Econometria

Parte 3

Prof. Adalto Acir Althaus Junior oe

Sumário

- Regressor irrelevante
- Multicolinearidade
- Interações
- Dummy
- Pooled OLS

Outras Questões

Regressor Irrelevante

- O que acontece se incluir um regressor que não deveria estar no modelo?
- Estimamos $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- Mas o modelo real é $y = \beta_0 + \beta_1 x_1 + u$

Resposta: Nós ainda temos uma estimativa consistente de todos os β, onde $\beta 2$ = 0, mas nossos erros padrão aumentam (tornando mais difícil encontrar efeitos estatisticamente significativos) ... veja os próximos slides.

Variância dos Regressores OLS

- Ocorrerá maior variância em suas estimativas dos β's, que são os $\widehat{\beta}_j$, aumentando seus erros padrão, tornando mais difícil encontrar estimativas estatisticamente significativas

- Então, útil saber o que aumenta $\boldsymbol{Var}(\widehat{\beta}_j)$

Variância dos Regressores OLS

- Estimando a variância da inclinação de um modelo OLS:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ij} - \overline{x}_j)^2 (1 - R^2{}_j)}$$

σ^2 é a variância do erro da regressão, u, σ_u^2

- para j = 1,..., k, onde $R^2{}_j$ é o R^2 da regressão x_j em todas as outras variáveis independentes incluindo o intercepto e σ^2 é a variância do erro de regressão, u, também representado por σ_u^2

$$\begin{aligned}
x_1 &= \alpha_0 + \alpha_1 x_2 + \cdots + \alpha_j x_j + e && \rightarrow R_1^2 \\
x_2 &= \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_j x_j + e && \rightarrow R_2^2 \\
&\vdots \qquad\qquad \vdots \\
x_j &= \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{j-1} x_{j-1} + e && \rightarrow R_j^2
\end{aligned}$$

Variância dos Regressores OLS

- Interpretação:
- Como mais variação no x afeta o SE? Por quê?
- Como maior σ^2 afeta o SE? Por quê?
- Como maior ${R^2}_j$ afeta o SE? Por quê?

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ij} - \overline{x}_j)^2 (1 - {R^2}_j)}$$

ou

$$Var(\widehat{\beta}_j) = \frac{\sigma_u^2}{\sum_{i=i}^{n}(x_{ij} - \overline{x_j})^2 (1 - R_j^2)}$$

Variância dos Regressores OLS

$$Var(\widehat{\beta}_j) = \frac{\sigma_u^2}{\sum_{i=i}^{n}(x_{ij} - \overline{x_j})^2 (1 - R_j^2)}$$

- Mais variação em x é bom, menores SE
  - ✓ Intuitivo: Mais variação em x, ajuda a identificar o efeito em y
  - ✓ É por isto que em amostras maiores nos fornecerão maiores variações em x_j

Variância dos Regressores OLS

$$Var(\widehat{\beta}_j) = \frac{\sigma_u^2}{\sum_{i=i}^{n}(x_{ij} - \overline{x_j})^2(1 - R_j^2)}$$

- Maior variância do erro (σ_u^2) levará a maiores SE
  - ✓ Intuitivo: grande parte da variação em y é explicado por outras coisas que não estão no modelo
  - ✓ Pode adicionar variáveis que afetam y (mesmo que não seja necessário para identificação) para melhorar o ajuste!

Variância dos Regressores OLS

$$Var(\widehat{\beta}_j) = \frac{\sigma_u^2}{\sum_{i=i}^{n}(x_{ij} - \overline{x_j})^2(1 - R_j^2)}$$

- Mas mais variáveis também podem ser ruins se forem colineares
  - ✓ Fica mais difícil de separar o efeito das variáveis que são altamente colineares
  - ✓ É por isso que não queremos adicionar variáveis que são "irrelevantes" (ou seja, elas não afetam y)

Devemos incluir variáveis que explicam e estão altamente correlacionadas com nosso interesse?

Discutiremos isso em "bad controls"

Multicolinearidade

- Variáveis altamente colineares podem inflar SEs
  - ✓ Mas, isso não causa um viés ou inconsistência!
  - ✓ O problema é realmente apenas se tiver uma amostra muito pequena; com uma amostra maior, pode-se obter maior variação nas variáveis independentes e obter estimativas mais precisas

Multicolinearidade

- Considere o seguinte modelo:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

  - ✓ onde x_2 e x_3 são altamente correlacionados
  - ✓ $Var(\beta_2)$ e $Var(\beta_3)$ podem ser grandes, mas a correlação entre x_2 e x_3 não tem efeito direto sobre $Var(\beta_1)$
  - ✓ Se x_1 não estiver correlacionado com x_2 e x_3, o $R^2{}_1$= 0 e $Var(\beta_1)$ não afetado

Multicolinearidade

- Principais conclusões
  - ✓ Não causa viés
  - ✓ Não inclua controles altamente correlacionados com as variáveis independentes de interesse, se eles não forem necessários para identificação [por exemplo, $E\ (u \mid x) = 0$ sem eles]
    - Mas obviamente, se $E\ (u \mid x) \neq 0$ sem esses controles, você precisa deles!
    - Uma amostra maior ajudará a aumentar a precisão

- Teste para detectar multicolinearidade: vif (fator de inflação de variância)
  - ✓ Quanto maior o coeficiente, maior a colinearidade entre os regressores (x's)

Interações

- Às vezes, é útil para identificação do problema a ser estudado, adicionar interações entre x's
- Ex. - a teoria sugere que empresas com um alto valor de x_1 devem ser mais afetadas por alguma mudança em x_2
- O modelo será parecido com algo como…

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

Interações

- De acordo com este modelo, qual é o efeito em y quando se aumenta x_1, mantendo todo o resto igual?

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

- **Answer:**

$$\Delta y = (\beta_1 + \beta_3 x_2)\Delta x_1$$

$$\frac{dy}{dx_1} = \beta_1 + \beta_3 x_2$$

Interações

- Se $\beta 3 < 0$, como um aumento em x_2 afeta o efeito parcial de x_1 em y, mantendo todo o resto igual?

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

$$\frac{dy}{dx_1} = \beta_1 + \beta_3 x_2$$

Resposta: O aumento em y para uma dada mudança em x_1 será menor em níveis (não necessariamente em magnitude absoluta) para firmas com um x_2 maior

Interações

- Suponhamos que β1> 0 e β3 <0, qual é o sinal do efeito de um aumento em x_1 para a empresa média na população?

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

$$\frac{dy}{dx_1} = \beta_1 + \beta_3 x_2$$

Resposta: É o sinal de: $\left.\frac{dy}{dx_1}\right| x_2 = \overline{x}_2 = \beta_1 + \beta_3 \overline{x}_2$

Interações – um erro comum...

- Pesquisadores afirmam que "desde $\beta 1$> 0 e $\beta 3$ <0, um aumento em x_1 aumentaria o y para a empresa média, mas o aumento é menor para empresas com um alto x_2".

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u \qquad \left.\frac{dy}{dx_1}\right| x_2 = \overline{x}_2 = \beta_1 + \beta_3 \overline{x}_2$$

- Errado!!! O efeito médio de um aumento em $x1$ pode ser negativo se $x2$ for muito grande!
- $\beta 1$ captura apenas efeito parcial quando $x2$ = 0, o que pode até não fazer sentido se $x2$ nunca for 0!

Interações – um erro comum...

- Para melhorar a interpretação de $\beta 1$, você pode reparametrizar novamente o modelo, demeaning (diminuindo da sua média) cada variável no modelo, e estimar

$$\tilde{y} = \delta_0 + \delta_1 \tilde{x}_1 + \delta_2 \tilde{x}_2 + \delta_3 \tilde{x}_1 \tilde{x}_2 + u$$

- Onde:

$$\tilde{y} = y - \mu_y$$
$$\tilde{x}_1 = x_1 - \mu_{x_1}$$
$$\tilde{x}_2 = x_2 - \mu_{x_2}$$

Interações – um erro comum...

- Você pode então mostrar: $\Delta y = (\delta_1 + \delta_3 \tilde{x}_2) + \Delta x_1$

- E então: $$\left.\frac{dy}{dx_1}\right| x_2 = \mu_2 = \delta_1 + \delta_3 (x_2 - \mu_2)$$
$$\left.\frac{dy}{dx_1}\right| x_2 = \mu_2 = \delta_1$$

- Agora, o coeficiente do $x1$ demeaned pode ser interpretado como efeito de $x1$ para a empresa média!

Interações – A principal regra para levar...

- Se você quiser coeficientes em variáveis não-interadas para refletir o efeito dessa variável para a empresa "média", demean (diminua da média) todas as suas variáveis antes de executar a especificação

- Por que há tanta confusão sobre isso? Provavelmente por causa das variáveis indicadoras (binárias)...

Variáveis Indicadores (binárias)

- Vamos agora falar sobre variáveis indicadoras
  - ✓ Interpretação das variáveis indicadoras
  - ✓ Interpretação quando você os interage
  - ✓ Quando demean (descontar da sua média) é útil
  - ✓ Quando usar um indicador em vez de uma variável contínua pode fazer sentido

Variáveis Indicadores (binárias)

- Motivação
- As variáveis indicadoras, também conhecidas como variáveis binárias, são bastante populares nos dias de hoje
  - ✓ Ex. # 1 - Sexo do CEO (masculino, feminino)
  - ✓ Ex. # 2 - Status de emprego (empregado, desempregado)
- Também visto em muitas especificações diff-in-diff
  - ✓ Ex. # 1 - Tamanho da empresa (acima vs. abaixo da mediana)
  - ✓ Ex. # 2 - Pagamento do CEO (acima vs. abaixo da mediana)

Variáveis Indicadores (Dummy) – como funciona

- Codifique a informação usando a variável dummy

  - ✓ Ex. # 1: $Male_1 = \begin{cases} 1, \text{ se a pessoa é homem} \\ 0, \text{ caso contrário} \end{cases}$
  - ✓ Ex. # 2: $Large_1 = \begin{cases} 1, \text{ se Ln(assets) da firma i} > \text{mediana} \\ 0, \text{ caso contrário} \end{cases}$

- A escolha de 0 ou 1 é relevante apenas para interpretação
- No caso de que as definições, como em Ex. # 2 são arbitrárias, você geralmente vai precisar de uma verificação de comprovação para o critério que você estabeleceu.

Variáveis Indicadores (Dummy)

- Considere: $wage = \beta_0 + \delta_0 female + \beta_1 educ + u$

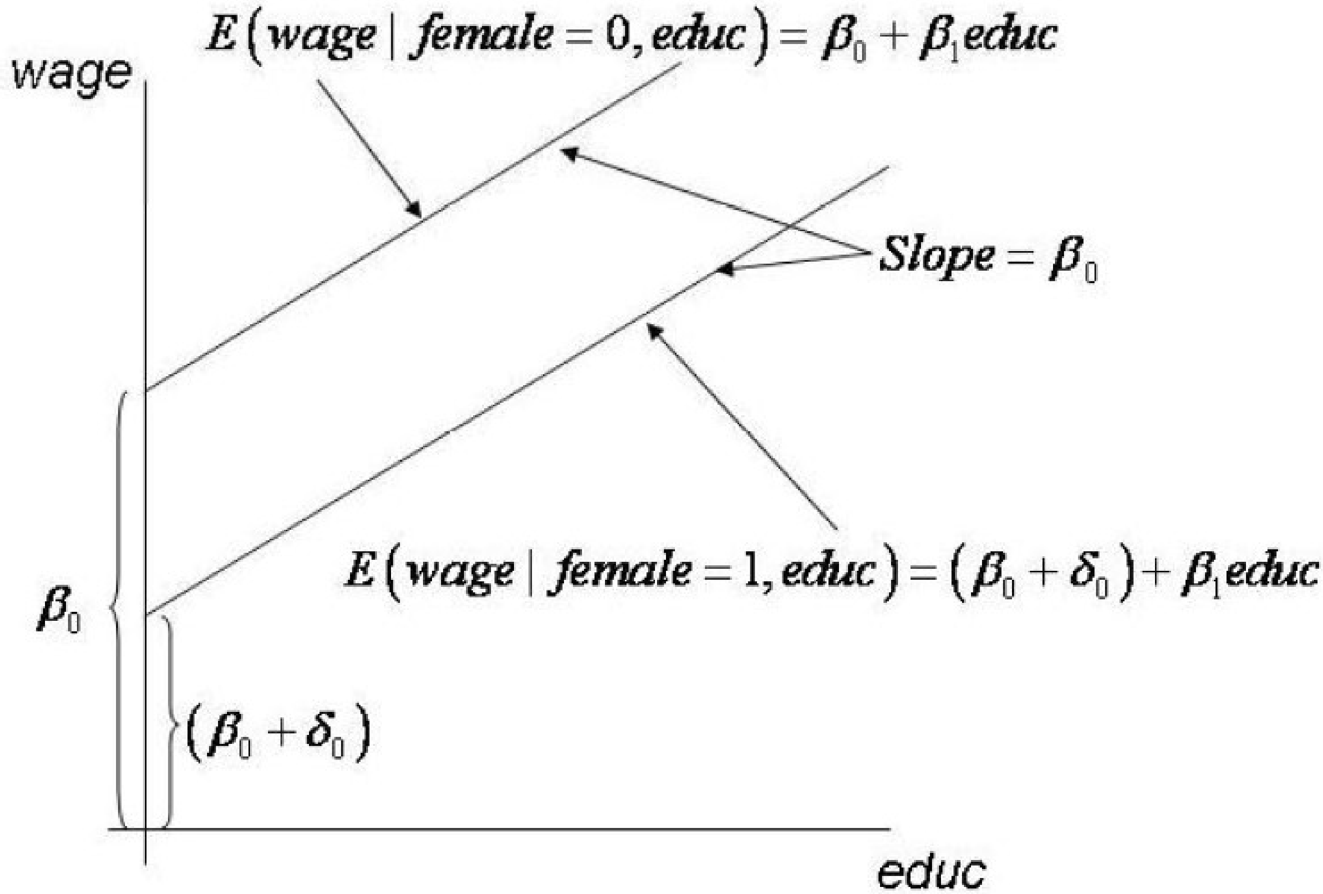
- δ_0 mede a diferença salarial entre homens e mulheres com o mesmo nível de escolaridade

  - $E(wage|female = o, educ) = \beta_0 + \beta_1 educ$
  - $E(wage|female = 1, educ) = \beta_0 + \delta_0 + \beta_1 educ$
  - Thus, $E(wage|f = 1, educ) - E(wage|f = 0, educ) = \boldsymbol{\delta_0}$

- O intercepto para $males = \beta_0, females = \beta_0 + \delta_0$

Variáveis Indicadores (Dummy)

- Quando δ_0 < 0, nós temos visualmente:



Variáveis Indicadores (Dummy)

- Suponha que estimamos o seguinte modelo salarial

$$Wage = -1.57 - 1.8female + 0.57educ + 0.03exp + 0.14tenure$$

- O Intercepto masculino é -1,57; isso é sem sentido, por quê?
- Como devemos interpretar o coeficiente de 1,8?

Resposta: As mulheres ganham $ 1,80 / hora menos que os homens com a mesma educação, experiência e ocupação

Variáveis Indicadores (Dummy) e Log

- Nada de novo; coeficiente no indicador tem% de interpretação. Considere o seguinte exemplo ...

$$\ln(price) = -1.35 + 0.17\ln(lotsize) + 0.71\ln(sqrft) \\ +0.03bdrms + 0.054colonial$$

- Mais uma vez, intercepto negativo sem sentido; todas as outras variáveis nunca são todas iguais a zero
- Interpretação = casa em estilo colonial custa <u>cerca de 5,4</u>% a mais do que casas “de outra forma semelhantes”

Múltiplas variáveis indicadoras(Dummy)

- Suponha que você queira saber quão mais baixos são os salários para as mulheres casadas e solteiras
- Agora temos 4 resultados possíveis
  - ✓ solteiros e masculinos
  - ✓ Casado e homem
  - ✓ solteira e feminina
  - ✓ Casada e feminino
- Para estimar, crie indicadores para três das variáveis e adicione-as à regressão

Múltiplas variáveis indicadoras(Dummy)

- Porque excluir uma dessas categorias??
- Temos que excluir um dos quatro porque eles são perfeitamente colineares com o intercepto, mas importa qual?

Resposta: Não, não importa realmente.

Apenas afeta a interpretação. As estimativas dos indicadores incluídos serão relativas ao indicador excluído
Por exemplo, se excluirmos "solteiro e masculino", estamos estimando uma mudança parcial no salário para as mulheres em relação à dos homens solteiros

Múltiplas variáveis indicadoras(Dummy)

- Porque excluir uma dessas categorias??

- **Nota: se você não excluir um, então os programas estatísticos como o Stata irão escolher um para você automaticamente. Para interpretar, você precisa descobrir qual deles foi descartado!**

Múltiplas variáveis indicadoras(Dummy)

- Considere os seguintes resultados de estimativa...

$$\ln(wage) = 0.3 + 0.21 marriedMale - .20 marriedFemale \\ -0.11 singleFemale + 0.08 education$$

- Eu omiti masculino solteiro; assim o intercepto será a média para homens solteiros
- E, pode interpretar outros coeficientes como...
  - ✓ Homens casados ganham ≈ 21% a mais do que homens solteiros, tudo o mais igual
  - ✓ As mulheres casadas ganham ≈ 20% menos do que os homens solteiros, tudo o mais igual

Interações com variáveis indicadoras(Dummy)

- Também poderíamos fazer regressão anterior usando interações entre dummies
  - ✓ Ex.: construir apenas dois indicadores, "feminino" e "casado", e estimar os seguintes

$$\ln(wage) = \beta_0 + \beta_1 female + \beta_2 married$$
$$+\beta_3(female \times married) + \beta_4 education$$

- **Como nossas estimativas e interpretações serão diferentes das estimativas anteriores?**

Interações com variáveis indicadoras(Dummy)

- Antes tínhamos:

$$\ln(wage) = 0.3 + 0.21 marriedMale - .20 marriedFemale \\ -0.11 singleFemale + 0.08 education$$

- Agora Temos:

$$\ln(wage) = 0.32 - 0.11 female + 0.21 married \\ -0.30(female \times married) + \cdots$$

- **Pergunta: Antes, mulheres casadas tinham salários que eram 0,20 mais baixos; quão mais baixos são os salários das mulheres casadas agora?**

Interações com variáveis indicadoras(Dummy)

- **Resposta: Será o mesmo!**

$$\ln(wage) = 0.32 - 0.11 female + 0.21 married$$
$$-0.30(female \times married) + \cdots$$

- Diferença para a mulher casada = −0,11 + 0,21 –0,30 = −0,20; exatamente o mesmo que antes

- Resumo = você pode fazer os indicadores de qualquer maneira; a interferência é não afetada

Interações com variáveis indicadoras(Dummy)

- Krueger (1993) encontrou...

$$\ln(wage) = \hat{\beta}_0 + 0.18compwork + 0.07comphome$$
$$+0.02(compwork \times comphome) + \cdots$$

- Categoria excluída = pessoas sem computador
- Como interpretamos essas estimativas?
  - ✓ Quão mais altos são os salários se possui computador no trabalho? ≈ 18%
  - ✓ Se possui computador em casa? ≈7%
  - ✓ Se tem computadores no trabalho e em casa? ≈18 + 7 + 2 = 27%

Interações com variáveis indicadoras(Dummy)

- Lembre-se, estas são apenas mudanças percentuais aproximadas ... Para obter uma mudança verdadeira, é preciso converter
  - ✓ % de variação nos salários por ter computadores em casa e no trabalho é dada por

$$100^{*}[\exp(0.18 + 0.07 + 0.02) - 1] = 31\%$$

Interações Dummy com variáveis contínuas

- Adicionando apenas dummies só irá mudar os interceptos para diferentes grupos

- Mas, se interagirmos com variáveis contínuas, podemos obter diferentes declives para diferentes grupos, bem

Interações Dummy com variáveis contínuas

- Considere o seguinte:

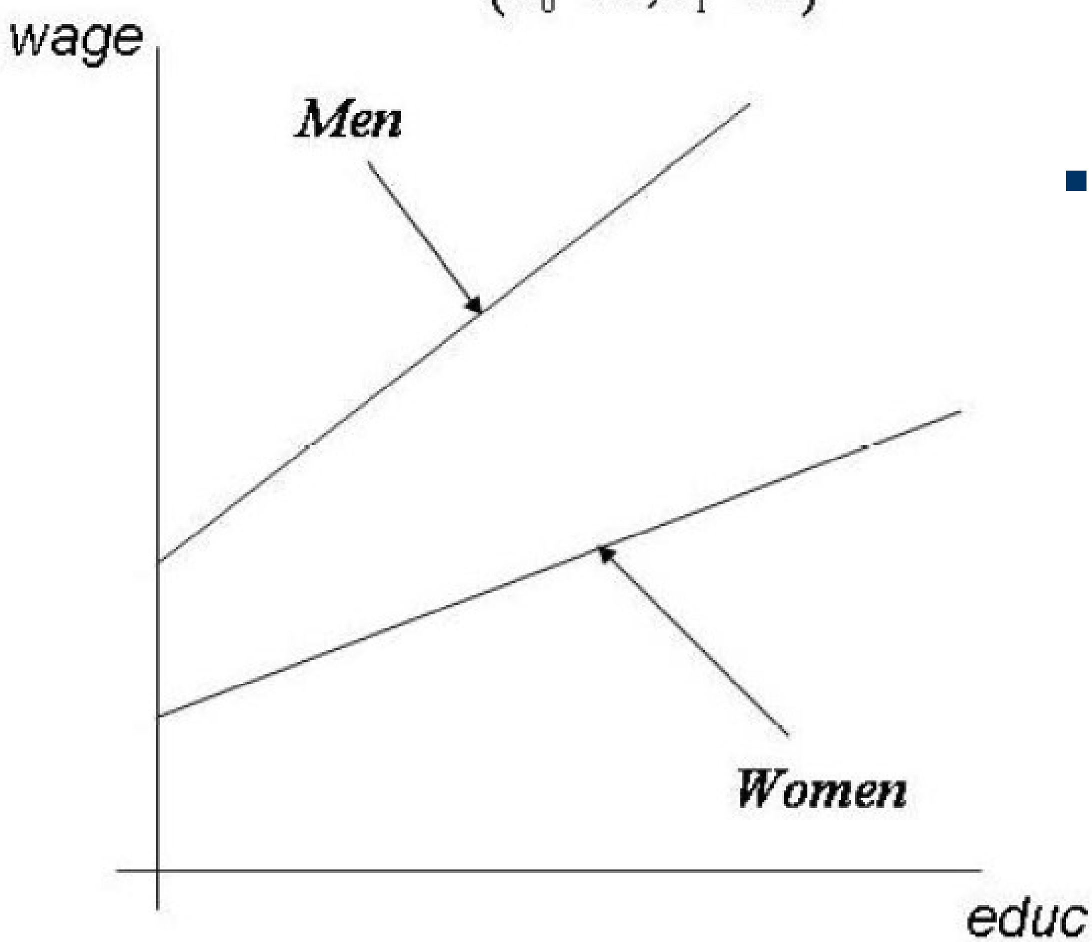
$$\ln(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 (female \times educ) + u$$

- ✓ Qual é o intercepto para os homens? β_0
- ✓ Qual é a inclinação para homens? β_1
- ✓ Qual é o intercepto para mulheres? $\beta_0 + \delta_0$
- ✓ Qual é a inclinação para as mulheres? $\beta_1 + \delta_1$

Interações Dummy com variáveis contínuas

$$\ln(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 (female \times educ) + u$$

$$(\delta_0 < 0, \delta_1 < 0)$$

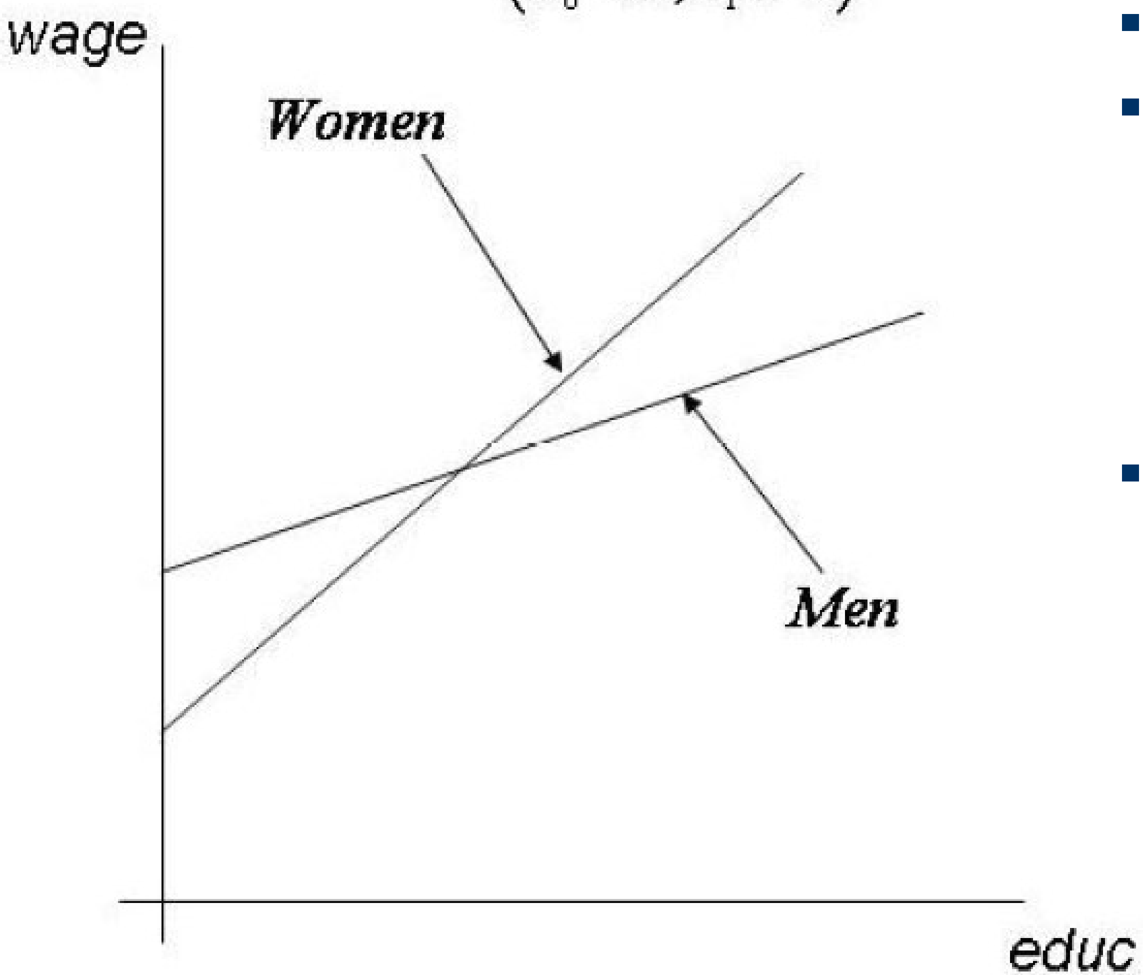


- Neste exemplo ...
  - ✓ As mulheres ganham salários mais baixos em todos os níveis de educação
  - ✓ Aumento médio por unidade de educação também é menor

Interações Dummy com variáveis contínuas

$$\ln(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 (female \times educ) + u$$

$$(\delta_0 < 0, \delta_1 > 0)$$



- Neste exemplo ...
- O salário é menor para as mulheres, mas apenas para os níveis mais baixos de educação, porque sua inclinação é maior
- **É justo concluir que as mulheres acabam ganhando salários mais altos com educação suficiente?**

Cuidado com inclinações diferentes

- Ponto de cruzamento (onde as mulheres ganham salários mais altos) pode ocorrer fora dos dados (ou seja, em níveis de educação que não existem)
- É Necessário resolver o ponto de intersecção antes de fazer essa afirmação sobre os dados

$$Women\text{: } \ln(wage) = \beta_0 + \delta_0 + (\beta_1 + \delta_1)educ + u$$

$$Men\text{: } \ln(wage) = \beta_0 + \beta_1 educ + u$$

- Serão iguais quando $\boldsymbol{educ}$=$\boldsymbol{\delta_0/\delta_1}$

Cuidado com inclinações diferentes

- Interpretação de termos não interagidos ao usar variáveis contínuas é complicado
- Por exemplo, considere as seguintes

$$\ln(wage) = 0.39 - 0.23female + 0.08educ - .01(female \times educ)$$

- A variação em wage da variável educ é de 8% para homens, 7% para mulheres
- Mas, no nível médio de escolaridade, quanto menos as mulheres ganham? [$-\mathbf{0023} - \mathbf{0.01} \times \boldsymbol{avg.educ}$]%

Cuidado com inclinações diferentes

- Novamente, a interpretação de variáveis não interagidas não é igual a efeito médio, a menos que você desconte a média (demean) das variáveis contínuas
- No exemplo anterior, estime o seguinte

$$\ln(wage) = \beta_0 + \delta_0 female + \beta_1(educ - \mu_{educ})$$
$$+\delta_1 female \times (educ - \mu_{educ})$$

- Agora, δ_0 nos diz quão menor o salário das mulheres no nível médio de escolaridade

Interações Dummy com variáveis contínuas

- Lembre! Como discutimos anteriormente, as inclinações não vão mudar por causa do desconto da média das variáveis
  - ✓ Somente o intercepto, β_0 e $\beta_0+\delta_0$ e seus erros padrões irão mudar

- Resumo = se você quiser interpretar dummies não-interagidos como o efeito dessas dummies na média das variáveis contínuas, você precisa descontar a média de todas as variáveis contínuas

Variáveis Ordinais

- Considere as classificações de ratings de crédito:
  - ✓ CR ∈ (AAA, AA, ..., C, D)

- Se quiser explicar a taxa de juros (IR) com classificações de rating, podemos converter CR em escala numérica, por exemplo, AAA = 1, AA = 2,… e estimar

$$IR_i = \beta_0 + \beta_1 CR_1 + \mu_1$$

- Mas, o que estamos implicitamente assumindo e como isso pode ser uma suposição problemática?

Variáveis Ordinais

- Resposta: Assumimos uma relação linear constante entre taxas de juros e CR
  - ✓ Mudar de AAA para AA produz a mesma alteração do movimento de BBB para BB

- Uma maneira melhor de se fazer é converter a variável ordinal em variáveis indicadoras

Variáveis Ordinais

- Faça CR_{AAA}=1 se CR=AAA; zero caso contrário
- Faça CR_{AA} = 1 se CR = AA, 0 caso contrário, etc.
- Em seguida, execute esta regressão

$$IR_i = \beta_0 + \beta_1 CR_{AAA} + \beta_2 CR_{AA} + \cdots + \beta_{m-1} CR_C + u_1$$

- Lembre-se de excluir um (por exemplo, "D")
- Isso permite que a mudança de IR de cada categoria de rating (em relação ao indicador excluído) seja de magnitude diferente!

Variáveis Ordinais – atenção aos dados

- Exemplo de dados $$IR_i = \beta_0 + \beta_1 CR_{AAA} + \beta_2 CR_{AA} + \cdots + \beta_{m-1} CR_C + u_1$$

Problema 1: mais categorias do que unidades de observação

	AAA	AA	A	BBB	BB	B	CCC	CC	C	D
Empresa A	1	0	0	0	0	0	0	0	0	0
Empresa B	0	0	1	0	0	0	0	0	0	0
Empresa C	0	1	0	0	0	0	0	0	0	0
Empresa D	0	0	0	1	0	0	0	0	0	0
Empresa E	0	0	0	0	1	0	0	0	0	0
Empresa F	0	1	0	0	0	0	0	0	0	0
Empresa G	0	0	1	0	0	0	0	0	0	0

Problema 2: Uma das colunas das dummies deve ser retirada

	AAA	AA	A	BBB	BB	B	CCC	CC	C	D
Empresa A	1	0	0	0	0	0	0	0	0	0
Empresa B	0	0	1	0	0	0	0	0	0	0
Empresa C	0	1	0	0	0	0	0	0	0	0
Empresa D	0	0	0	1	0	0	0	0	0	0
Empresa E	0	0	0	0	1	0	0	0	0	0
Empresa F	0	0	0	0	0	0	0	0	1	0
Empresa G	0	1	0	0	0	0	0	0	0	0
....	...	...	...	...	...	...	...	...	...	...
Empresa Y	0	0	0	0	0	0	0	1	0	0
Empresa X	0	0	0	0	0	1	0	0	0	0
Empresa Z	0	0	0	0	0	0	1	0	0	0

Apresentando Resultado das Regressões

- Tabela de resultados OLS geralmente deve mostrar o seguinte…
  - ✓ Variável dependente [claramente rotulada]
  - ✓ Variáveis independentes
  - ✓ Estimativa dos coeficientes, seus correspondentes erros padrão (ou t-stat) e estrelas indicando o nível de significância estatistica
  - ✓ R^2 ou R^2 ajustado ou ambos
  - ✓ Número de observações em cada regressão

Apresentando Resultado das Regressões

- Em corpo do artigo ou trabalho…
  - ✓ Concentre-se apenas na (s) variável (s) de interesse
  - ✓ Diga-nos seu sinal, magnitude, significado estatístico e econômico, interpretação, etc.

- Não desvie o foco para outros coeficientes, a menos que eles sejam "estranhos" (por exemplo, sinal errado, grande magnitude, etc)

Apresentando Resultado das Regressões

- E por último, mas não menos importante, não relate regressões em tabelas que você não vai discutir e / ou mencionar no corpo do texto

- Se não for importante o suficiente para mencionar no texto, não é importante o suficiente para estar em uma tabela

Pooled cross-section

- **Pooled OLS...**
- Consiste no “empilhamento” de cross-sections independentes coletadas em diferentes períodos de tempo.

- As cross-sections devem ser aleatórias (pelas mesmas razões de antes) e devem ser independentes, no sentido de que as observações coletadas em um período sejam independentes das coletadas nos demais períodos.

- Esta é uma diferença em relação a estrutura clássica de painel (em que as observações são escolhidas aleatoriamente no início do período e seguidas nos próximos períodos).

Pooled cross-section

- A principal razão para se construir um pool de cross-sections é aumentar o tamanho da amostra.

- Mas há um possível problema ao fazer isto: a distribuição conjunta de (y, x1,...,xk) pode mudar ao longo do tempo, ao passo que o modelo que estimamos mostra uma relação constante entre estas variáveis.

Pooled cross-section

- Tipos de dados

Cross Sectional

	Ano 2000					
	Variável Y	Variável X1	Variável X2	Variável X3	...	Variável Xn
Empresa A						
Empresa B						
Empresa C						
Empresa D						
Empresa E						
Empresa F						
...	...	...	...	...	...	...
Empresa Z						

Time Series

	Ano 2010	Ano 2009	Ano 2008	Ano 2007		Ano 20nn
	Variável Yt	Variável Yt-1	Variável Yt-2	Variável Yt-3	...	Variável Yt-n
Empresa A						

Time Series

	Empresa A					
	Variável Y	Variável X1	Variável X2	Variável X3	...	Variável Xn
Ano 2000						
Ano 2001						
Ano 2002						
Ano 2003						
Ano 2004						
Ano 2005						
...	...	...	...	...	...	...
Ano 20nn						

Pooled OLS

Panel Data

	Ano	Variável Y	Variável X1	Variável X2	...	Variável Xn
Empresa A	2000					
Empresa A	2001					
Empresa A	...					
Empresa A	20nn					
Empresa B	2000					
Empresa B	2001					
Empresa B	...					
Empresa B	20nn					
Empresa C	2000					
Empresa C	2001					
Empresa C	...					
Empresa C	20nn					
...	...	...	...	...	...	...
Empresa Z	2000					
Empresa Z	2001					
Empresa Z	...					
Empresa Z	20nn					

Pooled cross-section

- Para incorporar possíveis mudanças nestas relações podemos incorporar dummies de tempo ao modelo: valor 1 para observações coletadas em determinado ponto do tempo e valor zero para as demais.
  - ✓ Ex: D_2015 = 1; 0 otherwise
  - ✓ Ex: D_2016 = 1; 0 otherwise

- Podemos incorporar estas dummies tanto isoladas (para captar diferenças de intercepto no tempo) quanto interagindo com as variáveis explicativas (para captar mudanças dos interceptos entre os períodos).

Pooled cross-section

Pooled OLS

	Ano	Variável Y	D_Ano_2000	Variável X1	...	Variável Xn
Empresa A	2000		1			
Empresa A	2001		0			
Empresa A	...		0			
Empresa A	20nn		0			
Empresa B	2000		1			
Empresa B	2001		0			
Empresa B	...		0			
Empresa B	20nn		0			
Empresa C	2000		1			
Empresa C	2001		0			
Empresa C	...		0			
Empresa C	20nn		0			
...	...	...	...	...	...	...
Empresa Z	2000		1			
Empresa Z	2001		0			
Empresa Z	...		0			
Empresa Z	20nn		0			

Pooled cross-section

- Os métodos de estimação que abordamos até aqui (incluindo os testes de hipótese) podem ser aplicados a esta estrutura de dados e devem manter suas propriedades sob as mesmas hipóteses.

- Em geral, quando aplicamos os estimadores de MQO à estruturas de painel, damos o nome de Pooled MQO (POLS em inglês) ou **Pooled OLS**.

$$\boxed{\beta = (X^T X)^{-1} X^T Y}$$

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots \beta_n x_n + u}$$