Estatística II

1º Sem. 2015/2016

Lic. Economia e Finanças

Modelo de Regressão Linear Simples

Introdução ao modelo

- Vimos que a econometria estuda $E(Y|x)$.
- De uma forma mais geral, vamos estudar a análise de regressão.

  Regressão: A regressão de Y em X é qualquer característica da distribuição condicionada $f(y|x)$ expressa como função de x.

Introdução ao modelo

- Frequentemente o termo regressão é incorrectamente usado para designar outras coisas.
- Poderíamos estar interessados noutras medidas de localização, por exemplo a **mediana**.
- Nesta disciplina vamos estudar essencialmente o valor esperado condicionado.

Escolha do valor esperado condicionado

Estatística II 1^{o} Sem. 2015/2016

Lic. Economia e Finanças

Regressão Linear Simples

Introdução

Estimação OLS

Propriedades algébricas do OLS

Não-linearidades

- Porquê? Porque $E(Y|x)$ é a função de x que minimiza
$$E\left[(Y - h(x))^2\right].$$
- (A mediana minimiza $E[|Y - h(x)|]$, e podem usar-se funções perca não simétricas.)
- Vamos começar por considerar o caso em que x é univariado (isto é útil pois permite usar gráficos para ilustrar o estudo).

Definição do modelo

- Temos que definir 3 coisas:

    1. A forma de $E(Y|x)$ (vamos assumir linearidade)
    2. Como tratar o efeito das restantes variáveis em Y
    3. Como fazer análise *ceteris paribus*

Definição do modelo

- O nosso ponto de partida é:

$$
\begin{aligned}
E\left(Y|x\right) &= \beta_0 + \beta_1 x \\
Y &= \beta_0 + \beta_1 x + u
\end{aligned}
$$

- Exemplo: Salário em função da educação

Propriedades de $Y - E(Y|x) = u$

- Propriedades

  1. $E(u|x) = 0$
  2. $E(u) = 0$
  3. $Var(u|x) = \sigma^2_{Y|x}$
  4. $Cov(u, x) = 0$
  5. $Cov(u, g(x)) = 0$

- Só os desvios entre Y e $E(Y|x)$ gozam da propriedade 5.

Observações

- No caso do valor esperado condicional é possível fazer inferência *ceteris paribus* pois é possível considerar variações de x com u constante.

- O problema é que a economia nem sempre está interessada nos parâmetros de valores esperados condicionais, ou pelo menos daqueles que podemos estimar.

Hipóteses do modelo

- SLR.1: Modelo linear nos parâmetros
  $Y = \beta_0 + \beta_1 x + u$
- SLR.2: Amostra casual simples $\{(y_i, x_i) : i = 1 \ldots n\}$
  $y_i = \beta_0 + \beta_1 x_i + u_i$
- SLR.3: Variação amostral: x_i não é constante na amostra
- SLR.4: Erros não correlacionados com x: $E\left(u|x\right) = 0$
- SLR.5: Homocedasticidade: $Var\left(u|x\right) = \sigma^2$ constante

Estimação OLS

- Note-se que $E(u) = 0$ e que $Cov(u, x) = E(xu) = 0$

- Aplicando o método dos momentos:

$$\begin{aligned} \frac{1}{n}\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\ \frac{1}{n}\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i &= 0 \end{aligned}$$

Estimação OLS

- Da primeira equação tira-se

$$
\begin{aligned}
\bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}
$$

Estimação OLS

- Substituindo

$$\frac{1}{n}\sum\left(y_i-\left(\bar{y}-\hat{\beta}_1\bar{x}\right)-\hat{\beta}_1 x_i\right)x_i=0$$

$$\frac{1}{n}\sum\left(y_i-\bar{y}\right)x_i=\frac{1}{n}\sum\hat{\beta}_1\left(x_i-\bar{x}\right)x_i$$

$$\hat{\beta}_1=\frac{\sum\left(y_i-\bar{y}\right)x_i}{\sum\left(x_i-\bar{x}\right)^2}=\frac{\sum\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right)}{\sum\left(x_i-\bar{x}\right)^2}$$

- Isto é a covariância amostral de y e x a dividir pela variância amostral de x (necessariamente **positiva**).

Estimação OLS

- Valores ajustados (ou previstos) de y são dados por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

que se designa por recta de regressão.

- Dado o y "chapéu", podem calcular-se os resíduos

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

Estimação OLS

- Considere-se a soma dos quadrados dos resíduos

$$SSR = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- As estimativas dos mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$ minimizam a SSR.

Estimação OLS

- Para ver isso, considere-se

$$SSR = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = -2 \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

$$\frac{\partial SSR}{\partial \hat{\beta}_1} = -2 \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0$$

- A condição de segunda ordem está no livro.

Propriedades algébricas do OLS

- Propriedades:
  1. $\frac{1}{n}\sum \hat{u}_i = 0$
  2. $\frac{1}{n}\sum x_i \hat{u}_i = 0$
  3. $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
  4. $\bar{\hat{y}} = \bar{y}$
  5. $\frac{1}{n}\sum \hat{y}_i \hat{u}_i = 0$

Observação importante

Notando que $y_i = \hat{y}_i + \hat{u}_i$, esta última propriedade:

$$\frac{1}{n}\sum \hat{y}_i \hat{u}_i = 0$$

implica que o método dos mínimos quadrados decompõe a y_i em duas partes que são ortogonais na amostra.

Propriedades algébricas do OLS

- Considerem-se as seguintes quantidades:

$$
\begin{aligned}
SST &= \sum (y_i - \bar{y})^2 \\
SSE &= \sum (\hat{y}_i - \bar{y})^2 \\
SSR &= \sum \hat{u}_i^2
\end{aligned}
$$

- É possível provar que $SST = SSE + SSR$

Propriedades algébricas do OLS

$$
\begin{aligned}
SST &= \sum (y_i - \bar{y})^2 = \sum \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right]^2 \\
&= \sum \left[\hat{u}_i + (\hat{y}_i - \bar{y}) \right]^2 \\
&= \sum \hat{u}_i^2 + 2 \sum \hat{u}_i \left(\hat{y}_i - \bar{y} \right) + \sum \left(\hat{y}_i - \bar{y} \right)^2 \\
&= SSR + 2 \sum \hat{u}_i \left(\hat{y}_i - \bar{y} \right) + SSE
\end{aligned}
$$

Pelas propriedades 1 e 5 é fácil ver que $\sum \hat{u}_i \left(\hat{y}_i - \bar{y} \right) = 0$, pelo que $SST = SSE + SSR$.

Propriedades algébricas do OLS

- Esta decomposição permite avaliar a "qualidade do ajustamento":

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

$$0 \leq R^2 \leq 1$$

- No modelo de regressão simples $R^2 = \rho^2_{y,x}$.

Propriedades algébricas do OLS

- O que se passa se alterarmos as unidades de medida das variáveis?

$$
\begin{aligned}
Y &= \beta_0 + \frac{\beta_1}{\alpha}(\alpha x) + u \\
\alpha Y &= \alpha\beta_0 + \alpha\beta_1 x + \alpha u \\
\alpha Y + \theta &= (\alpha\beta_0 + \theta) + \alpha\beta_1 x + \alpha u
\end{aligned}
$$

- Em qualquer destes casos o R^2 não se altera.

Não-linearidades

- Aparentemente, modelos do tipo

$$Y = \beta_0 + \beta_1 x + u$$

  só permitem descrever relações lineares.

- Felizmente, não é assim!

Não-linearidades

- Por exemplo, o modelo

$$wage = \beta_0 + \beta_1 edu + u$$

implica que cada ano adicional de escolaridade tem o **mesmo** impacto no salário.

- Isto é pouco realista!

- Felizmente, o modelo tem que ser **linear nos parâmetros** mas não nas variáveis.

Não-linearidades

- Considere-se o modelo alternativo

$$\ln(wage) = \beta_0 + \beta_1 edu + u$$

- Agora

$$\Delta edu = 1 \Longrightarrow \Delta \ln(wage) = \beta_1$$

- Como $\Delta \ln(wage) \simeq \frac{\Delta wage}{wage}$ este modelo implica que cada ano adicional de escolaridade tem o **mesmo** impacto **percentual** no salário.

Exemplo 2.10

EXAMPLE 2.10 A LOG WAGE EQUATION

Using the same data as in Example 2.4, but using log(*wage*) as the dependent variable, we obtain the following relationship:

$$\widehat{\log(wage)} = 0.584 + 0.083\, educ \qquad [2.44]$$

$$n = 526,\ R^2 = 0.186.$$

The coefficient on *educ* has a percentage interpretation when it is multiplied by 100: $\widehat{wage}$ increases by 8.3% for every additional year of education. This is what economists mean when they refer to the "return to another year of education."

It is important to remember that the main reason for using the log of *wage* in (2.42) is to impose a constant percentage effect of education on *wage*. Once equation (2.44) is obtained, the natural log of *wage* is rarely mentioned. In particular, it is *not* correct to say that another year of education increases log(*wage*) by 8.3%.

The intercept in (2.44) is not very meaningful, because it gives the predicted log(*wage*), when *educ* = 0. The *R*-squared shows that *educ* explains about 18.6% of the variation in log(*wage*) (*not wage*). Finally, equation (2.44) might not capture all of the nonlinearity in the relationship between wage and schooling. If there are "diploma effects," then the twelfth year of education—graduation from high school—could be worth much more than the eleventh year. We will learn how to allow for this kind of nonlinearity in Chapter 7.

Não-linearidades

- Considere-se agora uma função procura

$$Q = \beta_0 + \beta_1 P$$

A elasticidade neste modelo é dada por

$$\varepsilon_{Q,P} = \frac{\beta_1 P}{\beta_0 + \beta_1 P}$$

Não-linearidades

- Para se obter uma elasticidade constante podemos usar a especificação

$$\ln(Q) = \beta_0 + \beta_1 \ln(P)$$

$$Q = \exp(\beta_0 + \beta_1 \ln(P))$$

$$\frac{\partial Q}{\partial P} = \frac{\beta_1}{P} Q \Longrightarrow \varepsilon_{Q,P} = \frac{\beta_1 Q}{P} \frac{P}{Q} = \beta_1$$

Isto é aquilo a que se chama um modelo linearizável.

Exemplo 2.11

EXAMPLE 2.11 **CEO SALARY AND FIRM SALES**

We can estimate a constant elasticity model relating CEO salary to firm sales. The data set is the same one used in Example 2.3, except we now relate *salary* to *sales*. Let *sales* be annual firm sales, measured in millions of dollars. A constant elasticity model is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u, \qquad [2.45]$$

where β_1 is the elasticity of *salary* with respect to *sales*. This model falls under the simple regression model by defining the dependent variable to be $y = \log(salary)$ and the independent variable to be $x = \log(sales)$. Estimating this equation by OLS gives

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales) \qquad [2.46]$$

$$n = 209, R^2 = 0.211.$$

The coefficient of log(*sales*) is the estimated elasticity of *salary* with respect to *sales*. It implies that a 1% increase in firm sales increases CEO salary by about 0.257%—the usual interpretation of an elasticity.

Não-linearidades

- Há modelos não linearizáveis. Por exemplo: probabilidade de um indivíduo ser vegetariano em função da idade

$$\Pr = \frac{\exp\left(\beta_0 + \beta_1 age\right)}{1 + \exp\left(\beta_0 + \beta_1 age\right)}$$

Tabela 2.3

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Exercício 2.6

6 Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312 \log(dist)$$

$$n = 135, R^2 = 0.162.$$

(i) Interpret the coefficient on log(*dist*). Is the sign of this estimate what you expect it to be?

(ii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision on where to put the incinerator.)

(iii) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

Exercício 2.8

8 Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 through SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by assuming the intercept is zero (see Section 2.6).

(i) Find $\mathrm{E}(\tilde{\beta}_1)$ in terms of the x_i, β_0, and β_1. Verify that $\tilde{\beta}_1$ is unbiased for β_1 when the population intercept (β_0) is zero. Are there other cases where $\tilde{\beta}_1$ is unbiased?

(ii) Find the variance of $\tilde{\beta}_1$. (*Hint:* The variance does not depend on β_0.)

(iii) Show that $\mathrm{Var}(\tilde{\beta}_1) \leq \mathrm{Var}(\hat{\beta}_1)$. [*Hint:* For any sample of data, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, with strict inequality unless $\bar{x} = 0$.]

(iv) Comment on the tradeoff between bias and variance when choosing between $\hat{\beta}_1$ and $\tilde{\beta}_1$.