ACH3657

Métodos Quantitativos para Avaliação de Políticas Públicas

Aula teórica 06

Regressão linear múltipla

Alexandre Ribeiro Leichsenring

alexandre.leichsenring@usp.br

Organização

1. Regressão Múltipla
   - O modelo com duas variáveis explicativas
   - O modelo com k variáveis explicativas

Regressão Múltipla

- A análise de regressão simples é usada para explicar uma variável dependente y como função de uma única variável independente y.
- Em trabalhos empíricos é muito difícil obter conclusões *ceteris paribus* sobre como x afeta y: a hipótese fundamental $Cov(u, x) = 0$ (todos os outros fatores que afetam y são não-correlacionados com x) é frequentemente irreal.
- A análise de regressão múltipla é mais receptiva à análise *ceteris paribus*, pois ela nos permite controlar explicitamente muitos outros fatores que, de maneira simultânea, afetam a variável dependente.
- Isso é importante para avaliar efeitos da política governamental quando nos baseamos em dados não-experimentais. Como os modelos de regressão múltipla podem acomodar muitas variáveis explicativas que podem estar correlacionadas, podemos esperar inferir causalidade nos casos em que a análise de regressão simples seria enganosa.
- Se adicionarmos ao nosso modelo mais fatores que são úteis para explicar y, então mais da variação de y poderá ser explicada. A análise de regressão múltipla pode ser usada para construir modelos melhores para prever a variável dependente.

- No modelo de regressão simples, somente a função de uma variável explicativa pode aparecer na equação. O modelo de regressão múltipla permite muito mais flexibilidade.
- O modelo de regressão múltipla é um veículo extensamente usado da análise empírica em economia e em ciências sociais.

O Modelo com duas variáveis explicativas

Quando há duas variáveis explicativas x_1 e x_2, o modelo de regressão usual é:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \ , \tag{1}$$

onde

y : resposta

x_1 e x_2 : valores das variáveis explicativas x_1 e x_2

β_0, β_1 e β_2 : parâmetros do modelo

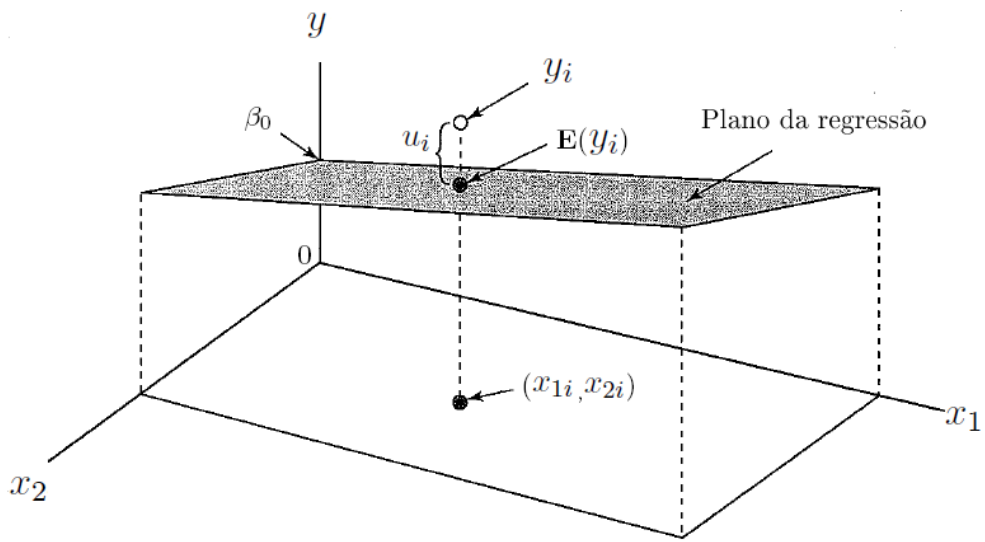
u : termo de erro

Equação de regressão populacional para o modelo

$$\mathbf{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{2}$$

Observação

- essa equação não depende de fatores aleatórios
- no caso simples (1 variável) a equação de regressão é uma reta
- equação com duas variáveis explicativas fornece um plano.



Exemplo

Variação simples da equação do salário para obter o efeito da educação sobre o salário-hora:

$$salarioh = \beta_0 + \beta_1\ educ + \beta_2 exper + u \tag{3}$$

onde

- salarioh é o salário por hora
- educ é a escolaridade, em anos de estudo
- exper representa anos de experiência no mercado de trabalho

- **salárioh** é determinado por duas variáveis explicativas (educação e experiência), e por outros fatores não-observados, contidos em u
- Estamos interessados no efeito de **educ** (β_2) sobre **salárioh**, mantendo fixos todos os outros fatores que afetam **salárioh**
- A equação 3 remove, efetivamente, **exper** do termo de erro e a coloca explicitamente na equação
- Como **exper** aparece na equação, seu coeficiente, β_2, mede o efeito *ceteris paribus* de **exper** sobre **salárioh**

Exemplo

Considere o problema de explicar o efeito do gasto público por estudante (gasto) sobre a nota média padronizada (notmed) do ensino médio. Suponha que a nota média dependa do gasto público, da renda familiar média (rendfam) e de outros fatores não-observáveis:

$$notmed = \beta_0 + \beta_1\ gasto + \beta_2\ rendfam + u. \tag{4}$$

- Para o propósito de análise da política pública, o coeficiente de interesse é β_1, o efeito *ceteris paribus* de gasto sobre notmed.
- Ao incluir rendfam explicitamente no modelo, somos capazes de controlar seu efeito sobre notmed.
- Importante, pois a renda familiar média tende a estar correlacionada com o gasto público por estudante

Nesses dois exemplos similares, mostramos que outros fatores observáveis, além da variável de interesse primordial podem ser incluídos em um modelo de regressão.

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

Exemplo

Suponha que o consumo da família ($cons$) é uma função quadrática da renda familiar ($rend$):

$$cons = \beta_0 + \beta_1 rend + \beta_2 rend^2 + u \tag{5}$$

em que u contém outros fatores que afetam o consumo.

- Nesse exemplo, o consumo depende somente de um fator observado: a renda
- Pareceria que ele pode ser tratado dentro do arcabouço da regressão simples
- No entanto, esse modelo está fora do padrão da regressão simples, porque ele contém duas funções da renda, $rend$ e $rend^2$
  - E, portanto, três parâmetros: β_0, β_1 e β_2
- A função consumo é facilmente escrita como modelo de regressão com duas variáveis, e fazendo:

$$\begin{aligned} x_1 &= rend \\ x_2 &= rend^2 \end{aligned}$$

- Há, entretanto, uma diferença importante em como interpretar os parâmetros.
- Na equação (4), β_1 é o efeito *ceteris paribus* de $educ$ sobre $salarioh$.
- O parâmetro β_1, não tem tal interpretação em (5)
- Em outras palavras, não faz sentido medir o efeito de $rend$ sobre $cons$ mantendo, ao mesmo tempo, $rend^2$ fixo!
- Em vez disso, a variação no consumo com respeito à variação na renda – a *propensão marginal a consumir* – é aproximada por:

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

- Em outras palavras, o efeito marginal da renda sobre o consumo depende tanto de β_2 como de β_1 e do nível de renda.

Suposições sobre o modelo

No modelo com duas variáveis independentes, a hipótese fundamental sobre como u está relacionado a x_1 e x_2 é:

$$\mathbf{E}(u|x_1, x_2) = 0 \tag{6}$$

- A interpretação da suposição (6) é similar à interpretação de RLS.3 na análise de regressão simples.
- Ela significa que, para qualquer valor de x_1, e x_2 na população, o fator não-observável médio é igual a zero.
- A parte importante da hipótese é que o valor esperado de u é o mesmo para todas as combinações de x_1 e x_2.

Como podemos interpretar a hipótese de média condicional zero no exemplo anterior?

- Na equação (3), a hipótese é $E(u|educ,\ exper) = 0$.
- Isso implica que outros fatores que afetam salárioh não estão, em média, relacionados a educ e exper.
- Portanto, se entendermos que aptidão inata é parte de u, então precisaremos que os níveis médios de aptidão sejam os mesmos em todas as combinações de educação e experiência na população que trabalha.
- Essa é a questão que precisamos fazer a fim de determinar se o método de produz estimadores não-viesados.

O exemplo que mede o desempenho dos estudantes (equação 4) é similar à equação do salário.

- A hipótese de média condicional zero é $E(u|gasto,\ rendfam) = 0$, o que significa que os outros fatores que afetam as notas - características das escolas e dos estudantes - são, em média, não-relacionados aos gastos públicos por estudante e à renda familiar média.

Exercício

Um modelo simples para explicar as taxas de homicídio nas cidades (taxahom) em termos da probabilidade de condenação (prcond) e da duração média da sentença (sentmed) é:

$$taxhom = \beta_0 + \beta_1\, prcond + \beta_2 sentmed + u$$

- Que fatores estão contidos em u?
- Você entende ser provável que a hipótese de média condicional zero se mantenha?

Modelo com k Variáveis Independentes

- A regressão múltipla permite que muitos fatores observados afetem y.
- No exemplo do salário, poderíamos também incluir semanas de treinamento de trabalho, anos de permanência com o empregador atual, medidas de aptidão e mesmo variáveis demográficas, como o número de irmãos ou a educação da mãe.
- No exemplo do gasto público por estudante, poderiam ser incluídos variáveis adicionais que medissem a qualidade dos professores e o tamanho das escolas.

Equação de regressão múltipla

O modelo de regressão linear múltipla geral (também chamado modelo de regressão múltipla) pode ser escrito, na população, como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \tag{7}$$

- A variável u é o termo de erro. Ele contém outros fatores, além de $x_1, x_2, \ldots, x_k$ que afetam y.
- Não importa quantas variáveis explicativas incluímos em nosso modelo, sempre haverá fatores que não podemos incluir, e eles estão contidos em u.

- A hipótese essencial para o modelo de regressão múltipla geral em termos de uma esperança condicional é:

$$\mathbf{E}(u|x_1, x_2, \ldots, x_k) = 0 \tag{8}$$

- a equação (8) requer que todos os fatores no termo erro não-observado sejam não-correlacionados com as variáveis explicativas.
- Ela também significa que consideramos corretamente a relação funcional entre as variáveis explicada e as explicativas.
- Qualquer problema que faça com que u seja correlacionado com qualquer variável independente fazem com que (8) não seja válida.

Estimação dos Parâmetros

Para encontrar os estimadores de Mínimos Quadrados para o modelo com 2 variáveis explicativas:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \tag{9}$$

é a função de regressão amostral.

Os erros u_i observados são dados por

$$\begin{aligned} u_i &= y_i - \hat{y}_i && (10) \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} && (11) \end{aligned}$$

O método de mínimos quadrados consiste em encontrar estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ que minimizem a soma dos quadrados dos resíduos. Ou seja, precisamos minimizar

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 \tag{12}$$

Devemos encontrar as derivadas parciais com relação a $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$, e igualar as 3 derivadas a zero. Fazendo

$$f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2,$$

queremos encontrar a solução do sistema:

$$\begin{aligned}
\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_0} &= 0 \\
\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} &= 0 \\
\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} &= 0
\end{aligned}$$

Resolvendo o sistema resultante, obtemos

Estimadores MQO para β_0, β_1 e β_2

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\
\hat{\beta}_1 &= \frac{\left(\sum_i \tilde{y}_i \tilde{x}_{1i}\right)\left(\sum_i \tilde{x}_{2i}^2\right) - \left(\sum_i \tilde{y}_i \tilde{x}_{2i}\right)\left(\sum_i \tilde{x}_{1i} \tilde{x}_{2i}\right)}{\left(\sum_i \tilde{x}_{1i}^2\right)\left(\sum_i \tilde{x}_{2i}^2\right) - \left(\sum_i \tilde{x}_{1i} \tilde{x}_{2i}\right)^2} \\
\hat{\beta}_2 &= \frac{\left(\sum_i \tilde{y}_i \tilde{x}_{2i}\right)\left(\sum_i \tilde{x}_{1i}^2\right) - \left(\sum_i \tilde{y}_i \tilde{x}_{1i}\right)\left(\sum_i \tilde{x}_{1i} \tilde{x}_{2i}\right)}{\left(\sum_i \tilde{x}_{1i}^2\right)\left(\sum_i \tilde{x}_{2i}^2\right) - \left(\sum_i \tilde{x}_{1i} \tilde{x}_{2i}\right)^2}
\end{aligned}
$$

Onde

$$
\begin{aligned}
\tilde{y}_i &= y_i - \bar{y} \\
\tilde{x}_1 &= x_{1i} - \bar{x}_1 \\
\tilde{x}_2 &= x_{2i} - \bar{x}_2
\end{aligned}
$$

Estimador Mínimos Quadrados para σ^2

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{\sum_{i=1}^{n} u_i^2}{n-3} \\
&= \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-3} \\
&= \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2}{n-3}
\end{aligned}
$$

Interpretação da Equação de Regressão

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:

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