AULA 09
Análise de regressão múltipla com informações qualitativas

Ernesto F. L. Amaral

25 de julho de 2013
Análise de Regressão Linear (MQ 2013)
www.ernestoamaral.com/mq13reg.html

Fonte:
Wooldridge, Jeffrey M. "Introdução à econometria: uma abordagem moderna". São Paulo: Cengage Learning, 2008. Capítulo 7 (pp.207-242).

VARIÁVEIS INDEPENDENTES QUALITATIVAS

– No decorrer do curso, já conversamos sobre incorporação de fatores qualitativos nos modelos de regressão.

– Variáveis independentes já foram utilizadas como informações qualitativas e não somente como quantitativas.

– Alguns exemplos foram variáveis de sexo, de opinião e de valores, além de diferentes categorizações de idade e escolaridade.

– As variáveis binárias ou variáveis *dummy* ou variáveis dicotômicas são formas de agregar informações qualitativas em modelos de regressão estatística.

DEFINIÇÃO DOS VALORES E NOMES

– É preciso definir qual evento será atribuído o valor um e qual será atribuído o valor zero.

– A maneira que denominamos nossas variáveis não tem importância para obter os resultados da regressão, mas ajuda a escolher seus nomes.

– Os valores zero e um são utilizados para facilitar a interpretação dos parâmetros da regressão.

– Outros valores diferentes serviriam para montar a variável, mas dificultaria o entendimento dos betas estimados.

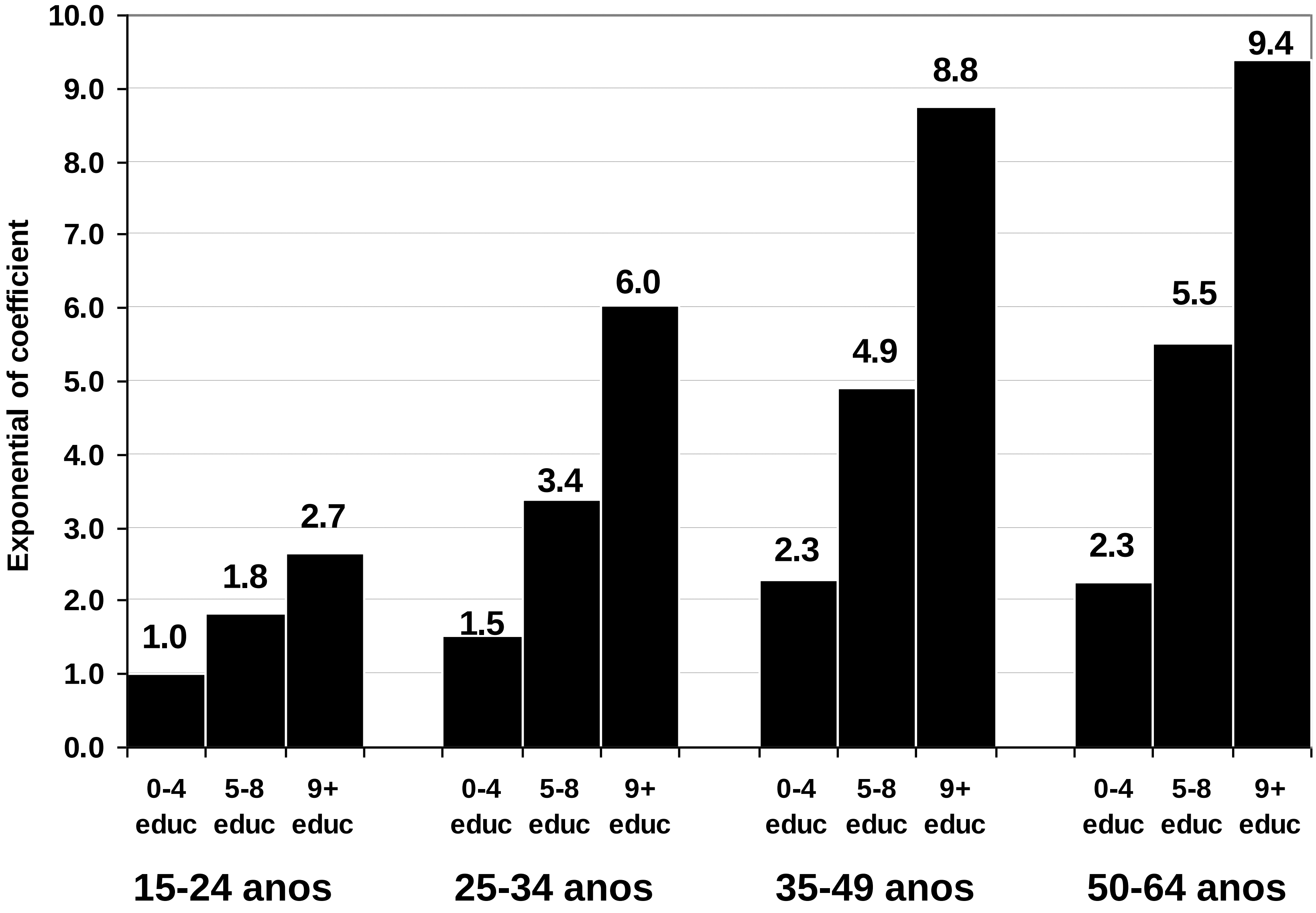
EXEMPLO DE DADOS DE CORTE TRANSVERSAL

Número da observação	Salário por hora	Sexo	Masculino	Feminino	Estado civil	Estado civil (casado)	Estado civil (outros)
1	3,10	4	0	1	3	0	1
2	3,24	4	0	1	2	1	0
3	3,00	2	1	0	4	0	1
4	6,00	2	1	0	2	1	0
5	5,30	2	1	0	2	1	0
...	...	...	...	...	...	...	...
525	11,56	2	1	0	2	1	0
526	3,50	4	0	1	4	0	1

DADOS DO CENSO DEMOGRÁFICO DO BRASIL, 1970

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).

UMA ÚNICA VARIÁVEL BINÁRIA INDEPENDENTE

– Com somente uma variável dicotômica explicativa, simplesmente adicionamos a variável à equação como uma variável independente:

$$\text{salário} = \beta_0 + \delta_0\ \text{feminino} + \beta_1\ \text{educação} + u$$

– É utilizado o "δ" para ressaltar que o parâmetro da variável "feminino" é interpretado como informação dicotômica.

– δ_0 é a diferença no salário entre mulheres e homens, dado o mesmo grau de educação e o mesmo termo de erro u.

– Se $\delta_0 < 0$, as mulheres ganham em média menos que os homens, para o mesmo nível dos outros fatores.

– Se $\delta_0 > 0$, as mulheres ganham em média mais que os homens, para o mesmo nível dos outros fatores.

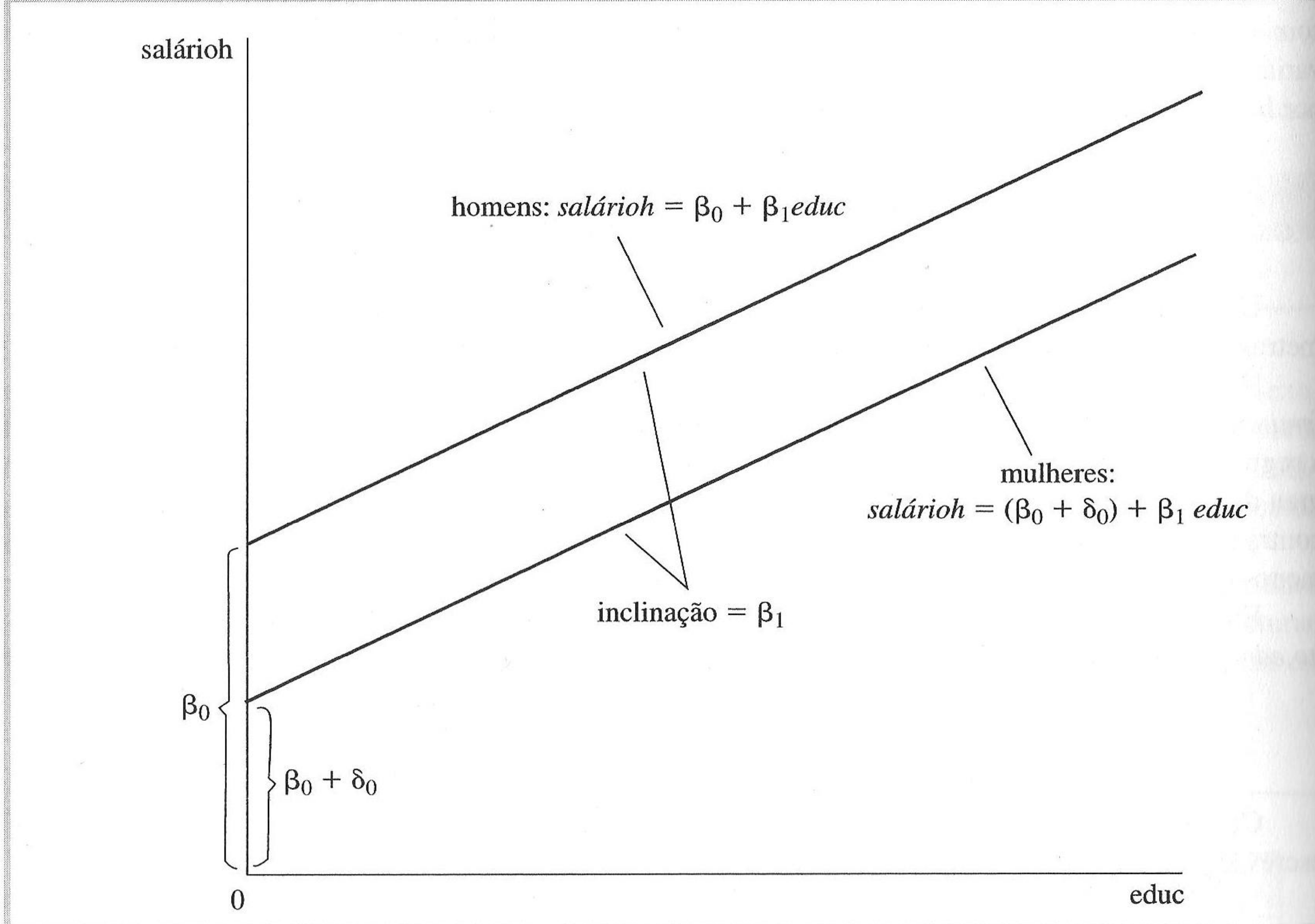
– A diferença entre homens e mulheres pode ser descrita graficamente como um **deslocamento de intercepto** entre as linhas que representam cada um dos sexos.

DESLOCAMENTO DE INTERCEPTO ENTRE SEXOS

– A diferença entre sexos não depende do nível de educação, por isso as retas são paralelas.

Figura 7.1

Gráfico de $salárioh = \beta_0 + \delta_0 feminino + \beta_1 educ$ para $\delta_0 < 0$.



ARMADILHA DA VARIÁVEL *DUMMY*

$$\text{salário} = \beta_0 + \delta_0 \text{ feminino} + \beta_1 \text{ educação} + u$$

- Na equação acima, o intercepto para homens é β_0 e o intercepto para mulheres é $\beta_0 + \delta_0$.
- Por isso, seria redundante incluir uma variável binária “masculino”, além da variável “feminino”.
- Como existem apenas dois grupos, são necessários apenas dois interceptos diferentes.
- O uso de duas variáveis binárias introduziria colinearidade perfeita, porque *feminino* + *masculino* = 1, o que significa que *masculino* é uma função linear perfeita de *feminino*.
- Os homens foram escolhidos para ser o grupo de referência (ou grupo base), que é o grupo contra o qual as comparações são realizadas:

$$\text{salário} = \alpha_0 + \gamma_0 \text{ masculino} + \beta_1 \text{ educação} + u$$

$$\alpha_0 + \gamma_0 = \beta_0 \quad \& \quad \alpha_0 = \beta_0 + \delta_0$$

RETIRADA DO INTERCEPTO GLOBAL

– Também é possível eliminar o intercepto global do modelo:

$$\text{salário} = \beta_0 \text{ masculino} + \alpha_0 \text{ feminino} + \beta_1 \text{ educação} + u$$

– Essa formulação não oferece uma maneira fácil de verificar diferenças nos interceptos.

– Lembremos que não existe uma maneira consensual de computar o R-quadrado em regressões sem intercepto.

– Por isso, geralmente é sempre incluído um intercepto global para o grupo de referência.

HIPÓTESE NULA E HIPÓTESE ALTERNATIVA

$$\text{salário} = \beta_0 + \delta_0 \text{ feminino} + \beta_1 \text{ educação} + u$$

- A hipótese nula de não-existência de diferença entre homens e mulheres será H_0: $\delta_0 = 0$.

- A hipótese alternativa de que existe discriminação contra as mulheres será H_1: $\delta_0 < 0$.

- Quando algumas variáveis independentes são binárias:
  - Nada muda: (1) na mecânica do MQO; (2) na teoria estatística; e (3) na estatística de *t*.
  - A única diferença é a interpretação do coeficiente da variável binária.

TESTE DE COMPARAÇÃO DE MÉDIAS

– A regressão simples sobre uma constante e uma variável binária é uma maneira objetiva de comparar as médias de dois grupos:

$$\text{salário} = \beta_0 + \delta_0 \text{ feminino} + u$$

– β_0 é o salário médio dos homens na amostra (feminino = 0).

– $\beta_0 + \delta_0$ é o salário médio das mulheres na amostra.

– Para que o teste *t* seja válido, é assumida a hipótese de homoscedasticidade, o que significa que a variância populacional dos salários dos homens é a mesma dos salários das mulheres.

ANÁLISE DE POLÍTICAS PÚBLICAS

– Variáveis binárias independentes refletem características predeterminadas (sexo), escolhas de indivíduos, unidades econômicas ou recebimento de políticas públicas:

$$\text{salário} = \beta_0 + \delta_0 \text{ bolsa família} + \beta_1 \text{ educação} + u$$

– É de se supor que indivíduos com baixa escolaridade tenham maior possibilidade de receber bolsa família.

– Por isso, é importante controlar o modelo por educação, porque gostaríamos de saber o efeito médio sobre salário se escolhermos um indivíduo aleatoriamente e dermos a ele o benefício do bolsa família.

– Na avaliação de políticas públicas, é interessante controlar por outros fatores, com o intuito de verificar se o efeito positivo sobre o salário de receber o bolsa família desaparece, ou se torna significativamente menor.

GRUPOS DE CONTROLE E DE TRATAMENTO

– Como vimos anteriormente, podemos classificar os indivíduos em grupos de controle (não recebeu a política) e grupo experimental ou de tratamento (recebeu a política).

– No entanto, sabemos que a escolha destes grupos não é realizada aleatoriamente, como é o caso das ciências naturais.

– Efeito causal da política será melhor estimado: (1) ao controlar modelo por uma maior quantidade de fatores; e (2) se beneficiários da política foram definidos aleatoriamente.

– De todo modo, a análise de regressão múltipla pode ser usada para controlar um número suficiente de outros fatores para estimar o efeito causal de uma política pública.

QUANDO VARIÁVEL DEPENDENTE É LOG(Y)

– Quando a variável dependente aparece na forma logarítmica, com uma ou mais variáveis binárias independentes, os coeficientes têm interpretação percentual.

– Ou seja, quando log(y) é a variável dependente em um modelo, o coeficiente de uma variável binária, quando multiplicado por 100, é interpretado como a diferença percentual em y, mantendo os outros fatores constantes.

– Quando o coeficiente de uma variável binária indica uma grande mudança proporcional em y, o cálculo da semi-elasticidade indica a diferença percentual exata:

$$100 * [\exp(\beta_1) - 1]$$

EXEMPLO 7.5 DA PÁGINA 215

log (salário-hora) = 0,417 - 0,297 feminino + ...

– 100 * 0,297 =>mulheres ganham 29,7% menos que homens.

– Queremos saber a diferença proporcional nos salários entre mulheres e homens, mantendo fixos todos os outros fatores:

$$\frac{sal.mul. - sal.hom.}{sal.hom} = \frac{sal.mul.}{sal.hom} - \frac{sal.hom.}{sal.hom} = \frac{sal.mul.}{sal.hom} - 1$$

– O modelo de regressão nos fornece:

$$log(sal.mul.) - log(sal.hom.) = \frac{\log(sal.mul.)}{\log(sal.hom.)} = -0{,}297$$

– Então:

$$\frac{sal.mul.}{sal.hom} - 1 = exp\left(\frac{\log(sal.mul.)}{\log(sal.hom.)}\right) - 1 = exp(-0{,}297) - 1 = 0{,}743 - 1 \approx -0{,}257$$

– 100 * 0,257 =>mulheres ganham 25,7% menos que homens.

VARIÁVEIS BINÁRIAS PARA CATEGORIAS MÚLTIPLAS

– Podemos utilizar mais de uma variável binária no modelo:

$$\log(\text{salário}) = \beta_0 + \beta_1 \text{ feminino} + \beta_2 \text{ casado}$$

– No exemplo acima, o prêmio por ser casado é assumido como o mesmo para homens e mulheres.

– Para considerar efeitos diferentes de ser casado por sexo, precisamos comparar quatro grupos: homens casados, mulheres casadas, homens solteiros e mulheres solteiras.

– Digamos que escolhemos homens solteiros como grupo de referência:

$$\log(\text{salário}) = \beta_0 + \beta_1 \text{ hcasados} + \beta_2 \text{ mcasadas} + \beta_3 \text{ msolteiras}$$

– As estimativas das três variáveis binárias medem a diferença proporcional nos salários relativamente aos homens solteiros.

– Como vimos, essa é a mesma idéia de realizar variáveis binárias combinando idade e escolaridade.

PRINCÍPIO GERAL PARA INCLUSÃO DE BINÁRIAS

– Se o modelo de regressão deve ter interceptos para g grupos, precisamos incluir g - 1 variáveis binárias e o intercepto.

– O intercepto do grupo de referência é o intercepto global no modelo.

– O coeficiente da variável binária representa a diferença estimada nos interceptos daquele grupo, em relação ao grupo de referência.

– A inclusão de g variáveis binárias juntamente com um intercepto resultará na armadilha da variável binária (colinearidade perfeita).

– Uma alternativa é incluir g variáveis binárias e excluir o intercepto global.

– No entanto, isso dificulta a interpretação de diferenças em relação ao grupo base e o R^2 é calculado diferentemente.

INFORMAÇÕES ORDINAIS COM VARIÁVEIS BINÁRIAS

– As categorias de variáveis ordinais podem ser organizadas em alguma ordem.

– Sabemos que há diferenças relativas entre os valores dos dados, mas não sabemos as magnitudes das diferenças.

– Por exemplo, na escala de frequência “pouco/médio/muito”, é possível ordenar os dados, mas não sabemos se a diferença entre “pouco” e “médio” é a mesma que aquela existente entre “médio” e “muito”.

– Aqui não faz sentido supor que o aumento de uma unidade nessa variável terá um efeito constante sobre outra variável.

– É possível criar três variáveis binárias, tomando uma como referência.

– No caso de variáveis com muitos valores, podemos dividi-la em categorias (escolaridade, por exemplo).

– Classificação pode afetar a significância de outras variáveis.

INTERAÇÕES ENTRE VARIÁVEIS BINÁRIAS

– Como vimos no exemplo de estado civil e sexo, podemos realizar interações entre variáveis binárias:

$$\log(\text{salário}) = \beta_0 + \beta_1 \text{ hcasados} + \beta_2 \text{ mcasadas} + \beta_3 \text{ msolteiras}$$

– A equação acima permite testar diretamente diferenças entre qualquer grupo e homens solteiros.

– Podemos adicionar um termo de interação diretamente:

$$\log(\text{salário}) = \beta_0 + \beta_1 \text{ feminino} + \beta_2 \text{ casado} + \beta_3 \text{ fem*casado}$$

– Os coeficientes de cada grupo serão:

- – Homens solteiros: β_0
- – Homens casados: $\beta_0 + \beta_2$
- – Mulheres solteiras: $\beta_0 + \beta_1$
- – Mulheres casadas: $\beta_0 + \beta_1 + \beta_2 + \beta_3$

– Nessa equação, β_3 permite testar diretamente se diferencial de sexo depende do estado civil e vice-versa.

INCLINAÇÕES DIFERENTES

– Existem casos de interação de variáveis binárias com variáveis explicativas que não são binárias para permitir diferença nas inclinações.

– Podemos testar se retorno da educação é o mesmo para homens e mulheres, considerando um diferencial de salários constante entre homens e mulheres:

$$\log(\text{salário}) = (\beta_0 + \delta_0 \text{ feminino}) + (\beta_1 + \delta_1 \text{ feminino})*\text{educ} + u$$

– Homens: intercepto (β_0) e inclinação (β_1)

– Mulheres: intercepto ($\beta_0 + \delta_0$) e inclinação ($\beta_1 + \delta_1$)

– δ_0: diferença nos interceptos entre mulheres e homens.

– δ_1: diferença no retorno da educação entre sexos.

– No Stata:

$$\log(\text{salário}) = \beta_0 + \delta_0 \text{ feminino} + \beta_1 \text{ educ} + \delta_1 \text{ fem*educ} + u$$

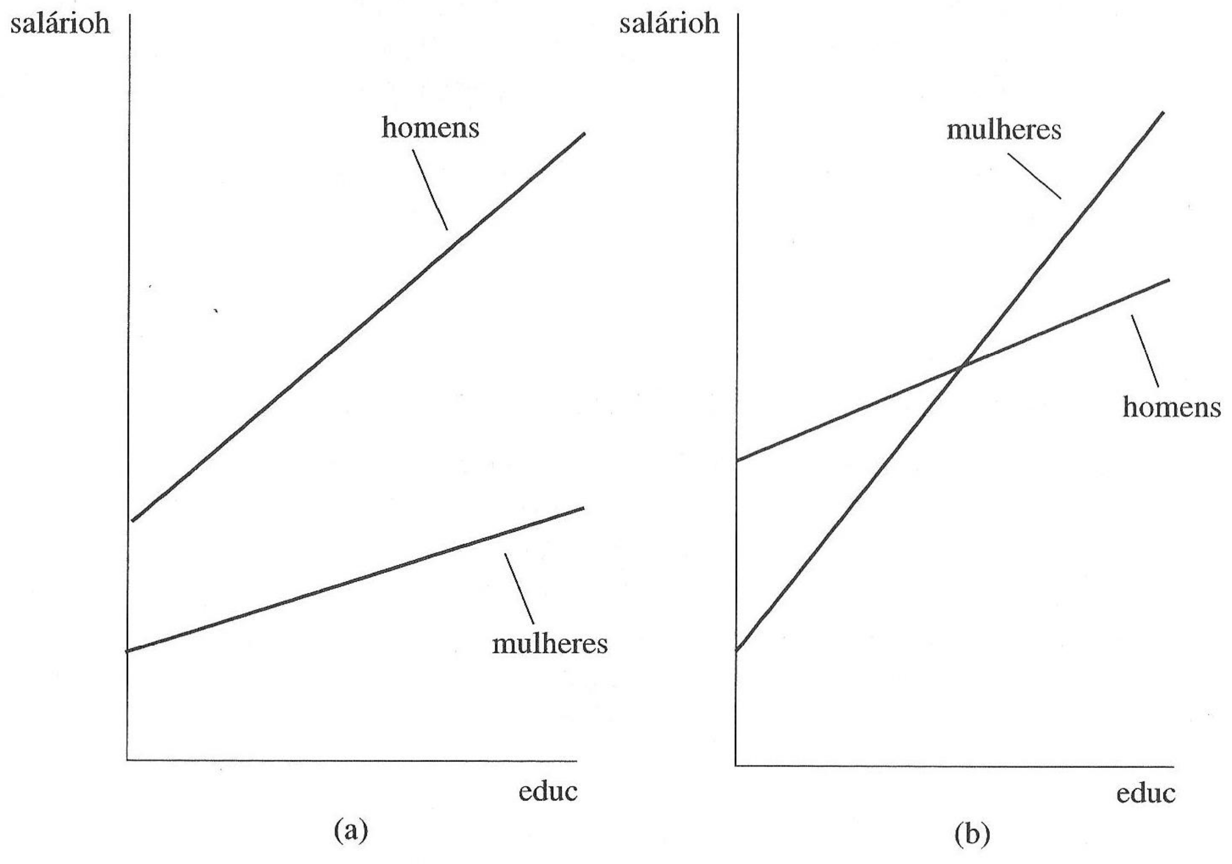
– Quando $\delta_0 + (\delta_1 * \text{educ}) = 0$, salário é igual entre sexos.

GRÁFICO A: intercepto e inclinação das mulheres é inferior.

GRÁFICO B: intercepto das mulheres é inferior, mas inclinação é superior.

Figura 7.2

Gráficos da equação (7.16). (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



REGRESSÕES ENTRE GRUPOS

– De uma forma geral, as interações permitem verificar se a relação (inclinação) entre uma variável independente e uma dependente é diferente para grupos distintos.

– Suponha que temos o seguinte modelo:

$$\text{salário} = \beta_0 + \beta_1 \text{ idade} + \beta_2 \text{ educ} + u$$

– Testar se qualquer uma das interações depende de sexo:

$$\text{salário} = \beta_0 + \delta_0 \text{ feminino} + \beta_1 \text{ idade} + \delta_1 \text{ feminino*idade} +$$
$$+ \beta_2 \text{ educ} + \delta_2 \text{ feminino*educ} + u$$

– δ_0: diferença nos interceptos entre mulheres e homens.

– δ_1: diferença de inclinação em relação à idade entre mulheres e homens.

– δ_2: diferença de inclinação em relação à escolaridade entre mulheres e homens.

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0$$

DIFERENÇA ENTRE GRUPOS

– Para testar a hipótese nula:

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0,$$

devemos estimar um modelo restrito (sem “feminino” e interações) e compará-lo ao modelo com interações, a partir da estatística F ou R^2 ajustado (modelos aninhados).

– Pode ser que o teste indique que devemos rejeitar a hipótese nula, mesmo que coeficientes de interação não sejam estatisticamente significantes individualmente.

– Este é nosso modelo:

$$\text{salário} = \beta_0 + \delta_0 \text{ feminino} + \beta_1 \text{ idade} + \delta_1 \text{ feminino*idade} + \\ + \beta_2 \text{ educ} + \delta_2 \text{ feminino*educ} + u$$

– A diferença entre homens e mulheres é dada por:

$$\delta_0 + \delta_1 \text{ idade} + \delta_2 \text{ educ},$$

considerando valores específicos para idade e escolaridade.

MODELO COM MUITAS VARIÁVEIS INDEPENDENTES

– Com poucas variáveis independentes, é fácil adicionar todas interações para testar diferenças entre grupos.

– Se tivéssemos muitas variáveis independentes, seria difícil a inclusão de muitas interações com “feminino”.

– Podemos utilizar a soma dos resíduos quadrados da estatística F para estimar esta diferença.

– Modelo restrito: hipótese é que cada beta é o mesmo nos dois grupos. Envolve k+1 restrições (tirar binária e interações).

– Modelo irrestrito:

  – (1) intercepto; (2) variável binária de grupo; (3) variáveis originais; e (4) k termos de interação.

  – Esse modelo tem [n - 2(k +1)] graus de liberdade.

  – Na verdade, queremos testar se intercepto e todas inclinações são iguais em dois grupos:

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + ... + \beta_{g,k}x_k + u, \text{ para g=1 e g=2}$$

ESTATÍSTICA DE CHOW

– A soma dos resíduos quadrados do modelo sem restrições é obtida de duas regressões separadas, uma de cada grupo:

  – SQR_1 é a soma dos resíduos quadrados do modelo para o primeiro grupo.

  – SQR_2 é a soma dos resíduos quadrados do modelo para o segundo grupo.

  – Modelo irrestrito: $SQR_{ir} = SQR_1 + SQR_2$

– Modelo restrito: SQR do agrupamento dos grupos em regressão única (SQR_P), sem *dummy* e interações.

– Estatística F pode ser calculada (estatística de Chow):

$$F = \frac{[SQR_P - (SQR_1 + SQR_2)]}{SQR_1 + SQR_2} \cdot \frac{[n - 2(k+1)]}{k+1}$$

– Devemos ter homoscedasticidade (variâncias dos erros dos dois grupos devem ser iguais).

LIMITAÇÕES DA ESTATÍSTICA DE CHOW

– Uma limitação do teste de Chow é a hipótese nula não permitir nenhuma diferença entre os grupos.

– Seria interessante considerar diferença nos interceptos entre os grupos e depois verificar diferenças das inclinações.

– Duas formas para interceptos diferirem sob hipótese nula:
  – Com poucas variáveis independentes, estimar modelo restrito que não possui interações, mas possui variável binária do grupo, e compará-lo com modelo irrestrito.
  – Com muitas variáveis independentes, estimar modelo restrito que não possui interações, mas possui variável binária do grupo (novo SQR_P), e incluir na fórmula da estatística de Chow.

ALTERNATIVA À ESTATÍSTICA DE CHOW

– Como vimos, se tivéssemos muitas variáveis independentes, seria difícil a inclusão de muitas interações com "feminino".

– Após a inclusão de tais interações, poderíamos utilizar a Estatística de Chow para estimar a diferença entre os modelos sem e com interações.

– Ou podemos usar a opção "i" em conjunto com interações:

```
xi: reg salario i.fem*i.idade i.fem*i.educ i.fem*i.raça
```

– Depois, faríamos um teste de F para testar se os efeitos das variáveis com interação são iguais a zero, em conjunto.

VARIÁVEL DEPENDENTE BINÁRIA

– Podemos usar uma regressão múltipla para explicar uma variável dependente binária, com valores zero ou um.

– Como *y* pode assumir somente dois valores, β_j não pode ser interpretado como a mudança em y devido ao aumento de uma unidade em x_j, mantendo fixos todos os outros fatores.

– Quando *y* é uma variável binária, assumindo valores zero e um, a probabilidade de sucesso (probabilidade de *y* ser igual a 1) é a mesma do valor esperado de *y*:

$$p(\mathbf{x}) = P(y=1|\mathbf{x}) = E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

onde **x** representa todas variáveis explicativas.

– $P(y=1|\mathbf{x})$ é uma função linear de x_j e é chamada de **probabilidade de resposta**.

– O modelo de regressão linear múltipla com uma variável dependente binária é chamado de **modelo de probabilidade linear (MPL)**, porque a probabilidade de resposta é linear nos parâmetros β_j.

INTERPRETAÇÃO DOS COEFICIENTES ESTIMADOS

– O β_0 estimado é a probabilidade de sucesso prevista quando cada x_j é definido como zero.

– O coeficiente de inclinação (β_1) estimado mede a mudança prevista na probabilidade de sucesso de y, quando x_1 aumenta em uma unidade, mantendo fixos os outros fatores.

– Para interpretarmos corretamente um modelo de probabilidade linear, precisamos saber o que constitui um “sucesso”.

– É uma boa idéia dar à variável dependente um nome que descreva o evento $y = 1$:

  – Participação na força de trabalho (partrab).
  – Concluiu ensino médio (ensmed).
  – Recebeu política pública (polpub).
  – Eleito das últimas eleições (eleito).
  – Praticou aborto (aborto).

EXEMPLO DE MODELO DE PROBABILIDADE LINEAR

– Suponha um modelo que explique a probabilidade de estar na força de trabalho (naft = 1):

$$naft = 0{,}586 + 0{,}038\ educ + 0{,}039\ exper - 0{,}0006\ exper^2 + ...$$

– O coeficiente de *educ* significa que, tudo o mais mantido fixo, mais um ano de educação, aumenta a probabilidade de participação na força de trabalho em 0,038.

  – Esse é o efeito marginal de mais um ano de educação na probabilidade de participação na força de trabalho.

– Há probabilidade negativa até 3,84 anos de estudo (gráfico a seguir), mas banco deste exemplo não tem indivíduos com menos de 5 anos de estudo.

– Experiência tem efeito positivo sobre participação na força de trabalho e depois passa a ser negativo.

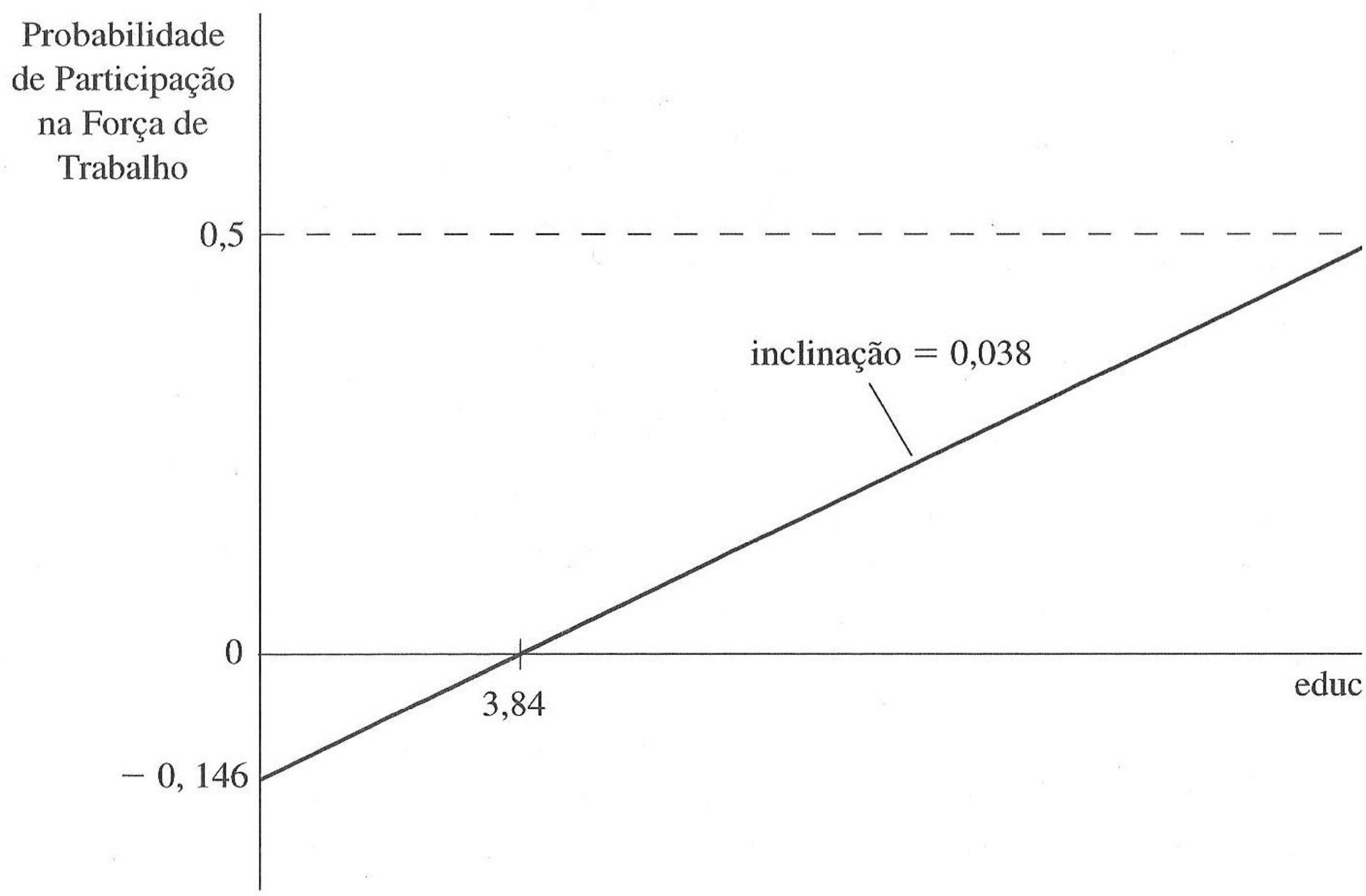
  – A curva muda de direção em 32,5 anos de experiência:

$$\beta \text{ de } exper \,/\, (2 * \beta \text{ de } exper^2) = [0{,}039/(2*0{,}0006)]$$

EXEMPLO DE MODELO DE PROBABILIDADE LINEAR

Figura 7.3

Relação estimada entre a probabilidade de estar na força de trabalho e anos de educação, com outras variáveis explicativas fixas.



DEFICIÊNCIAS DO MODELO DE PROBABILIDADE LINEAR

– É possível que certas combinações de valores das variáveis independentes gerem previsões menores que zero ou maiores que um:

  – Como estas são probabilidades, que devem estar entre zero e um, isso pode ser um pouco complicado.

– A probabilidade não pode ser linearmente relacionada com as variáveis independentes em todos os possíveis valores:

  – Por exemplo, se efeito de passar de 0 para 1 filho menor de 6 anos reduz probabilidade de trabalhar em 0,262, ao passar de 0 para 4 filhos, a probabilidade de trabalhar reduz em 1,048 (0,262*4), o que é impossível (p>1).

– Mesmo com estes problemas, o modelo de probabilidade linear é útil e frequentemente aplicado:

  – Ele funciona bem com valores das variáveis independentes próximos das médias na amostra.

VARIÂNCIA DA VARIÁVEL BINÁRIA

– Quando *y* é uma variável binária, sua variância, condicional em ***x***, é:

$$Var(y|\boldsymbol{x}) = p(\boldsymbol{x})\,[1 - p(\boldsymbol{x})],$$

onde $p(\boldsymbol{x})$ é a forma abreviada da probabilidade de sucesso.

– Haverá heteroscedasticidade no modelo de probabilidade linear.

– Isso não causa viés nos estimadores MQO dos coeficientes beta.

– Porém, somente com homoscedasticidade, os erros-padrão serão válidos e as estatísticas *t* e *F* poderão ser utilizadas.

– O próximo capítulo aborda como corrigir os erros-padrão quanto à heteroscedasticidade.

– Na prática, as estatísticas MQO habituais não ficam muito distorcidas e é aceitável usar estimativas padrão de um modelo de probabilidade linear.

INSERINDO VARIÁVEIS BINÁRIAS INDEPENDENTES

– Podemos incluir variáveis binárias independentes em modelos com variável binária dependente.

– O coeficiente mede a diferença prevista na probabilidade quando a variável binária independente vai de zero a um.

– Por exemplo, se adicionarmos variáveis binárias de raça (branco, negro, hispânico) na explicação da probabilidade de ser preso (pág. 235), obtemos:

$$prisão = 0{,}380 + 0{,}170\ negro + 0{,}096\ hispânico + \ldots$$

– O coeficiente de *negro* significa que, todos os outros fatores iguais, um homem negro tem uma probabilidade 0,17 maior de ser preso que um homem branco (grupo de referência).

– Outra forma de interpretação é dizer que probabilidade de prisão é 17% maior para os negros do que para os brancos.

– Para ser mais exato, negros têm 1,19 [exp(0,17)] vezes mais chance de serem presos do que brancos (ou 19% mais).

USO PRÁTICO DE VARIÁVEIS BINÁRIAS

– Precisamos ser cuidadosos ao avaliarmos variáveis binárias nas Ciências Sociais.

– Na maioria dos casos, as unidades de análise não foram selecionadas aleatoriamente para fazer parte de um grupo ou de outro.

– Devemos procurar saber se fatores não observados que afetam a variável dependente podem estar correlacionados com as variáveis independentes binárias de interesse.

– Ou seja, é preciso incluir fatores que possam estar relacionados com estas variáveis independentes.

– Introdução de informações de períodos anteriores podem ser úteis na estimação do impacto das independentes.

PROBLEMAS DE AUTO-SELEÇÃO

– Indivíduos se auto-selecionam para certos procedimentos ou programas, o que não é uma escolha aleatória:

  – Pessoas escolhem fazer uso de drogas.

  – Crianças entram em programas de saúde infantil por decisão dos pais.

– Ou seja, a participação não é determinada de forma aleatória, havendo problemas de **auto-seleção**.

– Mais uma vez, é preciso incluir outras variáveis na regressão, quando indicador binário de interesse (independente) for relacionado com fatores não-observados.

– Incluímos fatores relacionados com variável independente de interesse para corrigir endogeneidade, mas talvez somente métodos mais avançados serão eficazes.