

Inferência Causal e Dados em Painel - Aula 1 -

Prof. Mestre. Omar Barroso Khodr

Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

Introdução

O que é inferência Causal?

- A **inferência causal (IC)** envolve o uso de teoria(s) e conhecimento(s) institucional(is) para estimar o impacto de **eventos** e **escolhas** nos resultados de interesse.
- *Por exemplo, qual é o impacto dos ciclos de negócios sobre a taxa de câmbio.*
- *Qual é o impacto do aumento do salário mínimo sobre as taxas de pobreza de uma região específica?*
- A IC não é apenas um método ferramental para testar o impacto de fenômenos, mas também é uma maneira de dedução de ‘mistérios’ quando analisamos dados!

É elementar meu caro Watson...

- Nos livros de Sherlock Holmes (de Sr. Arthur Conan Doyle), o famoso investigador busca diversas pistas para bolar diferentes cenários explicativos sobre um mistério.
- Conforme a sua coleta de dados, Holmes busca em usar os fragmentos encontrados para chegar a premissas lógicas e explicar eventos complexos.
- Os métodos de IC se baseiam nas mesmas premissas usadas por grandes investigadores!



Fonte: Alamy

Correlação \neq Causalidade

- Devemos investigar exatamente por que as correlações, particularmente em dados observacionais, provavelmente **não** refletem uma relação causal.
- Por exemplo, quando o galo canta, o sol nasce logo após, mas sabemos que o galo não fez o sol nascer!
- Se o galo tivesse sido caçado pelo gato do fazendeiro, o sol ainda teria nascido.
- No entanto, muitas vezes cometemos esse tipo de erro ao interpretar ingenuamente correlações simples...

Correlação \neq Causalidade

- É possível que duas coisas tenham uma relação causal mesmo que **nenhuma correlação** seja observável entre elas.
- O exemplo entre o Marinheiro e o vento...
- Um marinheiro contraria o vento movendo o leme, mantendo o barco em um caminho reto.
- Um observador pode concluir incorretamente que o leme está quebrado devido à ausência de relação visível entre o movimento do leme e a direção do barco.
- O movimento aleatório do leme (por exemplo, usando cara ou coroa) revelaria a relação causal.



Fonte: Cunningham (Causal Inference: The Mixtape, 2021)

Correlação \neq Causalidade

- Os bancos centrais agem para neutralizar recessões por meio de operações de mercado aberto.
- Se as intervenções forem ótimas, nenhuma relação visível com a produção será observada.
- Mesmo intervenções agressivas para evitar recessões podem não mostrar evidências de eficácia, apesar de serem eficazes [segundo a teoria ou testes empíricos prévios].
- O **comportamento humano** frequentemente mascara relações causais devido a ações não aleatórias. (e.g., contratos de derivativos arriscados por agentes econômicos durante a crise de 2008).
- A **aleatoriedade** é crucial para identificar efeitos causais

A Otimização torna tudo Endógeno

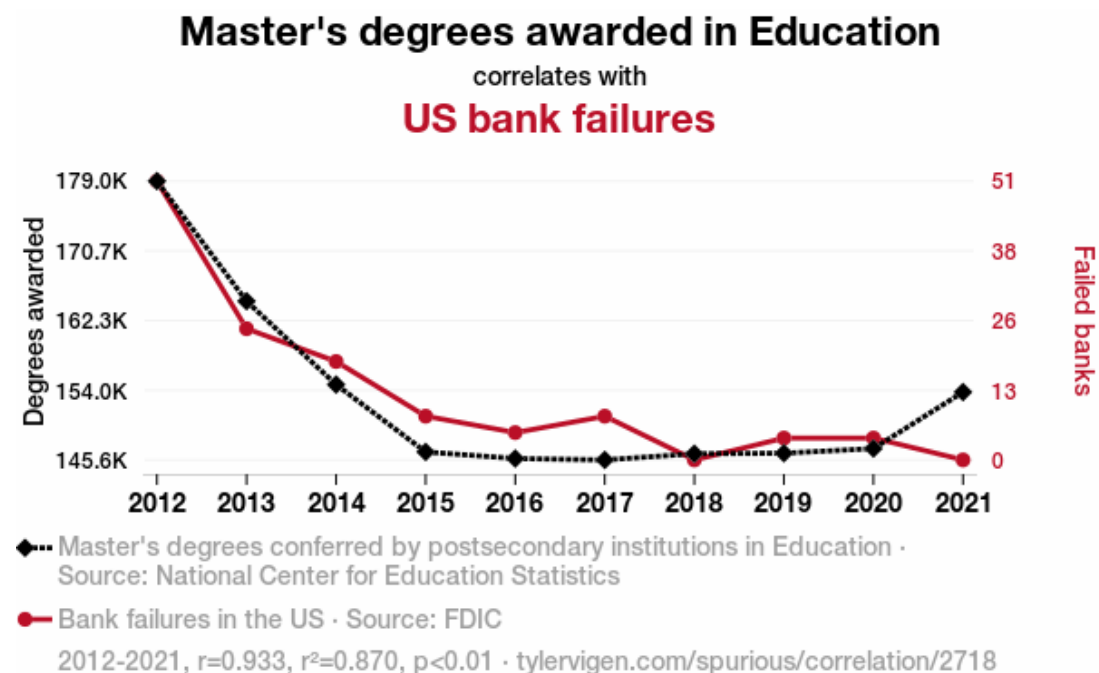
- **Metodologias** de inferência causal geralmente dependem de teoria e conhecimento institucional local.
- Conhecimento prévio é essencial para justificar alegações de descobertas causais. (e.g., como questões de estrutura, cultura, espaço e métodos científicos).
- Tipos de dados são amplamente categorizados em dados **experimentais** e **não experimentais (observacionais)**.
- Dados **experimentais** são coletados em ambientes controlados com participação ativa do pesquisador. (e.g., pesquisas de opiniões e coleta de dados).

A Otimização torna tudo Endógeno

- Dados **observacionais** são coletados passivamente, frequentemente por meio de pesquisas ou como subprodutos de outras atividades. (e.g., APIs ou Meta dados).
- Dados experimentais são mais **desafiadores** de obter em ciências sociais, mas estão se tornando mais comuns.
- Dados observacionais continuam sendo o tipo **mais comum** usado por pesquisadores.

A Otimização torna tudo Endógeno

- Correlações em dados observacionais frequentemente não refletem relações causais.
- Variáveis são escolhidas endogenamente por indivíduos que tomam decisões sob restrições, criando correlações espúrias.



Diplomas de Mestrados concedidos correlacionam com a quantidade de falências bancárias (EUA). Fonte: Spurious Correlations (2022)

A Otimização torna tudo Endógeno

- Uma correlação deve ser baseada em escolhas independentes para medir um efeito causal.(e.g., alguma variável exógena que seja correlacionada com um regressor, veremos mais sobre isso quando estudarmos as *variáveis instrumentais*).
- A teoria econômica sugere que as escolhas são endógenas.
- Correlações entre escolhas endógenas e resultados raramente representam efeitos causais.

A Otimização torna tudo Endógeno

- Identificar efeitos causais envolve suposições e uma crença específica sobre o trabalho científico.
- Pesquisas confiáveis priorizam a metodologia correta em vez de atingir os resultados desejados (evitando viés de confirmação).
- O verdadeiro trabalho científico é orientado para o **processo**, não para o resultado.
- Metodologias científicas permitem **respostas inesperadas e indesejáveis**.
- **Propaganda não é ciência**; métodos científicos devem formar crenças com base em evidências e metodologia rigorosa. [e.g., ao oposto de pesquisas tendenciosas que buscam amostras que potencialmente favorecem uma opinião].

Fundamentos

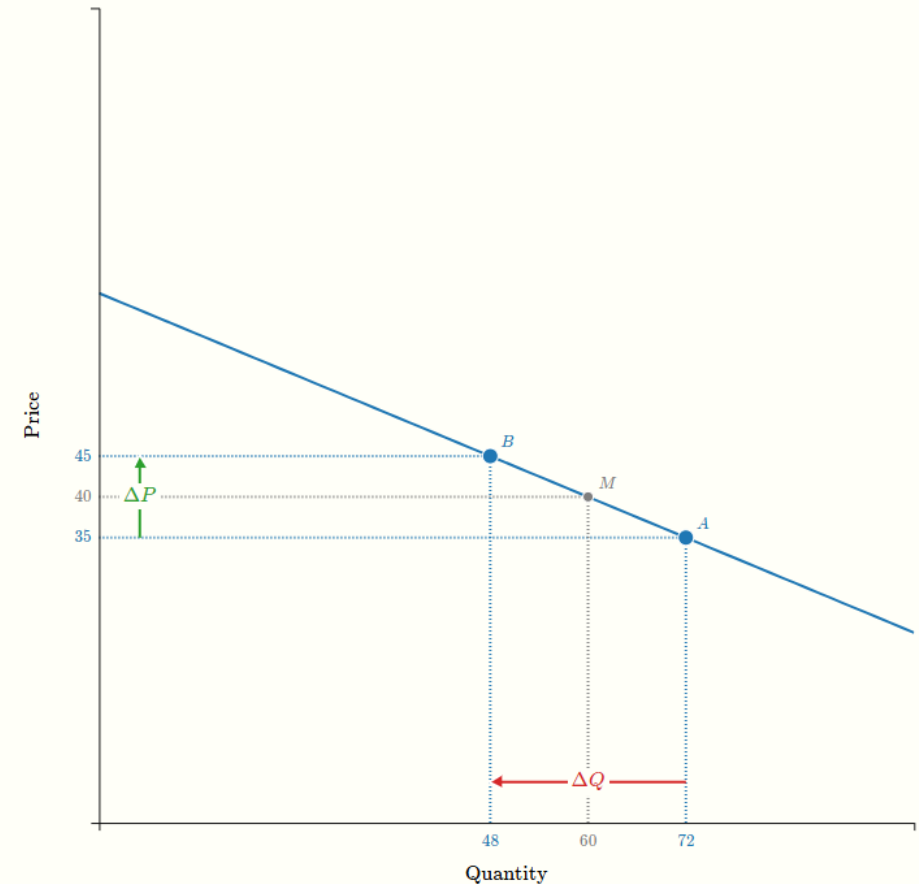
- **Análise Empírica:** Usando dados para testar teorias ou estimar relacionamentos entre variáveis. (e.g., correlação e covariância)
- **Formulação da Pergunta:** O passo inicial na análise econômica empírica; envolve *formulação cuidadosa da pergunta*.
- Por exemplo, quais são os efeitos de más safras sazonais [de trigo] na inflação dos alimentos? Ou, Qual é a função de programas de transferência de dinheiro para os níveis de desigualdade econômica em tempos de recessão?
- **Modelos Econômicos Formais:** Descrevem relacionamentos, comportamentos ou processos de forma matemática.

Fundamentos

- **Estatística Comparativa:** Uma descrição teórica dos efeitos causais dentro de um modelo, com base na suposição *ceteris paribus* (tudo o mais constante).
- **Inferência Causal:** Entender o efeito causal de uma **intervenção** pressupõe que outras variáveis relevantes no modelo não mudam.
- **Modelo de Oferta e Demanda:** Usado para ilustrar os desafios na estimativa da elasticidade-preço da demanda.
- **Interesse na Elasticidade:** Os formuladores de políticas e gerentes de negócios estão interessados na *elasticidade-preço* da demanda para maximizar os lucros e otimizar os impostos.

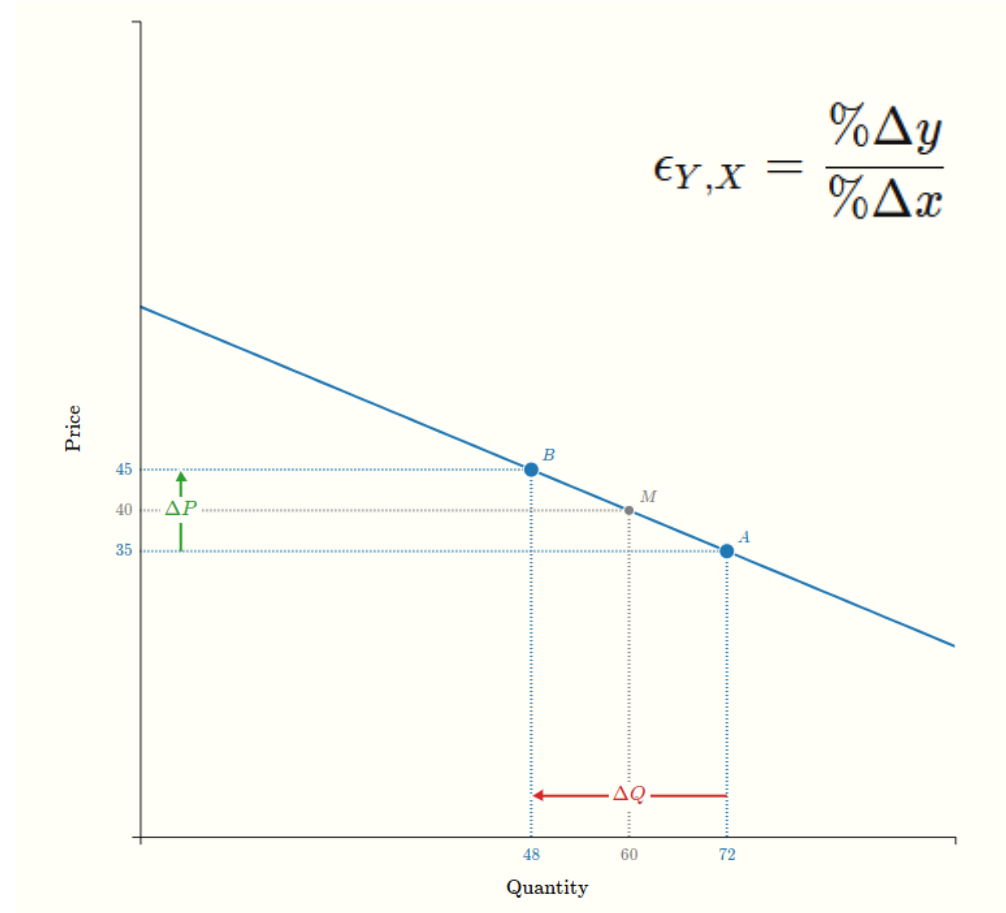
Fundamentos

- **Natureza Teórica das Curvas de Demanda:** As curvas de demanda são construções teóricas que representam resultados potenciais de preço e quantidade. Os valores observados são preço e quantidade de equilíbrio, não a curva de demanda inteira.
- **Cálculo da Elasticidade:** Requer plotar resultados potenciais ao longo de uma curva de demanda.



Fundamentos

- A elasticidade preço da demanda é a razão entre as mudanças percentuais na quantidade e no preço para uma única curva de demanda.
- No entanto, quando há mudanças na oferta e na demanda, uma sequência de pares de quantidade e preço emerge na história que não reflete nem a curva de demanda nem a curva de oferta.
- Na verdade, conectar os pontos não reflete nenhum objeto significativo ou útil.



Fonte: EconGraphs

Fundamentos

- Mas neste exemplo, a mudança em 'P' [preço] é exógena.
- Por exemplo, ela mantém a oferta fixa, os preços de outros bens fixos, a renda fixa, as preferências fixas, os custos de insumos fixos e assim por diante. Para estimar a elasticidade-preço da demanda, precisamos de mudanças em 'p' que sejam completa e totalmente independentes dos determinantes normais da oferta e dos outros determinantes da demanda.
- Caso contrário, obtemos mudanças na oferta ou na demanda, o que cria novos pares de dados para os quais qualquer correlação entre 'P' e 'Q' não será uma medida da elasticidade da demanda.

Fundamentos

- O problema é que a elasticidade é um objeto importante, e precisamos conhecê-la, e portanto precisamos resolver esse problema. Então, dado esse objeto teórico, precisamos escrever um modelo econométrico como ponto de partida. Um possível exemplo de um modelo econométrico seria uma função de demanda linear:

$$\log Q_d = \alpha + \delta \log P + \gamma X + u$$

Fundamentos

- No qual,
- α : É o intercepto
- δ : É a elasticidade da demanda.
- X : é a matriz de fatores que determina a demanda (como preço de outros bens e restrição orçamentária).
- γ : O coeficiente de relação entre X e Q_d (Quantidade Demandada).
- u : O termo de erro.

$$\log Q_d = \alpha + \delta \log P + \gamma X + u$$

Causalidade e Ceteris Paribus Na Econometria

Causalidade e Ceteris Paribus Na Econometria

- Associações por si só raramente são suficientes para a validação de políticas ou teorias.
- Lembrando *Ceteris Paribus*...
- "*outros fatores relevantes sendo iguais*".
- Essencial para isolar o efeito de uma variável (por exemplo, preço sob demanda ou outros fatores). *Teorema Frisch-Waugh-Lovell (FWL)*
- Sem manter outros fatores fixos, a causalidade não pode ser estabelecida.
- **Desafios na Análise Causal**
- Complexidade do mundo real: Muitos fatores influenciam os resultados (por exemplo, salários, criminalidade e tendências sociais).
- Impossível controlar todas as variáveis literalmente; a econometria simula condições *ceteris paribus*.

Teorema Frisch-Waugh-Lovell (FWL)

- Suponha que estamos utilizando a seguinte regressão.
- $y = D'\beta_1 + W'\beta_2 + \epsilon$
- Y = Variável dependente (e.g., Salário)
- D = Grupo de variáveis (e.g., gênero)
- W = Variáveis de controle (e.g., educação e experiência de trabalho).
- β_1 = O efeito de D em Y enquanto W é mantida constante.

Teorema Frisch-Waugh-Lovell (FWL)

- Nesse contexto, queremos saber como gênero (D), educação e experiência (W) influenciam os salários.
- Neste caso, removemos primeiro os efeitos de W.
- $y = \alpha + W \rightarrow \tilde{y} \rightarrow \tilde{y} = D' \widehat{\beta}_1 + \epsilon$
- $D = \alpha + W \rightarrow \tilde{D} \rightarrow \tilde{D} = \widehat{\beta}_1$
- Assim, obtemos os resíduos de ambos os modelos.
- Desta maneira, obtemos $\widehat{\beta}_1$ da regressão amostral utilizando os resíduos estimados \tilde{y} e \tilde{D} .

Causalidade e Ceteris Paribus Na Econometria

- Estudos de empíricos e de políticas
- Exemplo: O efeito do treinamento profissional sobre os salários depende do controle da educação/experiência.
- Pergunta-chave: Há fatores fixos suficientes para justificar a causalidade? Veremos na próxima aula...
- Dados Experimentais vs. Observacionais
- Cenário ideal: Experimentos randomizados (por exemplo, testes de fertilizantes em parcelas).
- Limitação prática: Dados do mundo real frequentemente carecem de controle experimental.
- A econometria ajuda a aproximar experimentos quando a randomização não é viável.

Causalidade e Ceteris Paribus Na Econometria

Exemplo: Fertilizantes e Rendimento das Culturas

- Estudos iniciais (por exemplo, *Griliches 1957*) testaram o efeito dos fertilizantes ceteris paribus.
- Desafios: A qualidade da terra, a precipitação pluvial, etc., variam, mas a atribuição aleatória ajuda a isolar o impacto dos fertilizantes.

Exercícios e Desafios para Causalidade

- **Exemplo:** Medindo o Retorno à Educação
- **Objetivo:** Estimar como um ano extra de educação afeta os salários (ceteris paribus).
- **Experimento Ideal:** Atribuir aleatoriamente níveis de educação e medir os salários.
- **Problema:** Antiético e impraticável — as pessoas selecionam a educação por conta própria nas pesquisas [coletas de dados].
- **Problemas com Dados do Mundo Real:**
- **Variáveis omitidas:** Resultados enviesados por experiência (correlação negativa com a educação) e habilidade inata (correlação positiva).
- **Desafio Principal:** Considerar fatores não observáveis (por exemplo, habilidade) é mais difícil do que fatores observados (por exemplo, experiência).

Exercícios e Desafios para Causalidade

- **2. Exemplo:** Presença Policial e Taxas de Criminalidade
- **Questão Causal:** Mais policiais reduzem a criminalidade?
- **Ideal Experimental:** Atribuir aleatoriamente o tamanho das forças policiais às cidades.
- **Realidade:** As cidades escolhem endogenamente os níveis de policiamento com base nas tendências de criminalidade → causalidade reversa.
- **Solução Econométrica:** Abordar a determinação simultânea (por exemplo, crime → contratação de policiais).

Exercícios e Desafios para Causalidade

- **3. Exemplo:** Salário Mínimo e Desemprego
- **Teoria:** Salários mínimos mais altos podem reduzir o emprego (oferta/demanda).
- **Desafio dos Dados:** As políticas de salário mínimo não são definidas aleatoriamente, são determinadas por fatores políticos/econômicos e podem confundir os resultados.
- **Complexidade de Séries Temporais:** Deve controlar as tendências macroeconômicas (por exemplo, períodos inflacionários e recessões).

Exercícios e Desafios para Causalidade

- **4. Exemplo:** Hipótese das *Expectativas* (Teoria Não Causal)
- **Teoria:** Os retornos esperados sobre os investimentos se igualam, considerando todas as informações disponíveis.
- **Teste:** Comparar os retornos de títulos do Tesouro de 3 meses com os de 6 meses vendidos antecipadamente.
- **Insight Principal:** Mesmo teorias não causais podem ser testadas com econometria. Por exemplo, eficiência dos mercados (Fama, 1965).

Estruturas de Dados Econômicos

- *Dados Transversais (Cross-Section)*
- Dados coletados em um único ponto no tempo (ou em um curto período) em diferentes unidades (por exemplo, indivíduos, empresas e países).
- Concentra-se na **variação entre unidades**, e não ao longo do tempo.
- **Análise instantânea:** Sem dimensão temporal (ao contrário de séries temporais ou dados em painel).
- **Exemplos de unidades:** Domicílios, cidades e empresas.
- Comum em pesquisas: Por exemplo, dados censitários, pesquisas sobre gastos de consumidores.

Estruturas de Dados Econômicos

- *Dados Transversais (Cross-Section).*
- **Vantagens:**
- Fácil de coletar e analisar.
- Útil para **análises estáticas** comparativas (por exemplo, diferenças regionais).
- **Limitações:**
- Não é possível rastrear mudanças ao longo do tempo (por exemplo, crescimento salarial).
- Risco de viés de variável omitida (por exemplo, capacidade não observada afetando salários).

TABLE 1.2 A Data Set on Economic Growth Rates and Country Characteristics				
obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
.
.
.
61	Zimbabwe	2.30	17	6

Fonte: Wooldridge (2013)

Estruturas de Dados Econômicos

- *Séries Temporais (Time-Series)*
- Dados coletados em **múltiplos períodos de tempo** para uma única unidade (por exemplo, país, empresa e indivíduo(s)).
- Concentra-se em tendências, padrões e mudanças ao longo do tempo.
- **Ordenação temporal:** Observações registradas em intervalos regulares (diário, mensal, anual).
- **Exemplos:** crescimento do PIB, preços de ações, taxas de desemprego, inflação.
- **Usado para previsões:** por exemplo, para prever condições econômicas futuras.

Estruturas de Dados Econômicos

- *Séries Temporais (Time-Series)*
- **Vantagens:**
- Captura comportamento dinâmico (por exemplo, ciclos de negócios e sazonalidade).
- Permite análise de efeitos defasados (por exemplo, impactos de políticas com atrasos).
- **Limitações:**
- Não estacionariedade: Tendências/volatilidade podem distorcer a análise (por exemplo e choques inflacionários).
- **Autocorrelação:** Valores passados podem influenciar valores futuros (por exemplo, recessões).

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics						
obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.
.
.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Fonte: Wooldridge (2013)

Estrutura de Dados Econômicos

- *Seção transversal combinada (Pooled Cross-Section)*
- Combina múltiplos conjuntos de dados transversais independentes coletados em diferentes períodos.
- Unidades (por exemplo, indivíduos, empresas) variam entre as amostras (sem acompanhamento ao longo do tempo).
- Principais Características
- Snapshots repetidos: por exemplo, pesquisas anuais com diferentes domicílios (2000, 2005, 2010).
- Sem acompanhamento: Unidades diferentes em cada período (vs. dados em painel).
- Equilibra a variação transversal e temporal: Mais observações do que uma única seção transversal.

Estrutura de Dados Econômicos

- *Seção transversal combinada (Pooled Cross-Section).*
- Vantagens
- Captura tendências amplas (por exemplo, impactos de políticas públicas, mudanças sociais).
- Mais eficiente do que seções transversais únicas (amostra maior).
- Limitações
- Não é possível analisar mudanças em nível individual (amostras diferentes em cada período).
- Efeitos temporais podem confundir os resultados (por exemplo, choques macroeconômicos).

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices						
obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.
.
.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.
.
.
520	1995	57200	16	1100	2	1.5

Fonte: Wooldridge (2013)

Estrutura de Dados Econômicos

- *(Dados em Painel)*
- Acompanha as mesmas unidades (indivíduos, empresas e países) em vários períodos.
- Combina dimensões transversais e de séries temporais.
- **Observações repetidas:** Mesmas entidades pesquisadas em intervalos regulares (por exemplo, anualmente).
- **Duas dimensões:** Transversal (unidades) + Tempo (períodos).
- **Painel balanceado:** Sem dados ausentes para nenhuma unidade.
- **Painel desbalanceado:** Observações ausentes para algumas unidades.

Estrutura de Dados Econômicos

- *(Dados em Painel)*
- **Vantagens**
- Controla a heterogeneidade não observada (por exemplo, capacidade inata).
- Analisa o comportamento dinâmico (por exemplo, efeitos defasados, adaptação).
- **Limitações**
- Custo e complexidade para coletar/manter.
- **Perda:** Os participantes podem desistir com o tempo.

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics						
obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.
.
.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Fonte: Wooldridge (2013)

Bibliografia

- Angrist, J.D. and Pischke, J.-S., 2009. Mostly Harmless Econometrics. 1st ed. Princeton University Press.
- Cunningham, S., 2021. Causal Inference: The Mixtape. Yale University Press. <https://doi.org/10.2307/j.ctv1c29t27>.
- Wooldridge, J.M. (2013) Introductory econometrics: a modern approach. 5th ed. Michigan State University.