica.
  - Idempotente ($P^2 = P$).
  - Os autovalores de P são ou 0 ou 1.
  - $\text{trace}(P) = k = \text{rank}(P)$.

Matriz residualizadora

- A matriz residualizadora (*residual-maker*) de X é dada por:

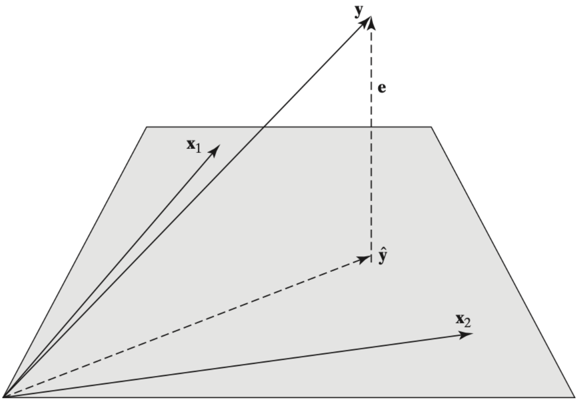
$$M = (I - P)$$

- Matriz residualizadora tem as seguintes propriedades:
  - Simétrica.
  - Idempotente ($M^2 = M$).
  - Os autovalores de M são ou 0 ou 1.
  - $\text{trace}(M) = n - k = \text{rank}(M)$.
  - $\boldsymbol{X}'M = 0$, $M\boldsymbol{X} = 0$, $PM = MP = 0$.
- Matriz residualizadora devolve o erro de projeção $\boldsymbol{z} - P\boldsymbol{z}$.
  - Como $P\boldsymbol{z}$ é o minimizador da distância Euclidiana no espaço gerado pelas colunas de $\boldsymbol{X}$,temos que:

$$(\boldsymbol{Mz}) \cdot (\boldsymbol{X\ell}) = \boldsymbol{z}'M\boldsymbol{X\ell} = 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^n, \boldsymbol{\ell} \in \mathbb{R}^k \,.$$

Visualização gráfica

FIGURE 3.2 Projection of **y** into the Column Space of **X.**



- Suponha que particionemos a matriz $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}$ onde $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ são matrizes de dimensão $n \times k_1$ e $n \times k_2$.
- Nesse caso, a fórmula da inversa particionada nos indica que:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \boldsymbol{X_1}'\boldsymbol{X}_1 & \boldsymbol{X_1}'\boldsymbol{X}_2 \\ \boldsymbol{X_2}'\boldsymbol{X}_1 & \boldsymbol{X_2}'\boldsymbol{X}_2 \end{bmatrix}^{-1} =$$

$$\begin{bmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{F}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} & -(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{F} \\ -\boldsymbol{F}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} & \boldsymbol{F} \end{bmatrix}$$

onde $\boldsymbol{F} = (\boldsymbol{X}_2'\boldsymbol{X}_2 - \boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2)^{-1} = (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}$ onde $\boldsymbol{M}_1$ é a residualizadora de $\boldsymbol{X}_1$.

  - Inversa de $\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2$ existe pois matriz $\boldsymbol{X}$ tem posto cheio.

Teorema de Frisch-Waugh-Lovell

- Da propriedade anterior, segue o importante resultado abaixo:

Teorema (Frisch-Waugh-Lovell)

Seja $\hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix}$. *Então:*

$$\hat{b}_2 = \left(\boldsymbol{X}_2' M_1 \boldsymbol{X}_2\right)^{-1} \left(\boldsymbol{X}_2' M_1 \boldsymbol{y}\right) = \left((M_1 \boldsymbol{X}_2)'(M_1 \boldsymbol{X}_2)\right)^{-1} \left((M_1 \boldsymbol{X}_2)'(M_1 \boldsymbol{y})\right)$$

- Resultado acima nos mostra que estimadores de MQO associados a $\boldsymbol{X}_2$ em uma regressão que inclui $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ são idênticos a:
  - Regredir $\boldsymbol{y}$ em $\boldsymbol{X}_1$, e guardar os resíduos $\boldsymbol{e}_y$.
  - Para cada $j = 1, \ldots k_2$, regredir a j-ésima coluna de $\boldsymbol{X}_2$ em $\boldsymbol{X}_1$, e guardar os resíduos $\boldsymbol{e}_j$.
  - Regredir $\boldsymbol{e}_y$ em $\boldsymbol{e}_1, \ldots \boldsymbol{e}_{k_2}$ e recuperar os coeficientes.

MQO: Propriedades Estatísticas em Amostras Finitas

Regressores fixos

- Nesta seção, analisaremos as propriedades de risco do estimador de MQO.
- Essas propriedades serão analisadas com respeito à distribuição de $\boldsymbol{y}$ condicionalmente a $\boldsymbol{X}$.
    - Em outras palavras, estamos pensando nas propriedades do estimador sob amostras repetidas (realizações alternativas da incerteza) em que o valor dos regressores é o mesmo da amostra observada.
    - Ou seja, estamos efetivamente tratando os regressores como fixos sob amostras repetidas.
    - **Exemplo:** se Y_i é a taxa de inflação no período i, e X_i a taxa de desemprego no período $i-1$, analisaremos as propriedades dos estimadores sob realizações alternativas do cenário econômico em que a taxa de desemprego é igual à observada nos períodos $i = 1, \ldots, n$.

Não viés

- Sob a restrição de que:

Hipótese (H2-EXOGENEIDADE)

$$\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0$$

- O estimador de MQO é não viciado para β, isto é, qualquer que seja o valor de $\beta \in \mathbb{R}^k$, temos que:

$$\mathbb{E}[\hat{b}|\boldsymbol{X}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \beta$$

- Observe que a condição de exogeneidade implica que:

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\beta$$

Interpretação da condição de exogeneidade

- Sob amostragem aleatória de uma população, i.e. $(X_i, Y_i) \overset{iid}{\sim} (X, Y)$, condição implica $\mathbb{E}[Y_i|\boldsymbol{X}] = \mathbb{E}[Y_i|X_i] = X_i'\beta \implies \mathbb{E}[Y|X] = X'\beta$.
  - A esperança condicional das variáveis (Y, X) que representam a distribuição das características na população de interesse é linear em X, e coincide com o modelo linear de interesse.
  - Se o modelo postulado é preditivo, isso signifca que o parâmetro-alvo do melhor preditor linear consiste também no melhor preditor dentro das funções não lineares de X (ou, de modo equivalente, a melhor aproximação linear a $\mathbb{E}[Y|X]$ é exata).
  - Se modelo postulado é causal, isso significa que as causas não observadas ϵ apresentam o mesmo valor médio nas diferentes subpopulações definidas pelos valores de X. Trata-se de condição mais forte que (i.e. que implica) a condição de identificação $\text{cov}(X, \epsilon) = 0$.
- Quando as observações (X_i, Y_i) apresentam dependência entre si, esta condição impõe restrições adicionais.
  - Por exemplo, se as observações estão ordenadas no tempo e o modelo postulado é causal, $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$ implica que causas não observadas do fenômeno no período i não exibem associação sistemática com as causas em i e em nenhum outro período.
    - Nesse caso, condição limita retroalimentação entre causas observadas e não observadas no tempo.

Viés de variável omitida

- Considere um modelo da forma:

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{\epsilon}\,,$$

em que $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}_1, \boldsymbol{X}_2] = 0$.

  - Por exemplo, num modelo causal, é suficiente observar as causas $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ para que as causas não observadas restantes estejam balanceadas nas subpopulações definidas pelos valores de $(\boldsymbol{X}_1, \boldsymbol{X}_2)$.

- Suponha que você não observe ou desconsidere $\boldsymbol{X}_2$, e considere o estimador de MQO $\tilde{b}_1$ de $\boldsymbol{y}$ em $\boldsymbol{X}_1$.
  - Sob quais condições esse estimador é não viciado para β_1?

- Um cálculo simples nos mostra que:

$$\tilde{b}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} = \beta_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\beta_2 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\epsilon$$
$$\tilde{b}_1 = \beta_1 + \hat{\gamma}\beta_2 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\epsilon}\,,$$

onde $\hat{\gamma}$ é uma matrix $k_1 \times k_2$ em que cada coluna representa os coeficientes do estimador de MQO da j-ésima coluna de $\boldsymbol{X}_2$ em $\boldsymbol{X}_1$.

Viés de variável omitida (cont.)

- Da propriedade da torre, temos que $\mathbb{E}[\epsilon|\boldsymbol{X}_1] = \mathbb{E}[\mathbb{E}[\epsilon|\boldsymbol{X}_1, \boldsymbol{X}_2]|\boldsymbol{X}_1] = 0$. Portanto, temos que: $\mathbb{E}[\tilde{b}_1|\boldsymbol{X}_1] = \beta_1 + \mathbb{E}[\hat{\gamma}|\boldsymbol{X}_1]\beta_2$
- Se uma das duas condições abaixo for satisfeita, estimador será não viciado para β_1:

1. $\beta_2 = 0$.
   - Se o modelo postulado é causal, essa hipótese significa que as variáveis $\boldsymbol{X}_2$ não possuem efeito causal sobre $\boldsymbol{y}$.
   - Se o modelo postulado é preditivo, essa hipótese significa que as variáveis $\boldsymbol{X}_2$ não possuem informação preditiva sobre $\boldsymbol{y}$, uma vez que usamos $\boldsymbol{X}_1$ na predição.
     - Nesse caso, estimador de MQO do melhor preditor linear que inclui $\boldsymbol{X}_2$ produzirá, em média, o mesmo resultado que o estimador que inclui $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$.
2. $\mathbb{E}[\hat{\gamma}|\boldsymbol{X}_1] = 0$.
   - Essa hipótese é satisfeita se as variáveis em $\boldsymbol{X}_1$ não possuem capacidade preditiva sobre nenhuma das variáveis em $\boldsymbol{X}_2$.

Estimador linear

- Para obter propriedades de otimalidade para o estimador de MQO, necessitamos introduzir definições adicionais.
- Um estimador do parâmetro β é dito linear (em $\boldsymbol{y}$) se:

$$\phi(\boldsymbol{X}, \boldsymbol{y}) = A(\boldsymbol{X})\boldsymbol{y}\,,$$

  onde $A : \mathbb{R}^{n \times k} \mapsto \mathbb{R}^{k \times n}$.
    - Note que o estimador de MQO é um estimador linear.

Homocedasticidade

- O estimador de MQO terá propriedades de otimalidade se, além de $H1$-$H2$ restringirmos que:

Hipótese (H3-Homocedasticidade)

Existe $\sigma^2 > 0$ *tal que* $\mathbb{V}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \sigma^2 \mathbb{I}_{n \times n}$.

- A hipótese de homocedasticidade requer que condicionalmente a $\boldsymbol{X}$ os termos de erro ϵ_i, ϵ_j, $j \neq i$, sejam não correlacionados e exibam variância idêntica e não dependente do valor de $\boldsymbol{X}$.
  - Sob amostragem aleatória de uma população (X, Y) em que vale para o modelo postulado que $\mathbb{E}[\epsilon|X] = 0$ (H2), temos que, para $i \neq j$, $\text{cov}(\epsilon_i, \epsilon_j|\boldsymbol{X}) = \mathbb{E}[\epsilon_i \epsilon_j|\boldsymbol{X}] = \mathbb{E}[\epsilon_i \epsilon_j|X_i, X_j] = \mathbb{E}[\mathbb{E}[\epsilon_i|\epsilon_j, X_i, X_j]\epsilon_j|X_i, X_j] = \mathbb{E}[\mathbb{E}[\epsilon_i|X_i]\epsilon_j|X_i, X_j] = 0$. Portanto, o requerimento de não covariância condicional não impõe restrições adicionais a H2.
  - Por outro lado, nesse caso, a hipótese impõe que $\mathbb{V}[Y|X] = \mathbb{V}[\epsilon|X] = \sigma^2$, i.e. que a dispersão de Y nas subpopulações definidas pelos diferentes valores de X sejam as mesmas.

Teorema de Gauss-Markov

- Defina o seguinte conjunto de distribuições condicionais de $\boldsymbol{y}$ dado $\boldsymbol{X}$:

$$\mathcal{F}(\boldsymbol{X}) = \{P_{\boldsymbol{y}|\boldsymbol{X}} : \exists \delta \in \mathbb{R}^k, \mathbb{E}_{P_{\boldsymbol{y}|\boldsymbol{X}}}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\delta, \exists \sigma^2 > 0, \mathbb{V}_{P_{\boldsymbol{y}|\boldsymbol{X}}}[y|\boldsymbol{X}] = \sigma^2 \mathbb{I}_{n\times n}\}$$

- Essa é a classe de distribuições condicionais de $\boldsymbol{y}$ dado $\boldsymbol{X}$ em que a esperança condicional é linear em $\boldsymbol{X}$ e a hipótese de homocedasticidade é satisfeita.
  - Denote, para uma distribuição condicional $F \in \mathcal{F}(\boldsymbol{X})$, $\delta(F)$ o parâmetro que entra na esperança condicional.
- Com base nas definições acima, temos o seguinte resultado.

Teorema (Gauss-Markov)

Sob H1, o estimador de MQO é o estimador linear não viciado para $\delta(F)$ de variância uniformemente mínima na classe $\mathcal{F}(\boldsymbol{X})$, i.e. para qualquer outro estimador linear ψ tal que $\mathbb{E}_F[\psi|\boldsymbol{X}] = \delta(F)$ para todo $F \in \mathcal{F}(\boldsymbol{X})$, temos que, para todo $F \in \mathcal{F}(\boldsymbol{X})$:

$$\mathbb{V}_F(\psi|\boldsymbol{X}) - \mathbb{V}_F(\hat{b}|\boldsymbol{X}) \quad \textit{é positiva semidefinida}$$

Etapas da demonstração

1. Mostrar que $\{\delta(F) : F \in \mathcal{F}(\boldsymbol{X})\} = \mathbb{R}^k$.
   - Basta considerar, para $\delta \in \mathbb{R}^k$, $\boldsymbol{y} = \boldsymbol{X}\delta + \boldsymbol{\epsilon}$, com $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I}_{n\times n})$ e $\boldsymbol{\epsilon}$ independente de $\boldsymbol{X}$.
2. Com base no resultado acima, mostrar que qualquer estimador linear $A(\boldsymbol{X})\boldsymbol{y}$ não viciado em $\mathcal{F}(\boldsymbol{X})$ deve satisfazer

$$A(\boldsymbol{X})\boldsymbol{X} = \mathbb{I}_{k\times k}$$

3. Mostrar que, para todo $F \in \mathcal{F}(\boldsymbol{X})$, $\text{cov}_F(A(\boldsymbol{X})\boldsymbol{y} - \hat{b}, \hat{b}|\boldsymbol{X}) = 0_{k\times k}$
4. Concluir que, para todo $F \in \mathcal{F}(\boldsymbol{X})$:

$$\mathbb{V}_F(A(\boldsymbol{X})\boldsymbol{y} - \hat{b}|\boldsymbol{X}) = \mathbb{V}_F(A(\boldsymbol{X})\boldsymbol{y}|\boldsymbol{X}) - \mathbb{V}_F(\hat{b}|\boldsymbol{X})$$

MQO: Inferência em Amostra Finita

O problema de teste de hipóteses

- Seja (Ω, Σ, P) um experimento estatístico, e $\mathcal{P}$ um *modelo*.
- Sejam $\mathcal{P}_0$ e $\mathcal{P}_1$ uma partição de $\mathcal{P}$.
- O problema de decisão estatística conhecido como teste de hipóteses consiste em afirmar se:

$$P \in \mathcal{P}_0$$

  ou

$$P \in \mathcal{P}_1$$

  - Classe $\mathcal{P}_0$ é conhecida como hipótese nula (H_0), e $\mathcal{P}_1$ é a classe de alternativas ou hipótese alternativa (H_1).
- Um teste não aleatorizado é uma regra de decisão $\phi : \Omega \mapsto \{0, 1\}$ mensurável (i.e. $\phi^{-1}(\{1\}) \in \Sigma$).
  - Se $\omega \in \Omega$ observado é tal que $\phi(\omega) = 1$, afirmamos H_1 (rejeitamos H_0), concluindo por $P \in \mathcal{P}_1$.
  - Por outro lado, se $\phi(\omega) = 0$, afirmamos H_0 (não rejeitamos H_0), concluindo por $P \in \mathcal{P}_0$.
    - Em contraste, em testes aleatorizados, permitimos que a hipótese nula seja rejeitada com uma probabilidade dada por $\phi(\omega) \in [0, 1]$.

Nível de significância, tamanho e poder

- A princípio, a definição de teste não restringe o comportamento do procedimento estatístico adotado.
- Entretanto, gostaríamos de procedimentos estatísticos que limitassem o erro tipo 1 de se rejeitar a hipótese nula, caso ela seja verdadeira.
- Especificamente, para um dado nível de significância $\alpha \in [0, 1]$, gostaríamos de que o teste satisfizesse:

$$\mathbb{E}_F[\phi] \leq \alpha, \quad \forall F \in \mathcal{P}_0$$

  - Probabilidade *ex-ante* de rejeição está limitada a uma probabilidade α.
  - À quantidade $\sup_{F \in \mathcal{P}_0} \mathbb{E}_F[\phi]$ damos o nome de tamanho de um teste.
- Por outro lado, dado um procedimento que controla o nível de significância, gostaríamos de maximizar o poder do teste, i.e., para todo $G \in \mathcal{P}_1$, gostaríamos de tornar:

$$\nu(G) := \mathbb{E}_G[\phi]$$

  o mais alto possível.
  - Maximizar o poder é equivalente a minimizar erro tipo 2 de se afirmar a hipótese nula quando ela é falsa.
  - Usualmente, isso requer que $\sup_{F \in \mathcal{P}_0} \mathbb{E}_F[\phi] = \alpha$.

Inferência no modelo linear

- Retomemos o modelo linear visto anteriormente:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}\,,$$

onde β é desconhecido.

- Para testes de hipótese em amostras finitas, requeremos:

Hipótese (H4-Normalidade)

$$\boldsymbol{\epsilon}|\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{n\times n})$$

.

- Hipótese não só requer que erro tenha média condicional zero e matriz de variância homocedástica, mas impõe normalidade da distribuição condicional.
  - Como parâmetros da normal não dependem de $\boldsymbol{X}$, hipótese implica que $\boldsymbol{\epsilon}$ é independente de $\boldsymbol{X}$.
- Classe de distribuições condicionais compatíveis com a hipótese são:

$$\mathcal{F}^N(\boldsymbol{X}) = \{\boldsymbol{P}_{\boldsymbol{y}|\boldsymbol{X}} : \boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\delta, \sigma^2 \mathbb{I}_{n\times n}), \quad \delta \in \mathbb{R}^k, \sigma^2 > 0\}$$

Estatística F

- Suponha que desejássemos testar:

$$H_0 : R\beta = c$$

contra a alternativa $$H_1 : R\beta \neq c$$

onde R é uma matriz $r \times k$ de posto r, e c é uma vetor $r \times 1$.
  - H_0 e H_1 formam partição de $\mathcal{F}^N(\boldsymbol{X})$
- Considere a seguinte estatística de teste:

$$W_{R,c} = \frac{(R\hat{b} - c)' \left(R\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}R'\right)^{-1}(R\hat{b} - c)}{r} = \frac{(R\hat{b} - c)' \left(R(\boldsymbol{X}'\boldsymbol{X})^{-1}R'\right)^{-1}(R\hat{b} - c)}{r\hat{\sigma}^2},$$

- $\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^n (Y_i - X_i'\hat{b})^2$, com $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ é estimador de $\mathbb{V}[\hat{b}|\boldsymbol{X}]$.
- Estatística mede o desvio de $R\hat{b}$ a c, onde a distância é "ponderada" pelo inverso da variância de $R\hat{b}$.
- Valores grandes da estatística formam evidência contra hipótese nula.

Distribuição da estatística de teste, sob a hipótese nula

Proposição

Suponha válidas as Hipóteses 1 e 4. Se a distribuição $\mathbb{P}$ satisfaz a hipótese nula, então:

$$W_{R,c}|\boldsymbol{X} \sim F(r, n-k)$$

- Etapas para demonstração do resultado:
  1. Mostrar que, sob a nula, $N := (R\hat{b} - c)' \left(R\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}R'\right)^{-1}(R\hat{b} - c)|\boldsymbol{X} \sim \chi^2(r)$.
  2. Mostrar que $(n-k)\hat{\sigma}^2/\sigma^2|\boldsymbol{X} \sim \chi^2_{n-k}$.
  3. Mostrar que N e $\hat{\sigma}^2$ são independentes (condicionalmente a $\boldsymbol{X}$).

Teste de hipótese

- O resultado anterior sugere considerar a seguinte regra de decisão, para um nível de significância $\alpha \in [0, 1]$:

$$\phi := \mathbf{1}\{W_{R,c} > q_F(1 - \alpha | r, n - k)\}$$

onde $q_F(1 - \alpha | r, n - k)$ é o quantil $1 - \alpha$ da distribuição $F(r, n - k)$.
  - Rejeitamos a nula se $W_{R,q}$ for suficientemente alta.
  - Com base no *slide* anterior, se nula for verdadeira, tem-se $\mathbb{E}[\phi | \boldsymbol{X}] = \alpha$.
- Estatística $W_{r,c}$ pode ser reescrita a partir da diferença relativa entre o R^2 do estimador de MQO irrestrito, relativamente ao estimador de MQO que impõe a hipótese nula.
  - Veja o livro do Greene para detalhes.
- No caso em que $r = 1$, teste é equivalente a rejeitar a nula se:

$$|\hat{t}_{R,q}| > q_t(1 - \alpha/2 | n - k)$$

onde $q_t(1 - \alpha/2 | n - k)$ é o quantil $1 - \alpha/2$ de uma t de Student com $n - k$ graus de liberdade, e:

$$\hat{t}_{R,c} = \frac{(R\hat{\beta} - c)}{\sqrt{R\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}R'}}$$

MQO: Propriedades Assintóticas

Regime Assintótico

- Nesta seção, analisaremos o comportamento do estimador de MQO em um regime em que o número de observações diverge.
  - Ideia de se trabalhar com um regime assintótico é de que isso nos seja informativo das propriedades do estimador quando o tamanho amostral é "grande".
- Para isso, focaremos no caso em que observamps uma amostra aleatória de uma população cuja distribuição é representada pelo para de variáveis aleatórias (X, Y).
  - Resultados para o caso não independente e/ou não identicamente distribuído podem ser encontrados, por exemplo, no livro de Halbert White, "Asymptotic Theory for Econometricians".
- Especificamente, trabalharemos com a seguinte hipótese.

Hipótese (Assintótico-1)

$(X_i, Y_i) \overset{iid}{\sim} (X, Y)$, *com* $\mathbb{E}[|Y|^2] < \infty$, $\text{trace}(\mathbb{E}[XX']) < \infty$ *e* $\mathbb{E}[XX']$ *tem posto* k.

Consistência

- Dizemos que um estimador é consistente para um parâmetro quando há convergência em probabilidade dele para o parâmetro.
  - Para qualquer tolerância e qualquer nível de probabilidade, existe um limiar n^* tal que, para tamanhos amostrais $n \geq n^*$, o estimador dista do parâmetro de interesse a menos da tolerância com probabilidade não menor que o nível de probabilidade pre-especificado.
- Propriedades do estimador de MQO em amostras grandes podem ser obtidas usando a expressão conveniente abaixo.

$$\hat{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

- No que segue, indexamos $\hat{b}$ explicitamente por n, denotando-o por $\hat{b}_n$, para explicitar a natureza do limite em consideração

Proposição

Suponha que Assintótico-1 seja satisfeita. Então:

- $\lim_{n\to\infty} \mathbb{P}[\{\hat{b}_n \text{ existe}\}] = 1$.
- $\hat{b}_n \xrightarrow{p} (\mathbb{E}[XX'])^{-1}\,\mathbb{E}[XY]$ *quando* $n \to \infty$.

Consistência para parâmetro de modelo

- Considere um modelo linear na população da forma:

$$Y = X'\beta + \epsilon\,.$$

- Resultado anterior implica que, se $\mathbb{E}[X\epsilon] = 0$, então $\hat{b}_n \xrightarrow{p} \beta$.
- Se $X = \begin{pmatrix} 1 & \tilde{X}' \end{pmatrix}'$, condição é equivalente a $\mathbb{E}[\epsilon] = 0$ e $\text{cov}(\tilde{X}, \epsilon) = 0$.
  - Ambas as hipóteses são satisfeitas por construção num modelo preditivo linear (com intercepto).
- Do resultado acima segue, portanto, que, quando X contém um intercepto, o estimador de MQO é consistente para os parâmetros do melhor preditor linear de Y em $\tilde{X}$ (e um intercepto).

Corolário

Suponha que $X = \begin{pmatrix} 1 & \tilde{X}' \end{pmatrix}'$, e que Assintótico-1 seja satisfeita. Então o estimador de MQO é consistente para os parâmetros do melhor preditor linear de Y em $\tilde{X}$ (e um intercepto). Consequentemente, num modelo causal, estimador de MQO será consistente para efeitos causais se as causas observadas e não observadas não covariarem, pois, nesse caso, parâmetros causais são identificáveis com base no melhor preditor linear.

Distribuição assintótica

- Se Y segue um modelo linear da forma:

$$Y = X'\beta + \epsilon$$

temos que

$$\hat{b}_n - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i\right)$$

- Sob a hipótese de consistência $\mathbb{E}[X\epsilon] = 0$, $\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i \xrightarrow{p} 0$
- Ademais, sob a hipótese de consistência, esperaríamos que $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i\right)$ convergisse em distribuição para uma normal multivariada, com base na aplicação do teorema central do limite.
  - Para isso $X\epsilon$ deve possuir segundo momento finito.

Normalidade assintótica de $\hat{b}_n$

- Elencamos hipóteses suficientes para aplicação de um TCL abaixo.

Hipótese (Assintótico-2)

$$\mathbb{E}[X\epsilon] = 0$$

Hipótese (Assintótico-3)

X e ϵ possuem quartos momentos finitos.

- Com base nas hipóteses 1-3, temos o seguinte resultado.

Proposição

Sob as hipóteses Assintótico 1-3, temos que:

$$\sqrt{n}(\hat{b}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[XX']^{-1}\mathbb{E}[XX'\epsilon^2]\mathbb{E}[XX']^{-1})$$

Variância assintótica

- Resultado anterior sugere que, em amostras grandes, dispersão da distribuição $\sqrt{n}(\hat{b}_n - \beta)$ é "aproximável" por:

$$\text{Avar}(\sqrt{n}(\hat{b}_n - \beta)) := \mathbb{E}[XX']^{-1}\mathbb{E}[XX'\epsilon^2]\mathbb{E}[XX']^{-1}$$

  - Note que, formalmente, **não** estamos dizendo que $\mathbb{V}\left[\sqrt{n}(\hat{b}_n - \beta)\right]$ converge para isso, visto que convergência em distribuição **não** implica convergência de momentos.
  - Não obstante, a variância da distribuição assintótica de $\sqrt{n}(\hat{b}_n - \beta)$, conhecida usualmente como variância assintótica de $\hat{b}_n$, nos será útil para inferência.

Estimador da variância assintótica de White

- Estimador consistente de $\text{Avar}(\sqrt{n}(\hat{b}_n - \beta))$ é dado por:

$$\hat{V}_{\text{HC}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{n} X_i X_i' \hat{e}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1},$$

onde $\hat{e}_i = Y_i - X_i'\hat{b}$ é o resíduo da regressão.

  - Estimador da variância "robusto à heterocedasticidade" de White (*heteroskedasticity-consistent variance estimator*), onde "robustez" advém do fato de que variância assintótica foi derivada sem recorrer à homocedasticidade (i.e. não precisamos supor em nenhum momento que $\mathbb{V}[Y|X] = \sigma^2$).
  - No comando `feols` do pacote `fixest` do `R`, opção `se = "hetero"` recupera este estimador (com uma correção de graus de liberdade que não faz diferença assintoticamente).
    - Se usarmos o comando `lm`, disponível no pacote básico do R, para implementar o estimador de MQO, precisamos usar o comando `vcovHC` do pacote `sandwich` para recuperar esse estimador da variância.

Estatística de Wald e Inferência assintótica

- Suponha que queiramos testar a hipótese nula:

$$H_0 : R\beta = c$$

contra a alternativa

$$H_1 : R\beta \neq c$$

onde R é uma matriz $r \times k$ com posto r.

- Nosso resultado de inferência assintótica sugere considerar a estatística de teste de Wald, dada por:

$$\tilde{W}_{R,c} = n(R\hat{b}_n - c)' \left(R\hat{V}_{\mathrm{HC}}R'\right)^{-1} (R\hat{b}_n - c),$$

Tamanho e poder assintóticos do teste

- Sob hipótese nula, não é difícil ver que:

$$\tilde{W}_{R,c} \xrightarrow{d} \chi^2_r$$

  - Desse modo, tomando $q_\chi(1-\alpha|r)$ o quantil $1-\alpha$ de uma chi-quadrado com r graus de liberdade, temos que, sob a nula:

$$\lim_{n\to\infty} \mathbb{P}\left[\tilde{W}_{R,c} > q_\chi(1-\alpha|r)\right] = \alpha\,.$$

- Por outro lado, não é difícil ver que, se nula não é verdadeira, a estatística $\tilde{W}_{R,c} \xrightarrow{p} \infty$.
  - Nesse caso, podemos mostrar que:

$$\lim_{n\to\infty} \mathbb{P}\left[\tilde{W}_{R,c} > q_\chi(1-\alpha|r)\right] = 1\,.$$

- No caso em que $r=1$ teste de Wald é equivalente a rejeitar a nula quando:

$$|\hat{t}_W| > q_Z(1-\alpha/2)\,,$$

onde q_Z é quantil da normal padrão e $\hat{t}_W = (R\hat{b}_n - c)/\sqrt{R\hat{V}_W R'/n}$.

Padrão de dependência *clustered*

- A derivação da variância assintótica do estimador presumiu que as observações eram independentes.
- No entanto, em diversos contextos, o padrão de amostragem ou de atribuição das causas induz dependência entre observações.
  - Por exemplo, se amostramos um conjunto de bairros, e para estes observamos (Y_i, X_i) de todos os moradores. Nesse caso, esperamos, sob realizações repetidas da amostragem, que haja correlação entre os X_i e Y_i de um mesmo bairro (por exemplo, se X_i denota a renda de i).
  - Por outro lado, se a incerteza amostral é de natureza econômica, podemos esperar que ϵ_i contenha fatores econômicos não observados comuns a um grupo, ou que X_i seja atribuído pela natureza de forma não independente, de modo que haja correlação sob realizações repetidas da incerteza.
- Nesses casos, pode ser razoável particionar a amostra em G clusters, $\mathcal{N}_1 \cup \mathcal{N}_2 \ldots \cup \mathcal{N}_G = \{1, \ldots, n\}$, com a propriedade de que observações no mesmo *cluster* apresentam dependência (arbitrária) entre si, mas observações entre clusters são aproximadamente independentes.
  - No exemplo de amostragem de bairros, hipótese pode ser razoável se amostramos de forma uniforme de uma população grande de bairros.

Estimador da variância robusto a *clusters*

- Sob o padrão de dependência de *clusters*, a variância do termo crucial para a distribuição assintótica do estimador de MQO fica na forma:

$$\mathbb{V}\left[\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i\right)\right] = \sum_{g=1}^{G} \frac{|\mathcal{N}_g|^2}{n} \mathbb{V}\left[\frac{1}{|\mathcal{N}_g|}\sum_{i \in \mathcal{N}_g} X_i \epsilon_i\right]$$

- O estimador da variância robusta a *clusters* (*cluster-robust variance estimator*) propõe estimar a variância assintótica do estimador de MQO como:

$$\hat{V}_{\text{CR}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{g=1}^{G}\frac{1}{n}\left(\sum_{i,j \in \mathcal{N}_g} X_i X_j' \hat{\epsilon}_i \hat{\epsilon}_j\right)\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}$$

- Sob restrições dos momentos de (X_i, Y_i), estimador produzirá inferência assintótica válida, num regime em que $G \to \infty$ e nenhum *cluster* "domina" os demais ($\max_g |\mathcal{N}_g|^2/n \to 0$).
  - Portanto, para utilizar esse método, necessitamos de bastantes *clusters* de tamanho comparável.
  - No pacote `feols`, opção `cluster` permite utilizar esse estimador. Para regressão obitda via `lm`, usar `vcovCL` do pacote `sandiwch`.