AULAS 16 E 17
Análise de regressão múltipla: estimação

Ernesto F. L. Amaral

25 e 30 de outubro de 2012
Avaliação de Políticas Públicas (DCP 046)

Fonte:

Cohen, Ernesto, e Rolando Franco. 2000. "Avaliação de Projetos Sociais". São Paulo, SP: Editora Vozes. pp.118-136 (capítulo 7).

Wooldridge, Jeffrey M. "Introdução à econometria: uma abordagem moderna". São Paulo: Cengage Learning, 2008. pp.64-109 (capítulo 3).

CAPÍTULO 7 - COHEN & FRANCO
MODELOS PARA A AVALIAÇÃO DE IMPACTOS

DESENHO DE PESQUISA DE AVALIAÇÃO DE IMPACTO

– Os métodos de estimação de impacto dependem do desenho da avaliação, isto é, se há dados para grupos de tratamento (beneficiários) e controle (comparação).

GRUPO	ANTES	POLÍTICA	DEPOIS
Tratamento	T_0	X	T_1
Controle	C_0		C_1

– "Diferença em diferenças" ou "dupla diferença" (DD) estima:

1) Diferença dentro de cada grupo (tratamento e controle).

2) Diferença dessas duas médias.

$$DD = (T_1 - T_0) - (C_1 - C_0)$$

DESENHOS EXPERIMENTAIS

– Atribuição aleatória, dentre determinados grupos, da oportunidade de participar em programas, definindo grupos de tratamento e controle:
  – Por exemplo, realização de pesquisa para averiguar as regiões pobres.
  – Seleção aleatória de regiões incluídas na política e daquelas que serão o controle.
  – Única diferença entre grupos é o ingresso no programa.
– Avaliação sistemática e mensuração dos resultados em distintos momentos da implementação do programa.
– Se a seleção é aleatória, pode-se dispensar a avaliação anterior à política para ambos os grupos.

	X	T_1
		C_1

DESENHOS QUASE-EXPERIMENTAIS

– O controle é construído com base na propensão do indivíduo de ingressar no programa.

– Busca-se obter grupo de comparação que corresponda ao grupo de beneficiários:

  – Com base em certas características (sociais, econômicas...) estima-se a probabilidade de um indivíduo de participar do programa.

  – Com base nessa propensão (exercício de emparelhamento), constitui-se o grupo de controle.

– Estima-se os efeitos na comparação entre o grupo de tratamento e o grupo de controle, antes e depois do programa.

T_0	X	T_1
C_0		C_1

DESENHOS NÃO-EXPERIMENTAIS

– Ausência de grupos de controle torna mais difícil isolar causas que geram impactos na variável de interesse.

– Pode ser realizada análise reflexiva para estimar efeitos dos programas, com comparação dos resultados obtidos pelos beneficiários antes e depois do programa.

– Modelo antes-depois:

T_0	X	T_1

– Modelo somente depois com grupo de comparação:

	X	T_1	T_2
		C_1	C_2

– Modelo somente depois:

	X	T_1	T_2

DESENHO DA AVALIAÇÃO	MÉTODO DE ESTIMAÇÃO DE IMPACTO
EXPERIMENTAL	COMPARAÇÃO DE MÉDIAS
QUASE-EXPERIMENTAL	REGRESSÃO MÚLTIPLA & DIFERENÇA EM DIFERENÇAS
NÃO-EXPERIMENTAL	REGRESSÃO MÚLTIPLA

CAPÍTULO 3 - WOOLDRIDGE
ANÁLISE DE REGRESSÃO MÚLTIPLA: ESTIMAÇÃO

MODELO DE REGRESSÃO MÚLTIPLA

- A desvantagem de usar análise de **regressão simples** é o fato de ser difícil que todos os outros fatores que afetam y não estejam correlacionados com x.
- Análise de **regressão múltipla** possibilita *ceteris paribus* (outros fatores constantes), pois permite controlar muitos outros fatores que afetam a variável dependente simultaneamente.
- Isso auxilia no teste de teorias e hipóteses, quando possuímos dados não-experimentais.
- Ao utilizar mais fatores na explicação de y, uma maior variação de y será explicada pelo modelo.
- Este é o modelo mais utilizado nas ciências sociais.
- O método de MQO é usado para estimar os parâmetros do modelo de regressão múltipla.

MODELO COM DUAS VARIÁVEIS INDEPENDENTES

$$salárioh = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

– Salário é determinado por educação, experiência e outros fatores não-observáveis (Equação Minceriana).

– β_1 mede o efeito de educação sobre salário, mantendo todos os outros fatores fixos (*ceteris paribus*).

– β_2 mede o efeito de experiência sobre salário, mantendo todos os outros fatores fixos.

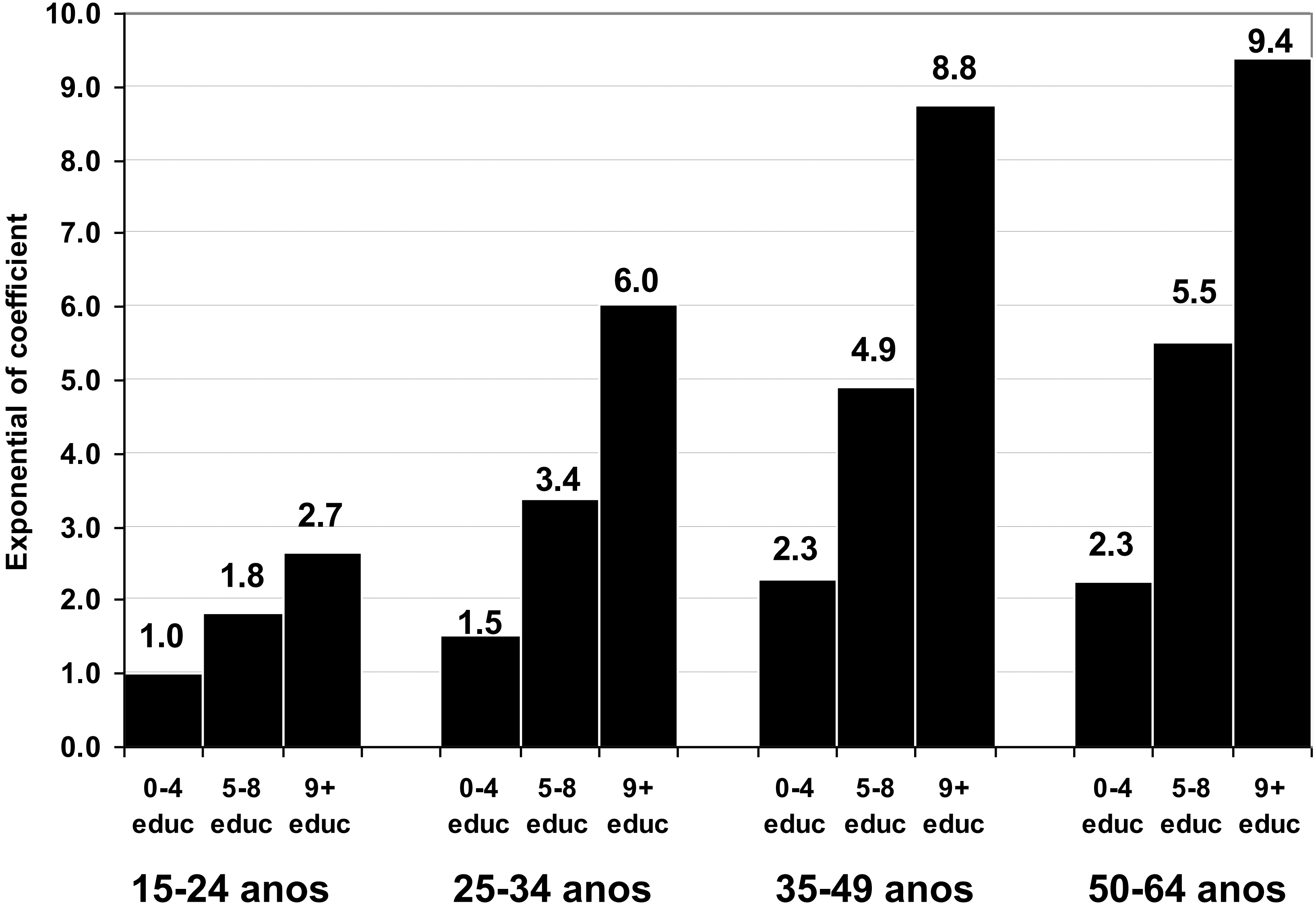
– Como experiência foi inserida na equação, podemos medir o efeito de educação sobre salário, mantendo experiência fixa.

– Na regressão simples, teríamos que assumir que experiência não é correlacionada com educação, o que é uma hipótese fraca.

APLICANDO A EQUAÇÃO DE MINCER (1974) NO BRASIL

Micro-região	Ano	Grupo Idade-Escol.	G11	G12	G13	...	G43	Log da Média da Renda
110006	1970	15-24 anos & 0-4 escol.	1	0	0	...	0	5.82
110006	1970	15-24 anos & 5-8 escol.	0	1	0	...	0	6.21
110006	1970	15-24 anos & 9+ escol.	0	0	1	...	0	6.75
...	...	...	...	...	...	...	...	...
110006	1970	50-64 anos & 9+ escol.	0	0	0	...	1	7.73
...	...	...	...	...	...	...	...	...

EFEITOS DE GRUPOS DE IDADE-ESCOLARIDADE NA RENDA DOS TRABALHADORES: BRASIL, 1970



Fonte: Censos Demográficos Brasileiros 1970 a 2000 (IBGE).

MODELO GERAL DE DUAS VARIÁVEIS INDEPENDENTES

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

– β_0 é o intercepto.

– β_1 mede a variação em y com relação a x_1, mantendo os outros fatores constantes.

– β_2 mede a variação em y com relação a x_2, mantendo os outros fatores constantes.

RELAÇÕES FUNCIONAIS ENTRE VARIÁVEIS

– A regressão múltipla é útil para generalizar relações funcionais entre variáveis.

– Por exemplo:

$$cons = \beta_0 + \beta_1 rend + \beta_2 rend^2 + u$$

– Variação no consumo decorrente de variação na renda é:

$$\frac{\Delta cons}{\Delta rend} \approx \beta_1 + 2\beta_2 rend$$

– O efeito marginal da renda sobre o consumo depende tanto de β_2 como de β_1 e do nível de renda.

– A definição das variáveis independentes é sempre importante na interpretação dos parâmetros.

HIPÓTESE SOBRE *u* EM RELAÇÃO A x_1 E x_2

$$E(u|x_1,x_2)=0$$

– Para qualquer valor de x_1 e x_2 na população, o fator não-observável médio é igual a zero.

– Isso implica que outros fatores que afetam *y* não estão, em média, relacionados com as variáveis explicativas.

– Os níveis médios dos fatores não-observáveis devem ser os mesmos nas combinações das variáveis independentes.

– A esperança igual a zero significa que a relação funcional entre as variáveis explicada e as explicativas está correta.

– No exemplo da renda ao quadrado, não é preciso incluir $rend^2$, já que ela é conhecida quando se conhece *rend*:

$$E(u|rend)=0$$

MODELO COM *k* VARIÁVEIS INDEPENDENTES

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

– Esse é o modelo de regressão linear múltipla geral ou, simplesmente, modelo de regressão múltipla.

– Há k + 1 parâmetros populacionais desconhecidos, já que temos k variáveis independentes e um intercepto.

– Os parâmetros β_1 a β_k são chamados de parâmetros de inclinação, mesmo que eles não tenham exatamente este significado.

– **A regressão é "linear" porque é linear nos β_j, mesmo que seja uma relação não-linear entre a variável dependente e as variáveis independentes:**

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_3 x_2^2 + u$$

OBTENÇÃO DAS ESTIMATIVAS DE MQO

– Reta de regressão de MQO ou função de regressão amostral (FRA):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

– O método de mínimos quadrados ordinários escolhe as estimativas que minimizam a soma dos resíduos quadrados.

– Dadas *n* observações de *y*, x_1, x_2, ... e x_k, as estimativas dos parâmetros são escolhidas para fazer com que a expressão abaixo tenha o menor valor possível:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}\right)^2$$

INTERPRETAÇÃO DA EQUAÇÃO DE REGRESSÃO

– Novamente a reta de regressão de MQO:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + u$$

– O intercepto é o valor previsto de *y* quando todas as variáveis independentes são iguais a zero.

– As estimativas dos demais parâmetros têm interpretações de efeito parcial (*ceteris paribus*).

– Da equação acima, temos:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \cdots + \hat{\beta}_k \Delta x_k$$

– O coeficiente de x_1 mede a variação em *y* devido a um aumento de uma unidade em x_1, mantendo todas as outras variáveis independentes constantes:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1, sendo: \Delta x_2 = \cdots = \Delta x_k = 0$$

SIGNIFICADO DE “MANTER OUTROS FATORES FIXOS”

– Regressão múltipla permite interpretação *ceteris paribus* mesmo que dados não sejam coletados de maneira *ceteris paribus*.

– Os dados são coletados por amostra aleatória que não estabelece restrições sobre os valores a serem obtidos das variáveis independentes.

– Ou seja, a regressão múltipla permite simular situação de outros fatores constantes, sem restringir a coleta de dados.

– Essa modelagem permite realizar em ambientes não-experimentais o que cientistas naturais realizam em experimentos de laboratório (mantendo outros fatores fixos).

– A **avaliação de impacto de políticas** pode ser realizada com regressão múltipla, mensurando relação entre variáveis independentes e dependente, com noção de *ceteris paribus*.

GRAU DE AJUSTE

– O R^2 nunca diminui quando outra variável independente é adicionada na regressão.

– Isso ocorre porque a soma dos resíduos quadrados nunca aumenta quando variáveis explicativas são acrescentadas ao modelo.

– Essa característica faz de R^2 um teste fraco para decidir pela inclusão de variáveis no modelo.

– O efeito parcial da variável independente (β_k) sobre y é o que deve definir se a variável deve ser inserida no modelo.

– R^2 é um grau de ajuste geral do modelo, assim como um teste para indicar o quanto um grupo de variáveis explica variações em y.

REGRESSÃO ATRAVÉS DA ORIGEM

– Em alguns modelos, pode-se avaliar que o ideal seria ter β_0 igual a zero:

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \cdots + \tilde{\beta}_k x_k$$

– R^2 pode ser negativo, o que significa que a média amostral de y "explica" mais da variação em y_i do que as variáveis independentes.

– Nesse caso, devemos incluir um intercepto ou procurar novas variáveis explicativas.

– Se β_0 for diferente de zero na população, a regressão através da origem gera estimadores dos parâmetros de inclinação (β_k) viesados.

– Se β_0 for igual a zero na população, a regressão com intercepto gera maiores variâncias dos estimadores de inclinação.

VALOR ESPERADOS DOS ESTIMADORES DE MQO

HIPÓTESE RLM.1 (LINEAR NOS PARÂMETROS)

– Modelo na população pode ser escrito como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

– β_0, β_1,..., β_k são parâmetros desconhecidos (constantes) de interesse, e *u* é um erro aleatório não-observável ou um termo de perturbação aleatória.

HIPÓTESE RLM.2 (AMOSTRAGEM ALEATÓRIA)

– Temos uma amostra aleatória de *n* observações do modelo populacional acima.

HIPÓTESE RLM.3 (MÉDIA CONDICIONAL ZERO)

– O erro *u* tem um valor esperado igual a zero, dados quaisquer valores das variáveis independentes:

$$E(u|x_1,x_2,...,x_k)=0$$

HIPÓTESE RLM.4 (COLINEARIDADE NÃO PERFEITA)

– Na amostra e na população, nenhuma das variáveis independentes é constante, e não há relações lineares exatas entre as variáveis independentes.

– As variáveis independentes devem ser correlacionadas entre si, mas não deve haver **colinearidade perfeita** (por exemplo, uma variável não pode ser múltiplo de outra).

– Altos graus de correlação entre variáveis independentes e tamanho pequeno da amostra aumentam variância de beta.

– Correlação alta (mas não perfeita) entre duas ou mais variáveis não é desejável (**multicolinearidade**).

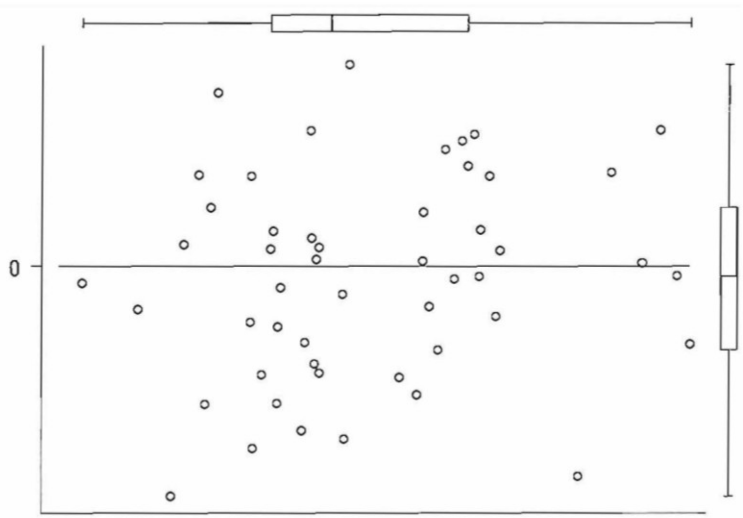
– Por outro lado, **se a correlação for nula**, não é necessário regressão múltipla, mas sim regressão simples, já que o termo de erro englobaria todos fatores não-observáveis e não-relacionados com as variáveis independentes.

HIPÓTESE RLM.5 (HOMOSCEDASTICIDADE)

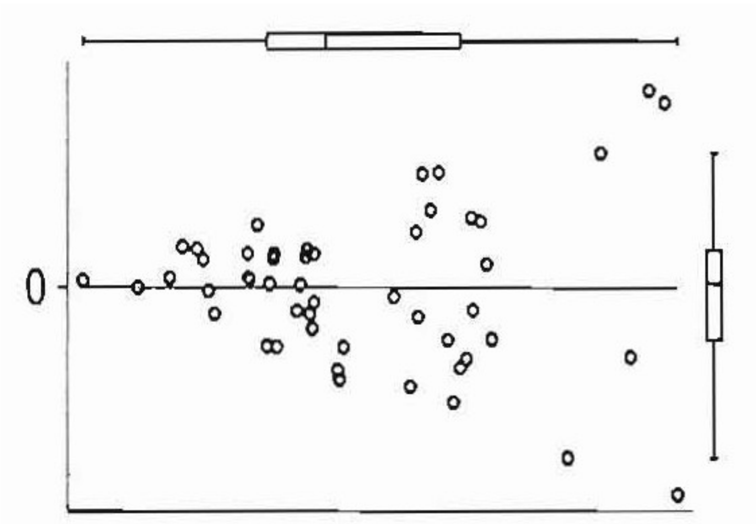
– A variância do termo erro (*u*), condicionada às variáveis explicativas, é a mesma para todas as combinações de resultados das variáveis explicativas.

– Se essa hipótese é violada, o modelo exibe heteroscedasticidade.

HOMOSCEDASTICIDADE



HETEROSCEDASTICIDADE



Fonte: Hamilton, 1992: 52-53.

TEOREMA DE GAUSS-MARKOV

– Sob as hipóteses RLM.1 a RLM.5, os parâmetros estimados do intercepto e de inclinação são os melhores estimadores lineares não-viesados dos parâmetros populacionais:

Best Linear Unbiased Estimators (BLUEs)

– Em outras palavras, os estimadores de mínimos quadrados ordinários (MQO) são os melhores estimadores lineares não-viesados.