Mínimos Quadrados Ordinários (MQO)

Vinicius Limeira

Abstract

Uma introdução ao método de mínimos quadrados ordinários.

Contents

1 Introdução

Modelos de regressão linear estimados por meio do método de mínimos quadrados ordinários constituem a pedra angular da econometria. Dessa forma, para um curso que se propõe introdutório, devemos partir daí. Com esse intuito, vamos nos basear em Verbeek (2012) de forma a ressaltar o método de MQO em termos algébricos.

Suponha, com efeito, que você tenha uma amostra com N observações de salários individuais e algumas características de fundo. Nosso objetivo principal é relacionar os salários dessa amostra a essas características, conforme vimos na seção anterior. Em termos um pouco mais formais, vamos chamar os salários de y e as $K-1$ características por $x_2, ..., x_k$. Nesses termos, podemos nos perguntar qual a combinação linear de $x_2, ..., x_k$ e uma constante dá uma boa aproximação de y. Para responder essa pergunta, primeiro, considere uma combinação linear arbitrária, incluindo a constante, que pode ser escrita como

$$\tilde{\beta}_1 + \tilde{\beta}_2 x_2 + ... + \tilde{\beta}_k x_k \tag{1}$$

Onde, $\tilde{\beta}_1, ..., \tilde{\beta}_k$ são constantes a serem escolhidas, de modo que a diferença entre y_i e essa combinação linear pode ser dada por:[1]

$$y_i - [\tilde{\beta}_1 + \tilde{\beta}_2 x_{i2} + ... + \tilde{\beta}_k x_{ik}] \tag{2}$$

Ou,

$$y_i - x_i^{'}\tilde{\beta} \tag{3}$$

Nesses termos, devemos escolher valores para $\tilde{\beta}_1, ..., \tilde{\beta}_k$ de modo que a diferença dada por 3 seja a menor possível.[2] Isto é, devemos determinar $\tilde{\beta}$ de modo a minimizar a seguinte função objetivo

$$S(\tilde{\beta}) \equiv \sum_{i=1}^{N} (y_i - x_i^{'}\tilde{\beta})^2 \tag{4}$$

Em resumo, minimizar a soma dos erros da aproximação ao quadrado. Esse método é, precisamente, o que chamamos de **mínimos quadrados ordinários**. Para resolver esse problema de minimização, consideramos a condição de primeira ordem obtida pela derivação de $S(\tilde{\beta})$ com respeito ao vetor $\tilde{\beta}$. Isto é,

$$-2 \sum_{i=1}^{N} x_i (y_i - x_i^{'}\tilde{\beta}) = 0 \tag{5}$$

Ou

$$(\sum_{i=1}^{N} x_i x_i^{'})\tilde{\beta} = \sum_{i=1}^{N} x_i y_i \tag{6}$$

A solução, assim, para o problema de minimização pode ser dado por:

$$b = (\sum_{i=1}^{N} x_i x_i^{'})^{-1} \sum_{i=1}^{N} x_i y_i \tag{7}$$

[1] Observe que indexamos as observações por i, dado $i = 1, ..., N$.

[2] Nessa apostila, vamos seguir a convenção de que que x_i é o vetor coluna $\begin{pmatrix} 1 \\ x_{i2} \\ ... \\ x_{ik} \end{pmatrix}$ e $x_i^{'}$ é o transposto de x_i, isto é, um vetor-linha.

Tomando a condição de segunda ordem, é fácil verificar que b, de fato, corresponde ao mínimo de 4. A combinação linear resultante de x_i é então dada por

$$\widehat{y}_i = x_i^{'} b \tag{8}$$

que é a melhor aproximação linear de y dado $x_2, ..., x_k$ e uma constante. Ou seja, a soma das diferenças ao quadrado entre os valores observados de y_i e os valores *estimados* $\widehat{y}_i$ será mínima para a solução b.

Na derivação da aproximação linear, nós não utilizamos nenhum conceito estatístico ou econômico, diga-se. O que fizemos foi apenas uma *manipulação algébrica*, não relacionada a forma como os dados foram gerados. Ou seja, dado um conjunto de variáveis, podemos sempre determinar a melhor aproximação linear de uma variável usando as demais variáveis.[3]

Definindo, assim, um **resíduo** e_i como a diferença entre os valores observados e aqueles aproximados, $e_i = y_i - \widehat{y} = y_i - X_i b$, nós podemos decompor os valores observados como

$$y_i = \widehat{y} + e_i = x_i^{'} b + e_i \tag{9}$$

Isso nos permite escrever o valor mínimo para a função objetivo 4 como

$$S(b) = \sum_{i=1}^{N} e_i^2 \tag{10}$$

O que pode ser descrito como o *somatório dos resíduos ao quadrado*. Com efeito, pode ser mostrado que o valor aproximado $X_i b$ e o resíduo e_i satisfazem certas propriedades por construção. Por exemplo, se nós reescrevermos 5, substituindo a solução b, nós obtemos

$$\sum_{i=1}^{N} x_i (y_i - x_i^{'} \tilde{\beta}) = \sum_{i=1}^{N} x_i e_i = 0 \tag{11}$$

Isso significa que o vetor $\varepsilon = (e_1, ..., e_N)$ é ortogonal para cada vetor de observações nas $x - variveis$.[4]

2 Pacotes utilizados nessa seção

```
library(wooldridge)
library(ggplot2)
library(stargazer)
library(foreign)
```

[3] A única hipótese que fizemos aqui é que a matriz $KxK \sum_{i=1}^{N} x_i x_i^{'}$ é inversível. Isso é chamado, em geral, de **premissa de não existência de multicolinearidade**.

[4] Dois vetores x e y são ditos ortogonais se $x'y = 0$, assim $\sum_i x_i y_i = 0$.

3 Regressão linear simples

Para o caso onde $K = 2$, nós temos apenas um regressor e uma constante. Nesse caso, as observações (y_i, x_i) podem ser desenhadas em um gráfico de duas dimensões. A melhor aproximação de y por x e uma constante é obtido pela minimização da soma dos resíduos ao quadrado, o que no caso de duas variáveis é igual à distância vertical entre uma observação e o valor ajustado. Todos os valores ajustados (*fitted values*) estão em uma linha reta, chamada de **reta de regressão**.

Dado que a matrix 2×2 pode ser invertida analiticamente, nós podemos derivar soluções para b_1 e b_2 nesse caso especial a partir da expressão geral para b dada por 7. De forma equivalente, nós podemos minimizar a soma dos resíduos ao quatro com respeito aos parâmetros desconhecidos de forma direta. Assim, teremos:

$$S(\tilde{\beta}_1, \tilde{\beta}_2) = \sum_{i=1}^{N} (y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i)^2 \tag{12}$$

Os elementos básicos na derivação das soluções do MQO serão as condições de primeira ordem

$$\frac{\partial S(\tilde{\beta}_1, \tilde{\beta}_2)}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^{N} (y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i) = 0 \tag{13}$$

$$\frac{\partial S(\tilde{\beta}_1, \tilde{\beta}_2)}{\partial \tilde{\beta}_2} = -2 \sum_{i=1}^{N} x_i (y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i) = 0 \tag{14}$$

3.1 Salários versus Experiência no emprego

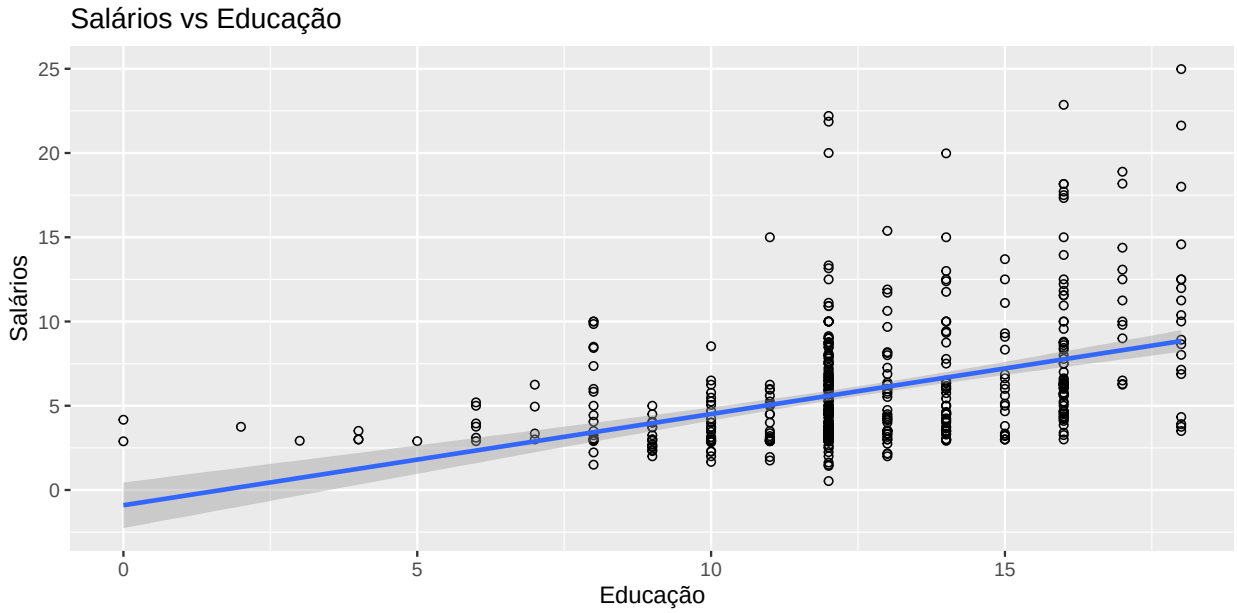
Para ilustrar a **reta de regressão**, vamos pegar aquele mesmo conjunto de dados da seção anterior e tomar o modelo representado por

$$\text{Salário}_i = \beta_0 + \text{Experiência}_i + \varepsilon_i$$

```
data(package = "wooldridge")
data(wage1)
```

Uma vez carregado esse conjunto de dados, podemos plotar a reta de regressão entre `salários` e `experiência no emprego` como abaixo.

```
ggplot2::ggplot(wage1, aes(x=educ, y=wage))+
  geom_point(shape=1)+
  geom_smooth(method = lm)+
  xlab('Educação')+
  ylab('Salários')+
  ggtitle('Salários vs Educação')
```



Observe que, de fato, há uma correlação positiva entre Educação e Salários. Será que nossa hipótese está no caminho correto para não ser rejeitado pela evidência empírica? Para que consigamos encontrar uma evidência de que a Educação de fato tem efeito positivo sobre os Salários, precisaremos estimar o parâmetro β_1.

O trabalho, entretanto, não acaba aí. É preciso verificar se o valor obtido para o parâmetro é estatisticamente significativo. Isso será feito através de testes de hipóteses. Por fim, podemos usar nosso modelo para fins de previsão ou mesmo para implementação de alguma política pública

3.2 Modelo Para a Relação anterior

Para ilustrar a reta de regressão, vamos pegar aquele mesmo conjunto de dados da seção anterior e tomar o modelo representado por:

$$\text{Salário}_i = \beta_0 + \text{Experiência}_i + \varepsilon_i$$

```
wage1 <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")


reta = lm(wage~exper, data = wage1)
```

A linha é o que chamamos de reta de regressão, enquanto a distância entre as observações e essa reta são justamente os resíduos. A expressão da reta de regressão é dada, então, por $Salrios = 5.37 + 0.03Experincia$.

3.3 Aproximação Linear

Vamos ilustrar, agora, o que queremos dizer com *aproximação linear*. Considere o conjunto de dados que importamos acima e regrida os valores dos `salários individuais` contra uma **variável dummy** que associa 1 para o gênero feminimo e 0 para o gênero masculino. Isso é feito abaixo.

```
reg <- lm(wage1$wage ~ wage1$female)
stargazer(reg, title = 'Salários contra dummy de gênero',
          header = FALSE)
```

Table 1: Salários contra dummy de gênero

	Dependent variable:
	wage
female	-2.512^{***}
	(0.303)
Constant	7.099^{***}
	(0.210)
Observations	526
R^2	0.116
Adjusted R^2	0.114
Residual Std. Error	3.476 (df = 524)
F Statistic	68.537^{***} (df = 1; 524)
Note:	*p<0.1; **p<0.05; ***p<0.01

Ao considerar apenas uma dummy de gênero, podemos dizer que a melhor aproximação para ossalários das mulheres seria 4.59. Enquanto para os homens, seria 7.1.

3.4 O modelo de Regressão Linear

Usualmente, nós podemos querer mais do que simplesmente encontrar a melhor aproximação linear de uma variável dado um conjunto de outras variáveis. Às vezes podemos estar interessados em relações mais gerais do que aquelas proporcionadas pela amostra disponível. Às vezes podemos estar interessados em verificar o efeito de uma mudança em alguma das nossas variáveis. Em outras palavras, podemos querer fazer afirmações sobre coisas que ainda não observamos efetivamente. Para isso, nós precisamos que as relações encontradas reflitam mais do que apenas uma coincidência histórica e sim uma relação fundamental.[5]

Para fazer isso, devemos assumir que existe uma relação geral que é válida para todas as possíveis observações de uma bem definida população. Restringindo para o caso de relações lineares, nós especificamos um **modelo estatístico** do tipo:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \varepsilon_i \tag{15}$$

[5]Para maiores detalhes sobre essa seção, ver Verbeek (2012).

ou

$$y_i = x_i^{'}\beta + \varepsilon_i \tag{16}$$

onde y_i e x_i são variáveis observáveis e ε_i é não observável e refere-se a um **termo de erro**. Os elementos em β são **parâmetros populacionais** não conhecidos. A igualdade em 16 é supostamente válida para qualquer possível observação, ainda que tenhamos acesso a uma amostra com N observações. Nós, aliás, consideráremos essa amostra como uma realização de todas as potenciais amostras de tamanho N que poderíamos tomar da mesma população. Nesse caso, nós podemos considerar y_i e ε_i como variáveis aleatórias.[6] Nós podemos, a propósito, representar a equação 16 em notação matricial, como abaixo:

$$y = X\beta + \varepsilon \tag{17}$$

onde X é uma matriz $N \times K$ e β é uma matriz $K \times 1$. O **processo de amostragem**, por suposto, descreve como a amostra é tomada da população e, como resultado, determina a aleatoriedade dessa amostra. Numa primeira vista, as x_i variáveis são consideradas fixas ou *não estocásticas*, o que significa que toda nova amostra terá a mesma matriz X. Nesse caso, consideramos x_i como sendo **determinístíca**. Desse modo, uma nova amostra apenas implica em novos valores para ε_i e, portanto, para y_i.

Chamamos atenção, a propósito, que o único caso relevante onde $x_i s$ são realmente determinísticas é em laboratório, onde o pesquisador pode determinar as condições de um dado experimento (ex. temperatura ou pressão). Em economia nós teremos que lidar, de maneira geral, com dados não experimentais. Apesar disso, é conveniente e em casos particulares apropriado a determinado contexto econômico agir como se as x_i variáveis fossem determinísticas. Nesse caso, teremos que fazer algumas suposições sobre a distribuição de ε_i. Uma conveniente corresponde à **amostragem aleatória** onde cada erro ε_i é um desenho aleatório tomado da distribuição da população, independente de outros termos de erro.

Em um segundo olhar, uma nova amostra implica em novos valores tanto para x_i quanto para ε_i, assim a cada tempo um novo conjunto com N observações para (y_i, x_i) é desenhado. Nesse caso, amostragem aleatória signfica que cada conjunto (y_i, x_i) é um desenho aleatório da distribuição da população. Nesse contexto, será importante fazer suposições sobre a distribuição conjunta de x_i e ε_i, em particular em respeito à extensão a que a distribuição de ε_i é deixada a depender sobre X. A ideia de uma amostra (aleatória) é mais facilmente entendida no contexto de corte transversal, onde o interesse reside em uma população grande e fixa, por exemplo, todas as famílias brasileiras em setembro de 2016 ou todas as ações listadas na BOVESPA em um dado momento. No contexto de séries temporais, diferentes observações fazem referência a diferentes períodos do tempo, e não faz sentido assumir que nós temos uma amostra aleatória de períodos do tempo. Ao invés disso, nós tomaremos a visão de que a amostra que temos é apenas uma realização do que poderia ocorrer em um dado intervalo de tempo e a aleatoriedade se refere a estados alternativos do mundo. Nesse caso, teremos de fazer algumas suposições sobre como os dados foram gerados (ao invés de como os dados foram *amostrados*).

É importante perceber que sem nenhuma restrição adicional, o modelo *estatístico* proposto em 16 é uma tautologia: para cada valor de β pode-se definir um conjunto de $\varepsilon_i s$ tal que 16 se mantém exata para cada observação. Nós, assim, precisamos impor algumas suposições para dar ao modelo um significado. Uma suposição comum é que o valor esperado de ε_i dadas todas as variáveis explanatórias em x_i será zero, i.e., $E[\varepsilon_i | x_i] = 0$. Usualmente, nos referimos a essa suposição dizendo

[6] Em outras palavras, cada observação corresponderá a uma realização dessas variáveis aleatórias.

que as variáveis explanatórias são **exógenas**. Sob essa suposição, temos que

$$E[y_i|x_i] = x_i^{'}\beta \tag{18}$$

desse modo, a linha de regressão $x_i^{'}\beta$ descreve a esperança condicional de y_i, dados os valores de x_i. **Os coeficientes β_k medem como o valor esperado de y_i é afetado se o valor de x_{ik} mudar, mantendo os demais elementos em x_i constantes**.[7] A teoria econômica, contudo, frequentemente sugere que o modelo contido em 16 descreve uma relação causal, no qual os β coeficientes medem a mudança em y_i *causadas* por a ceteris paribus mudança em x_{ik}. Nesses casos, ε_i tem uma interpretação econômica (e não apenas estatística) e impondo que ele não é correlacionado com x_i, como nós fazemos ao impor $E[\varepsilon_i|x_i] = 0$, pode não ser justificado. Pela razão de em muitas aplicações podermos argumentar que variáveis não observadas presentes no termo de erro estarem relacionadas às variáveis x_i, nós devemos ter cuidado ao interpretar os coeficientes da regressão como medidas de efeito causal.[8]

Agora que nossos β coeficientes possuem um significado, nós podemos tentar usar a amostra (y_i, x_i), $i = 1, ..., N$ para dizer alguma coisa sobre eles. A regra que diz como uma dada amostra é traduzida em um valor aproximado para β é referida como um **estimador**. O resultado para uma dada amostra é chamado uma **estimativa**. O *estimador* é um vetor de variáveis aleatórias porque a amostra pode mudar. A *estimativa*, por sua vez, é um vetor de números. O mais amplamente estimador usado em econometria é o estimador de **Mínimos Quadrados Ordinários** ou simplesmente estimador de MQO. Isso é somente a regra de mínimos quadrados ordinários descrita na seção anterior aplicada a uma amostra disponível. O estimador de MQO para β é dado por

$$b = \left(\sum_{i=1}^{N} x_i x_i^{'}\right)^{-1} \sum_{i=1}^{N} x_i y_i \tag{19}$$

Dado que temos assumido o modelo descrito por 16 como verdadeiro, combinado com um esquema de amostragem, b é agora um vetor de variáveis aleatórias. Nosso interesse se posiciona no verdadeiro vetor de parâmetro β não conhecido, e b é considerado uma aproximação a ele. Enquanto uma dada amostra somente produz uma única estimativa, nós avaliamos a qualidade dela através das propriedades do estimador subjacente. O estimador b tem uma distribuição de amostragem porque esse valor depende da amostra que é escolhida (aleatoriamente) da população.

É extremamente importante entender a diferença entre o estimador b e o verdadeiro coeficiente β da população. O primeiro é um vetor de variáveis aleatórias, o resultado disso depende da amostra que é empregada (e, em termos gerais, do método empregado). O segundo é um conjunto fixo de números desconhecidos, caracterizando o modelo populacional descrito em 16. Da mesma forma, a distinção entre o termo de erro ε_i e os resíduos e_i é importante. Termos de erro são não observáveis e suposições distribucionais sobre eles são necessárias para derivar as propriedades de amostragem dos estimadores para β. A próxima seção trata desse aspecto. Os resíduos, por sua vez, são obtidos após a estimação, e seus valores dependerão do valor estimado para β e, por conseguinte, dependerão da amostra e do método de estimação.

As propriedades do termo de erro ε_i e dos resíduos e_i não são as mesmas e ocasionalmente são

[7] Os economistas costumam se referir a isso como **ceteris paribus**.

[8] Voltaremos a esse problema quando estivermos tratando de *endogeneidade*.

bastante diferente. Por exemplo,

$$\sum_{i=1}^{N} x_i(y_i - x_i'\tilde{\beta}) = \sum_{i=1}^{N} x_i e_i = 0 \tag{20}$$

é tipicamente não satisfeita quando os resíduos são substituídos pelos termos de erro.[9]

Referências

Verbeek, M. 2012. *A Guide to Modern Econometrics*. Editora Wiley.

[9] Ao longo do nosso curso, usaremos essa notação: ε_i refere-se ao termo de erro e e_i aos resíduos.