

ACH3657

Métodos Quantitativos para Avaliação de Políticas Públicas

Aula teórica 01
Análise de Regressão

Alexandre Ribeiro Leichsenring
alexandre.leichsenring@usp.br



Introdução

- Imagine que você seja contratado pelo governo do Estado para avaliar a eficácia de um programa de capacitação profissional financiado com recursos públicos.
- O programa é voltado para qualquer trabalhador desempregado, e a adesão é voluntária.
- Você deve determinar qual efeito, se houve, do programa de treinamento na inserção no mercado formal de trabalho.

Outros exemplos:

- Efeitos de gastos de campanhas políticas sobre os resultados de eleições
- Efeito de gastos públicos com escolas sobre o desempenho de estudantes

Introdução

- Popularização do uso de modelos matemáticos nas ciências sociais nas últimas décadas
- Disponibilidade crescente de dados quantitativos sobre fenômenos sociais gerados por pesquisas quantitativas (Censo, Pnad, etc) em larga escala e por registros administrativos de políticas públicas
- Embora modelos para dados sociais variem bastante em complexidade e sofisticação, muitos podem ser considerados variantes de técnicas conhecidas como *regressão linear*.
- O objetivo do curso é uma apresentação dos modelos de regressão passando por alguns tipos de variáveis resposta

Introdução

- Modelos estatísticos são relevantes em diversos ramos das ciências sociais e econômicas aplicadas
- Aparecem frequentemente para testar uma determinada teoria econômica ou para revelar a relação entre variáveis com alguma importância na análise de políticas públicas ou na tomada de decisão
- Uma análise empírica usa dados para testar a teoria ou para estimar a relação entre variáveis

Modelos estatísticos e econométricos

- A econometria evoluiu como uma disciplina separada da estatística matemática porque enfoca problemas inerentes à coleta e análise de dados não-experimentais.
- **Dados não experimentais** não são coletados por meio de experimentos controlados (são também chamados de dados observacionais)
- **Dados experimentais** são frequentemente coletados em ambientes de laboratórios nas ciências naturais, mas são muito difíceis de se obter nas ciências sociais.

- Na ciências sociais, um modelo é dado por uma ou mais equações descrevendo o processo gerador da variável resposta
- Frequentemente busca-se compreender a relação causal entre um conjunto de *variáveis* (*variáveis explicativas*) e a variável resposta
- Como muitos *modelos* de fenômenos do mundo real, tais modelos são apenas aproximações da realidade e deve-se ter cautela no seu uso

“Todos os modelos estão errados. Alguns são úteis”
(atribuído a George Box)

- Não obstante, esses modelos fornecem uma ideia abrangente das dinâmicas por trás dos fenômenos sociais e podem ser úteis na compreensão e descoberta de fatores relacionados

Passos na análise econométrica empírica

- O primeiro passo na análise empírica é a formulação cuidadosa da questão de interesse
- A questão pode lidar com testar um determinado aspecto da teoria econômica, ou com testar os efeitos de uma determinada política pública
- Métodos estatísticos podem ser usados para responder um grande conjunto de questões

Exemplo

- Considere o seguinte problema: um gestor quer avaliar os efeitos de um curso de educação profissional na produtividade do trabalhador
- Uma conjectura básica é que fatores como educação, experiência e treinamento afetam a produtividade do trabalhador
- Além disso, sabemos que a remuneração do trabalho está relacionada à produtividade do trabalhador
- Esse simples raciocínio leva para um modelo como:

$$\text{salário} = f(\text{educação, experiência, treinamento}) \quad (1)$$

- Obviamente outros fatores afetam a remuneração mas a equação acima captura a essência do problema

Outro exemplo

- O prêmio Nobel Gary Becker postulou quadro conceitual para descrever a participação de um indivíduo no crime
- Alguns crimes claramente apresentam recompensa econômica
- A maioria dos crimes, entretanto, tem custos associados
- Os custos de oportunidade do crime impedem a participação do criminoso em outras atividades (como o emprego formal)
- Além disso, há custos associados à possibilidade de ser pego e, se condenado, custos associados ao encarceramento
- Da perspectiva de Becker, a decisão de adotar uma atividade ilegal está relacionada à alocação de recursos, em que são levados em consideração benefícios e custos relacionados à atividade

- Sob suposições gerais, podemos derivar uma equação descrevendo a quantidade de tempo empenhada em atividade criminosa como uma função de vários fatores

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \quad (2)$$

onde

- y = horas empenhadas em atividades criminosas
- x_1 = “remuneração” por uma hora empenhada no crime
- x_2 = remuneração num emprego legal
- x_3 = renda de outras fontes
- x_4 = probabilidade de ser pego
- x_5 = probabilidade de ser condenado em caso de prisão
- x_6 = sentença esperada em caso de condenação
- x_7 = idade

- Outros fatores geralmente afetam a decisão de participar no crime, mas a lista acima é ilustrativa de uma análise formal
- Podemos usar a teoria para estimar o efeito que cada uma dessas variáveis teria na atividade criminosa.
- Essa é a base para a análise estatística da atividade criminal

Passos na análise econométrica (continuação)

- A modelagem teórica (social ou econômica) é o ponto de partida da análise **empírica** (é comum usar a teoria de modo pouco formal ou contar com a intuição)
- Você pode concordar quanto aos determinantes da produtividade que aparecem na equação (1) serem razoavelmente baseados no senso comum

$$\text{salário} = f(\text{educação, experiência, treinamento})$$

- Precisamos então voltar à especificação do modelo *econométrico*
- A forma da equação $f(\cdot)$ deve ser especificada antes de podermos empreender uma análise econométrica
- Como lidar com variáveis que não podem ser adequadamente observadas?
E.g.: rendimento obtido com uma atividade criminosa.
- Mesmo variáveis como a probabilidade de ser preso não podem ser obtidas de um modo realista (mas nesse caso, podemos observar estatísticas de detenção e derivar uma variável que se aproxime da probabilidade de prisão)
- Muitos outros fatores que não podemos observar interferem na questão, mas devem ser considerados



- As ambiguidades são resolvidas ao se especificar um modelo econométrico em particular, tal como:

$$\begin{aligned} \textit{crime} = & \beta_0 + \beta_1 \textit{salario}_m + \beta_2 \textit{outrenda} + \beta_3 \textit{freqpris} + \beta_4 \textit{freqcond} + \\ & + \beta_5 \textit{sentmed} + \beta_6 \textit{idade} + u \end{aligned}$$

crime = algum indicador sobre a frequência da atividade criminal

salario = o salário que se pode ganhar em emprego legal

outrenda = renda de outras fontes (ativos, herança, etc)

freqpris = frequência de prisões por infrações anteriores

freqcond = frequência de condenação

sentmed = duração média da sentença após a condenação

- A escolha dessas variáveis é determinada pela teoria social/econômica e por considerações sobre os dados
- O termo u contém fatores não observados como “rendimento” da atividade criminosa, o caráter moral, a formação da família e erros de mensuração
- Podemos acrescentar outras variáveis mas nunca poderemos eliminar u inteiramente
- u é chamado *termo de erro*, e é, talvez, o componente mais importante na análise
- $\beta_0, \beta_1, \dots, \beta_6$ são os *parâmetros do modelo*: descrevem as direções e as intensidades das influências da relação entre *crime* e as variáveis usadas para explicar *crime* no modelo



- Um modelo econométrico para o exemplo da capacitação profissional poderia ser:

$$salarario = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 treina + u$$

em que o termo u contém fatores tais como “aptidão inata”, qualidade da educação, formação da família e uma miríade de outros fatores que podem influenciar o salário de uma pessoas

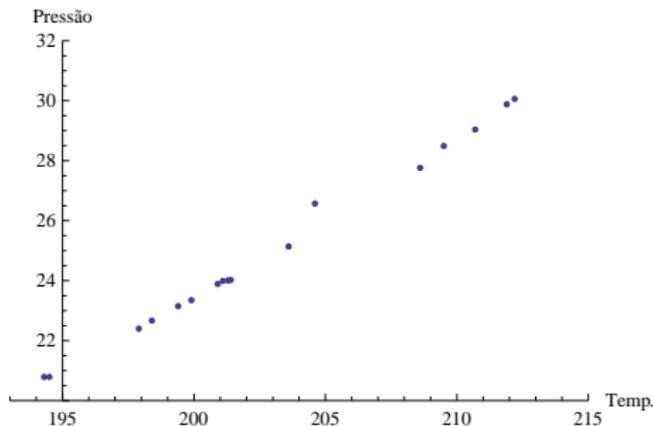
- Se estivermos especialmente interessados nos efeitos do treinamento (capacitação), então β_3 é o parâmetro de interesse
- Uma vez que um modelo econométrico tenha sido especificado, várias hipóteses de interesse pode ser formuladas em termos dos parâmetros desconhecidos
- Por exemplo, no exemplo do crime, poderíamos levantar a hipótese de que $salario_m$, o salário que poderia ser ganho no emprego legal, não tem efeito no comportamento criminoso. No contexto do modelo, a hipótese é equivalente a

$$\beta_1 = 0$$

Exemplo nas ciências naturais

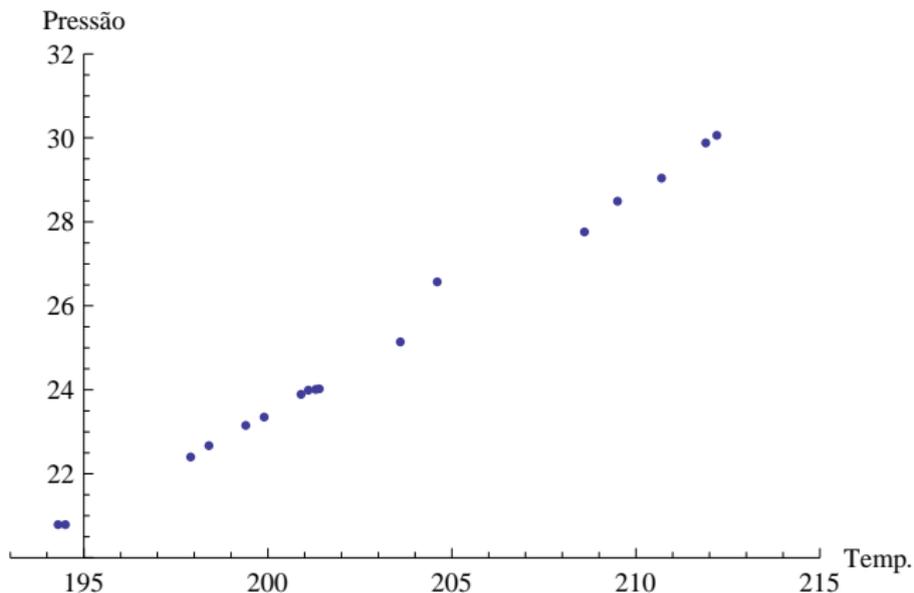
Em um artigo em 1857, o físico escocês James Forbes discutia uma série de experimentos que ele havia feito sobre a relação entre pressão atmosférica e o ponto de ebulição da água. Ele sabia que a altitude podia ser determinada a partir da pressão atmosférica, medida com barômetro, com pressões baixas correspondendo a altas altitudes.

Na metade do século dezenove os barômetros eram instrumentos frágeis, e ele conjecturava que uma medida da temperatura de ebulição da água (mais simples) poderia substituir medições por barômetro.

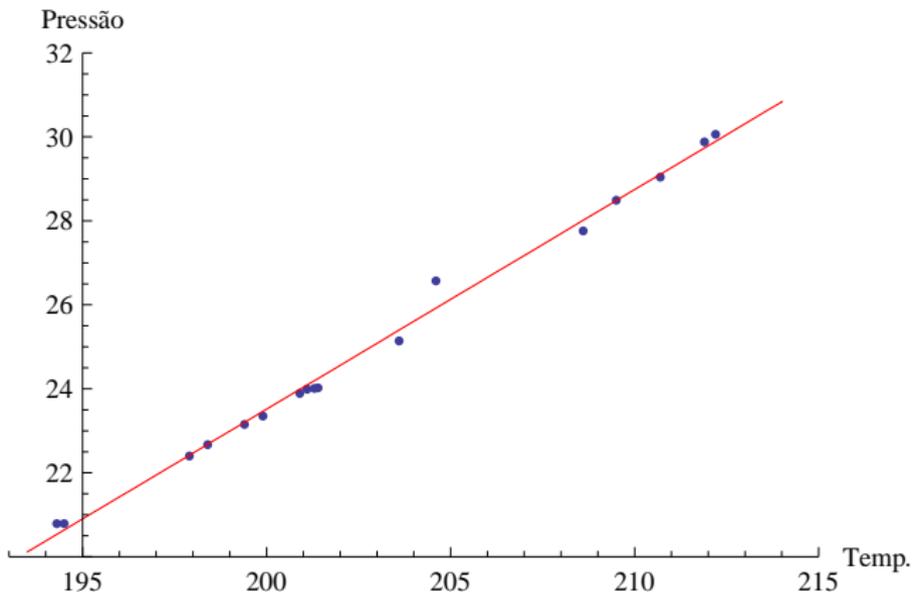


Forbes coletou dados nos Alpes e na Escócia, medindo em cada lugar a pressão através de um barômetro.

Os pontos na figura parecem dispostos ao longo de uma reta, o que nos encoraja a pensar que, para uma dada temperatura, a pressão média pode ser modelada através de uma reta.



Os pontos na figura parecem dispostos ao longo de uma reta, o que nos encoraja a pensar que, para uma dada temperatura, a pressão média pode ser modelada através de uma reta.



Análise baseada em dados

- Uma análise empírica requer dados
- Após a coleta dos dados sobre as variáveis relevantes, os métodos econométricos são usados para estimar os parâmetros do modelo e formalmente testar as hipóteses de interesse
- Em alguns casos, o modelo é usado para se fazer previsão
- A coleta de dados é muito importante em trabalhos empíricos

Estrutura dos dados socioeconômicos

- Dados de Corte Transversal (*cross-section*)
- Cortes Transversais Agrupados
- Dados em painel (ou Dados Longitudinais)
- Dados de séries temporais

Dados de Corte Transversal

- Consiste em uma amostra de uma unidade de análise, tomada em um determinado ponto no tempo
- Dados muito utilizados nas ciências sociais e na avaliação de políticas públicas
- É importante avaliar o universo de onde foi extraída a amostra (e.g., informações sobre salários serão obtidas a partir das pessoas que trabalham)
- Bases nacionais de dados de corte transversal: cada edição da PNAD, do Censo, da Pesquisa de Orçamento Familiar, etc, é uma base de dados de corte transversal

Dados de Corte Transversal: exemplo

Base de dados de corte transversal sobre salários e outras características dos indivíduos

<i>nobsi</i>	<i>salárioh</i>	<i>educ</i>	<i>exper</i>	<i>feminino</i>	<i>casado</i>
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
.
.
.
525	11,56	16	5	0	1
526	3,50	14	5	1	0

Dados de Corte Transversal: exemplo

Base de dados sobre taxas de crescimento econômico e características de um país

<i>nobsp</i>	<i>país</i>	<i>cpibpcr</i>	<i>consgov60</i>	<i>second60</i>
1	Argentina	0,89	9	32
2	Áustria	3,32	16	50
3	Bélgica	2,56	13	69
4	Bolívia	1,24	18	12
.
.
.
61	Zimbábue	2,30	17	6

Séries temporais

- Observações sobre uma (ou mais) variável(is) ao longo do tempo
- Exemplo: oferta de moeda, taxa de juros, PIB, índice de preços ao consumidor, taxa anual de homicídios, número de automóveis vendidos
- Eventos passados podem influenciar eventos futuros
- Tempo é uma dimensão importante
- Característica essencial das observações econômicas: raramente é possível assumir que as informações são independentes ao longo do tempo
- Informações econômicas geralmente estão relacionadas com seu histórico recente
- Dados tem periodicidade (diária, semanal mensal, anual, etc)
- Comum apresentarem sazonalidade

Séries temporais: exemplo

Salário mínimo, desemprego e dados relacionados para Porto Rico

<i>nobsa</i>	<i>ano</i>	<i>minmed</i>	<i>cobmed</i>	<i>desemp</i>	<i>pnb</i>
1	1950	0,20	20,1	15,4	878,7
2	1951	0,21	20,7	16,0	925,0
3	1952	0,23	22,6	14,8	1.015,9
.
.
.
37	1986	3,35	58,1	18,9	4.281,6
38	1987	3,35	58,2	16,8	4.496,7

Corte Transversal Agrupados

- Características de corte transversais e de séries temporais
- Exemplo: suponha dois estudos sobre famílias de corte transversal com intervalo de 2 anos com famílias diferentes (PNAD, por exemplo)
- Podemos formar um corte transversal agrupado combinando os dois anos
- Modo eficaz de analisar os efeitos de uma nova política de governo: coletando dados anteriores e posteriores a uma determinada mudança de política governamental

Corte Transversal Agrupados: exemplo

Cortes transversais agrupados: 2 anos de preços de moradias

<i>nobsm</i>	<i>ano</i>	<i>preçoc</i>	<i>imppro</i>	<i>arquad</i>	<i>qtdorm</i>	<i>banhos</i>
1	1993	85.500	42	1.600	3	2,0
2	1993	67.300	36	1.440	3	2,5
3	1993	134.000	38	2.000	4	2,5
.
.
.
250	1993	243.600	41	2.600	4	3,0
251	1995	65.000	16	1.250	2	1,0
252	1995	182.400	20	2.200	4	2,0
253	1995	97.500	15	1.540	3	2,0
.
.
.
520	1995	57.200	16	1.100	2	1,5

Dados longitudinais

- Consiste de uma série temporal para cada registro
- Ex: Suponha histórico do salário, educação e emprego para um conjunto de indivíduos ao longo de um período de 10 anos
- Ex: Coleta de informações (investimento, finanças, etc) para um conjunto de empresas por um período de 5 anos
- Também podemos coletar dados em painel para recortes geográficos (como município, estado, país)
- característica essencial: as *mesmas unidades* são acompanhadas ao longo de um período
- A repetição das unidades de observação constituem uma complicação da obtenção
- Por outro lado: observar as mesmas unidades traz várias vantagens
- Ter múltiplas observações dos mesmos indivíduos nos permite controlar certas características não observáveis
- Pode facilitar inferência causal

Dados longitudinais: exemplo

Conjunto de dados em painel sobre estatísticas de crime nas cidades para dois anos

<i>nobsc</i>	<i>cidade</i>	<i>ano</i>	<i>homicds</i>	<i>população</i>	<i>desemp</i>	<i>polícia</i>
1	1	1986	5	350.000	8,7	440
2	1	1990	8	359.200	7,2	471
3	2	1986	2	64.300	5,4	75
4	2	1990	1	65.100	5,5	75
.
.
.
297	149	1986	10	260.700	9,6	286
298	149	1990	6	245.000	9,8	334
299	150	1986	25	543.000	4,3	520
300	150	1990	32	546.200	5,2	493

Causalidade e a noção de *Ceteris Paribus*

- Na avaliação de políticas públicas, o objetivo é inferir se uma variável tem **efeito causal** sobre outra variável
- Encontrar simplesmente uma relação entre duas ou mais variáveis pode ser sugestivo mas não conclusivo
- A menos que se possa estabelecer causalidade

O conceito *Ceteris Paribus*

“Se todos os demais fatores relevantes permanecem constantes”

- O conceito *ceteris paribus* tem um papel importante na análise causal
- A maior parte das questões de análise de políticas públicas têm um caráter *ceteris paribus*
- No exemplo de capacitação profissional, interessa conhecer, por exemplo, o efeito de uma semana adicional de capacitação sobre o salário, quando todos os demais componentes permanecem inalterados.
- Se for possível manter constantes todos os demais fatores relevantes e se encontra uma relação entre capacitação e salário, pode-se concluir que a capacitação tem um efeito causal sobre a produtividade dos trabalhadores
- A pergunta fundamental na maior parte dos estudos empíricos é:
“Fora mantidos constantes fatores suficientes para que se justifique a causalidade?”
- Poucas vezes um estudo estatístico/econométrico será avaliado sem que surja esta pergunta.
- Se os demais fatores não permanecem constantes, então não se pode saber qual é o efeito específico da variável explicativa de interesse?

- Na maioria das aplicações sérias, a quantidade de fatores que podem ter influencia sobre uma variável é imensa
- Isolar qualquer variável particular pode parecer uma tarefa impossível
- A aplicação cuidadosa de métodos econométricos pode simular experimentos *ceteris paribus*