

ACH3657

# Métodos Quantitativos para Avaliação de Políticas Públicas

Aula teórica 07  
Regressão Múltipla

Alexandre Ribeiro Leichsenring  
[alexandre.leichsenring@usp.br](mailto:alexandre.leichsenring@usp.br)



## 1 Regressão Múltipla

## Interpretação da Equação de Regressão com duas variáveis

Mais importante que os detalhes subjacentes à computação dos  $\hat{\beta}_j$  é a interpretação da equação estimada. No caso de duas variáveis independentes:

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

- O intercepto  $\hat{\beta}_0$  na equação é o valor previsto de  $y$  quando  $x_1 = 0$  e  $x_2 = 0$ .
- As estimativas  $\hat{\beta}_1$  e  $\hat{\beta}_2$  têm interpretações de efeito parcial, ou *ceteris paribus*.
- Da equação de regressão, temos:

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

► podemos obter a variação prevista em  $y$  dadas as variações em  $x_1$  e  $x_2$ .

- Em particular quando  $x_2$  é mantido fixo, então:

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1$$

► Ao incluir  $x_2$  no nosso modelo, obtemos um coeficiente de  $x_1$  com uma interpretação *ceteris paribus*. Essa é a razão de a análise de regressão múltipla ser tão útil.

- Analogamente, quando  $x_1$  é mantido fixo, então:

$$\Delta \hat{y} = \hat{\beta}_2 \Delta x_2$$

## Determinantes da nota média em curso superior nos EUA

As variáveis do arquivo **gpa1.RData** incluem a nota média em um curso superior (*nmgrad*), a nota média do ensino médio (*nmem*), e a nota do teste de avaliação de conhecimentos para ingresso em curso superior (*tac*) para uma amostra de 141 estudantes de uma grande Universidade dos Estados Unidos; tanto *nmgrad* como *nmem* estão baseados em uma escala de quatro pontos. Obtemos a seguinte equação de regressão MQO para estimar *nmgrad* a partir de *nmem* e *tac*:

$$\widehat{nmgrad} = 1,29 + 0,453 nmem + 0,0094 tac$$

Como interpretamos essa equação?

- $\beta_0 = 1,29$  é o valor previsto de *nmgrad* se tanto *nmem* como *tac* forem iguais a zero (como ninguém que frequenta um curso superior teve nota média no ensino médio igual a zero ou uma nota no teste de ingresso no curso superior igual a zero, o intercepto nessa equação não é, por si mesmo, significativo)
- Estimativas mais interessantes: coeficientes de *nmem* e *tac*. Há relação parcial positiva entre *nmgrad* e *nmem*: mantendo *tac* fixo, um ponto adicional em *nmem* está associado a 0,453 de um ponto em *nmgrad*

$$nm\hat{g}rad = 1,29 + 0,453 nmem + 0,0094 tac$$

- O sinal de *tac* implica que, mantendo *nmem* fixo, uma variação de 10 pontos na nota em *tac* afeta *nmgrad* em menos de um décimo de um ponto
  - ▶ Efeito pequeno, sugere que, uma vez considerada a *nmem*, a nota do *tac* não é um forte predictor de *nmgrad*
- Se ajustarmos um modelo de regressão simples relacionando somente *nmgrad* e *tac*, obtemos:

$$nm\hat{g}rad = 2,40 + 0,0271 tac$$

- ▶ O coeficiente de *tac* é quase três vezes maior que a estimativa no modelo múltiplo. No entanto, essa equação não nos permite comparar duas pessoas com o mesmo *nmem*; ela corresponde a um experimento diferente. Mais adiante, falaremos mais sobre as diferenças entre as regressões múltipla e simples.

## Interpretação da Equação de Regressão Múltipla ( $k$ variáveis)

- O caso com mais de duas variáveis independentes é similar.
- A reta de regressão de MQO é:

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

- Em termos de variações:

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

- O coeficiente de  $x_1$  mede a variação em  $\hat{y}$  devido a um aumento de uma unidade em  $x_1$ , tudo o mais constante. Isto é, mantendo  $x_2, x_3, \dots, x_k$  fixos:

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1$$

- Controlamos as variáveis  $x_2, x_3, \dots, x_k$  ao estimar o efeito de  $x_1$ , sobre  $\hat{y}$ .
- Os outros coeficientes têm interpretação similar.

## Exemplo: Equação do salário horário

Usando as 526 observações de trabalhadores do arquivo `wage1.RData`, incluímos *educ* (anos de educação formal), *exper* (anos de experiência no mercado de trabalho) e *perm* (anos com o empregador atual) na equação que explica  $\log(\text{salarioh})$ . A equação estimada é:

$$\log(\hat{\text{salarioh}}) = 0,284 + 0,092 \text{ educ} + 0,0041 \text{ exper} + 0,022 \text{ perm}.$$

- Os coeficientes têm uma interpretação de porcentagem
- Mas também têm uma interpretação *ceteris paribus*
- O coeficiente 0,092 significa que, mantendo *exper* e *perm* fixos, um ano a mais de educação formal aumenta o valor esperado de  $\log(\text{salarioh})$  em 0,092, o que se traduz em um aumento aproximado de 9,2% [ $100(0,092)$ ] em *salarioh*.
- Se considerarmos duas pessoas com os mesmos níveis de experiência e permanência no trabalho, o coeficiente de *educ* é a diferença proporcional no salário horário previsto quando seus níveis de educação diferem em um ano.
- Essa medida de retorno da educação mantém fixos ao menos dois importantes fatores de produtividade
- Saber se ela é uma boa estimativa do retorno *ceteris paribus* de mais um ano de educação formal requer que estudemos as propriedades estatísticas de MQO.

## Sobre “Manter Outros Fatores Fixos” na Regressão Múltipla

- Em anterior, observamos que o coeficiente **tac** mede a diferença prevista em **nmgrad**, mantendo **nmem** fixo.
- O poder da análise de regressão múltipla é que ela proporciona uma interpretação *ceteris paribus* mesmo que os dados não sejam coletados de uma maneira *ceteris paribus*
- Ao dar ao coeficiente de **tac** uma interpretação de efeito parcial, pode parecer que, realmente, saímos a campo e extraímos amostras compostas de pessoas com a mesma **nmem** e, possivelmente, com diferentes notas do **tac**.
- Isso não é verdade. Os dados são uma amostra aleatória de uma universidade grande: não há restrições colocadas sobre os valores amostrais de **nmem** ou **tac** na obtenção dos dados.
- De fato, raramente temos o luxo de manter certas variáveis fixas na obtenção de nossa amostra.
- Se pudéssemos coletar uma amostra de indivíduos com a mesma **nmem**, então poderíamos realizar uma análise de regressão simples relacionando **nmgrad** a **tac**.
- A regressão múltipla nos permite, efetivamente, simular essa situação sem restringir os valores de quaisquer variáveis independentes.
- O poder que a análise de regressão múltipla tem é que ela nos permite fazer, em ambientes não- experimentais, o que os cientistas naturais são capazes de fazer em um ambiente controlado de laboratório: manter outros fatores fixos.



## Variação de mais de uma Variável Independente Simultaneamente

Se queremos variar mais que uma variável independente ao mesmo tempo para encontrar o efeito resultante sobre a variável dependente

- Quando um indivíduo permanece na mesma empresa por mais um ano:
  - ▶ Ambos *exper* (experiência geral da força de trabalho) e *perm* aumentam em um ano.
- O efeito total (mantendo *educ* fixo) é

$$\begin{aligned}\Delta \log(\hat{\text{salario}}_h) &= 0,0041 \Delta \text{exper} + 0,022 \Delta \text{perm} \\ &= 0,0041 + 0,022 = 0,0261\end{aligned}$$

## Valores Estimados e Resíduos

- Para a observação  $i$ , o valor ajustado é:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_k x_{ki}$$

- No exemplo da nota no curso superior, temos:

$$nm\hat{grad} = 1,29 + 0,453 nm\text{mem} + 0,0094 tac$$

Se, por exemplo, temos:

$$\begin{aligned} nm\text{mem} &= 3,5 \\ tac &= 24 \end{aligned}$$

Então:

$$\begin{aligned} nm\hat{grad} &= 1,29 + 0,453(3,5) + 0,0094(24) \\ &= 3,101 \end{aligned}$$

- O resíduo da observação  $i$  é definido exatamente como no caso da regressão simples:

$$\hat{u}_i = y_i - \hat{y}_i$$

- Há um resíduo para cada observação
- Os valores estimados de MQO e os resíduos têm algumas propriedades importantes que são extensões imediatas do caso da variável única:
  - 1 A média amostral dos resíduos é zero.
  - 2 A covariância amostral entre cada variável independente e os resíduos de MQO é zero. Conseqüentemente, a covariância amostral entre os valores estimados de MQO e os resíduos de MQO é zero
  - 3 O ponto  $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, \bar{y})$  está sempre sobre a reta de regressão

## Comparação das Estimativas das Regressões Simples e Múltipla

- Equação de regressão simples de  $y$  sobre  $x_1$ :

$$\hat{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

- Equação de regressão múltipla:

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

- Relação entre  $\tilde{\beta}_1$  e  $\hat{\beta}_1$ :

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1,$$

em que  $\tilde{\delta}_1$ , é o coeficiente de inclinação da regressão simples de  $x_{2i}$  sobre  $x_{1i}$ ,  $i = 1, \dots, n$ , isto é, do seguinte modelo:

$$x_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$

►  $\hat{\beta}_2 \tilde{\delta}_1$  pode causar confusão!

- Portanto, há duas situações em que  $\tilde{\beta}_1$  e  $\hat{\beta}_1$  são iguais:

- 1 O efeito parcial de  $x_2$  sobre  $y$  é zero na amostra, isto é,  $\hat{\beta}_2 = 0$ .
- 2  $x_1$  e  $x_2$  são não-correlacionados na amostra, isto é,  $\tilde{\delta}_1 = 0$ .



## De volta ao exemplo das notas no ensino superior

Ainda que as estimativas das regressões múltipla e simples quase nunca sejam idênticas, podemos usar a fórmula anterior para caracterizar o motivo pelo qual elas deveriam ser diferentes ou similares.

No exemplo das notas no ensino superior, temos:

$$nm\hat{g}rad = 1,286 + 0,453 nmem + 0,009tac$$

$$nm\hat{g}rad = 1,415 + 0,482 nmem$$

- ▶ A estimativa dos coeficientes para  $nmem$  nos dois casos é parecida. A correlação linear entre  $nmem$  e  $tac$  é 0,346 (não trivial). Por quê os coeficientes estimados para  $nmem$  na regressão múltipla e na simples são próximos?

## Participação nos Planos de Pensão 401 (k)

Usando os dados do arquivo em 401k.RData para estimar o efeito da taxa de contribuição ( $taxcont$ ) sobre a taxa de participação ( $taxap$ ) dos trabalhadores nos planos de pensão de contribuição definida existentes nos Estados Unidos.

Há 1.534 planos no banco de dados, a  $taxap$  média é 87,36, a  $taxcont$  média é 0,732 e a idade média é 13,2.

A equação de regressão estimada é:

$$\widehat{taxap} = 80,12 + 5,52 taxcont + 0,243 idade$$

- $taxcont$  e  $idade$  têm os efeitos esperados. O que aconteceria se não controlássemos a variável  $idade$ ?
- O efeito estimado de  $idade$  é não trivial, então poderíamos esperar uma alteração no efeito estimado de  $taxcont$  se  $idade$  fosse excluída da regressão.
- Entretanto, a regressão simples de  $taxap$  sobre  $taxcont$  produz:

$$\widehat{taxap} = 83,08 + 5,86 taxcont$$

- A estimativa de regressão simples do efeito de  $taxcont$  sobre  $taxap$  é diferente da estimativa de regressão múltipla, mas a diferença não é muito grande.
- Isso pode ser explicado pelo fato de a correlação amostral entre  $taxcont$  e  $idade$  ser somente de 0,12.

