Econometria Aplicada [Módulo em R]

Regressão Linear Simples

Vítor Wilher

Cientista de Dados | Mestre em Economia

Plano de Voo

Pacotes utilizados nessa seção

Introdução ao mundo da econometria

Estruturas de Dados

O escopo do nosso curso

Regressão linear simples

Salários versus Experiência no emprego

Entendendo melhor o que é uma aproximação linear

Exemplo 2.5 do Wooldridge

Exemplo 2.6 do Wooldridge

Pacotes utilizados nessa seção

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

Introdução ao mundo da econometria

A econometria é a união entre teoria econômica, dados observados e métodos estatísticos. Por tratar de uma interação incomum, é uma das disciplinas mais temidas e desafiadoras das faculdades de economia. Nosso objetivo nesse curso é, por suposto, romper com esse temor, apresentando uma teoria regada de exemplos intuitivos, que façam parte do dia a dia dos alunos.

Para ilustrar, suponha que você esteja interessado em entender como salários e escolaridade se relacionam ao longo do tempo. Intuitivamente, é possível dizer que à medida que as pessoas estudam mais, elas conseguem auferir uma renda mensal mais elevada. De fato, é isso que a gente observa, não é mesmo?

Introdução ao mundo da econometria

Imagine, por suposto, que você formule uma teoria que relacione os salários a algumas variáveis, entre elas a escolaridade obtida por uma pessoa ao longo da vida.[1] Sua hipótese básica, nascida da observação, é de que maior escolaridade está relacionada a maiores salários.

Uma vez que você tenha uma teoria, você pode especificá-la em uma equação, como abaixo:

$$\text{Salários} = \beta_0 + \beta_1 \text{Educação} \tag{1}$$

Onde os `Salários` representam o que os economistas chamam de **variável endógena**, a `Educação` é a **variável exógena**, β_0 é uma constante e β_1 é um parâmetro, que mede o efeito da educação nos salários. O que você tem aqui é um `modelo matemático`, que relaciona a variável educação à variável salários.

[1] De fato, essa teoria já existe. Veja o trabalho pioneiro aqui.

Introdução ao mundo da econometria

Essa relação *determinística* entre salários e educação é de interesse limitado, entretanto. Isto porque, nada garante que no mundo real, a educação seja a única variável a afetar os salários. É possível que, entre outras, os anos de experiência de uma pessoa também influenciem no salário obtido em um mês. Ou, para não ficar apenas nas flores, que os próprios relacionamentos que uma determinada pessoa cultive no ambiente de trabalho influencie no salário auferido.

Introdução ao mundo da econometria

Nesse contexto, caso você esteja interessado em construir uma análise empírica que busque *embasar* sua teoria, será preciso construir um modelo econométrico, em que admite-se a influência da educação sobre os salários, mas trabalha-se com um *termo de erro*, uma medida do nosso desconhecimento sobre o *conjunto de todas as variáveis* que de fato influenciam os salários. Isso pode ser ilustrado como abaixo:

$$\text{Salários} = \beta_0 + \beta_1 \text{Educação} + \varepsilon \tag{2}$$

Onde ε representa o *termo de erro*. Em outras palavras, saímos de um *modelo determinístico* para um modelo *estocástico*.

Introdução ao mundo da econometria

Especificado o seu modelo econométrico, é hora de obter os dados, não é mesmo? Com o código abaixo, nós acessamos o *dataset* `wage1`, que faz parte do livro de Wooldridge [2013] e que possui um `pacote` no `R` de mesmo nome.[2]

```
data(wage1)
```

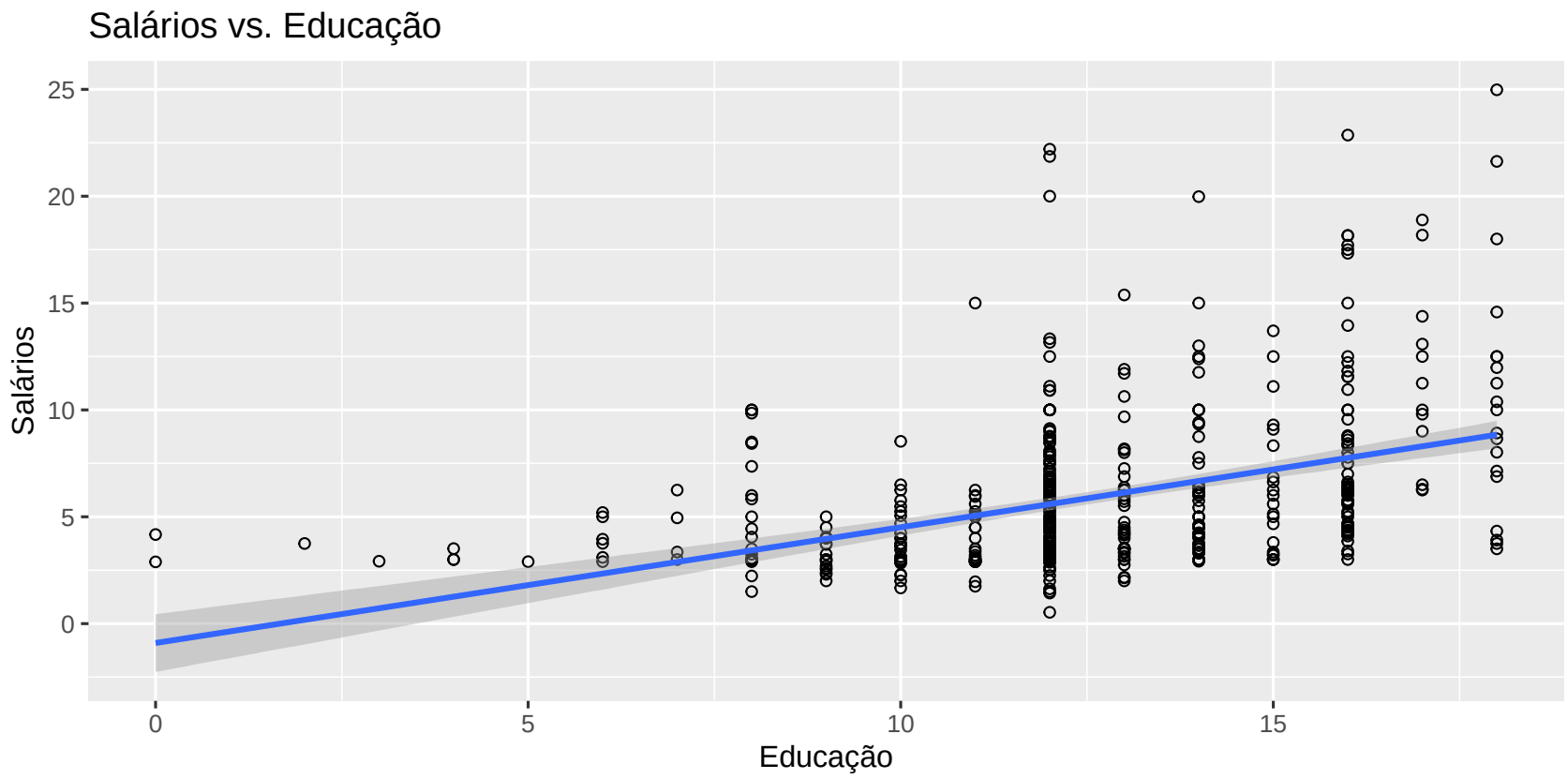
[2] Caso ainda não tenha instalado, o faço com a função `install.packages()`.

Introdução ao mundo da econometria

Com os dados em mãos, talvez seja interessante verificar como `Salários` e `Educação` estão relacionados. Para tal, vamos plotar um gráfico de correlação?

Introdução ao mundo da econometria

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



Introdução ao mundo da econometria

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.

Introdução ao mundo da econometria

Em resumo, portanto, e com base em Gujarati [2006], o método econométrico é composto por alguns passos:

1. Formulação de uma teoria ou hipótese;
2. Construção de um modelo matemático;
3. Especificação econométrica;
4. Coleta de dados;
5. Estimação dos parâmetros do modelo;
6. Teste de Hipóteses;
7. Uso do modelo para fins de previsão e de formulação de políticas.

Introdução ao mundo da econometria

A econometria é, em geral, associada aos economistas, mas pode e tem sido utilizada em um amplo número de disciplinas e por uma gama grande de profissionais. Formalizar relações entre variáveis em modelos matemáticos e, posteriormente, em modelos estatísticos sujeitos à estimação pode, afinal, ser estendido sem maiores dificuldades para diversos campos de estudo, como a metereologia ou a física, por exemplo.

Introdução ao mundo da econometria

O trabalho do econometrista consiste em especificar e quantificar relações entre variáveis distintas. Econometristas formulam modelos estatísticos, geralmente baseados em alguma teoria, confrontam esses modelos com dados observados, tentando encontrar uma especificação que melhor se adeque. Os elementos desconhecidos nessas especificações, os parâmetros, são então estimados a partir de uma amostra de dados. Por fim, o trabalho consiste em verificar a adequação dos resultados encontrados. Isto é, cabe ao econometrista julgar se o modelo em questão pode ser utilizado seja para fins de previsão ou de avaliação de uma determinada política.

Estruturas de Dados

O modelo econométrico mais apropriado estará fortemente associado à estrutura de dados disponível. Há, em termos gerais, quatro grandes grupos de estruturas. Uma primeira *classe* de modelos descreverá relações entre diversos períodos de tempo, isto é, entre o passado e o presente. Quanto, por exemplo, a inflação passada influencia a inflação atual? Esse tipo de modelo é o que chamamos de *série temporal*, tendo problemas e discussões bastante específicas.

Estruturas de Dados

Um segundo tipo de modelo estará interessado nas relações entre variáveis em um determinado ponto do tempo. Dada uma quantidade de pessoas com salários e escolaridade distintas, como essas variáveis estão relacionadas. Esse tipo de estrutura é chamada de *corte transversal* ou *cross-section*.

Uma terceira classe de modelos diz respeito a *amostras distintas* em diferentes anos, chamados frequentemente de *cortes transversais agrupados*. Para um ano x considera-se uma amostra de pessoas com salários e escolaridade, enquanto para um ano y considera-se outra amostra com outras pessoas com salários e escolaridade. Por fim, é possível acompanhar características de uma mesma amostra ao longo de vários anos. É o que chamamos de *dados em painel*.

O escopo do nosso curso

Nossa intenção é possibilitar ao leitor uma **base** para que ele possa avançar na construção de modelos econométricos envolvendo essas diferentes estruturas de dados. Para tal, vamos introduzir o aluno aos modelos de regressão linear estimados via o clássico método de mínimos quadrados ordinários.

Uma vez que o aluno tenha intimidade com as propriedades dos estimadores de MQO, baseadas nas premissas de **Gauss-Markov**, será possível avaliar em quais estruturas de dados problemas como heterocedasticidade, multicolinearidade, autocorrelação e endogeneidade são mais comuns.

O escopo do nosso curso

Ademais, o aluno poderá identificar se de fato o modelo representado pela equação 2 é o mais adequado para representar os salários. Se não seria mais adequado acrescentar outras variáveis, considerando, claro, que essas estejam disponíveis. Por exemplo, poderíamos considerar a experiência, modificando o modelo proposto pela 2 por:

$$\text{Salários} = \beta_0 + \beta_1 \text{Educação} + \beta_2 \text{Experiência} + \varepsilon \qquad (3)$$

Uma vez consideradas essas variáveis, poderíamos *estimar* os dois modelos, como abaixo.

O escopo do nosso curso

```
modelo1 <- lm(wage~educ, data=wage1)
modelo2 <- lm(wage~educ+exper, data=wage1)
```

O escopo do nosso curso

```
stargazer(modelo1, modelo2,
          title='Modelos para os Salários',
          font.size = 'tiny', header=FALSE)
```

Table 1: Modelos para os Salários

	Dependent variable:	
	wage	
	(1)	(2)
educ	0.541^{***}	0.644^{***}
	(0.053)	(0.054)
exper		0.070^{***}
		(0.011)
Constant	-0.905	-3.391^{***}
	(0.685)	(0.767)
Observations	526	526
R^2	0.165	0.225
Adjusted R^2	0.163	0.222
Residual Std. Error	3.378 (df = 524)	3.257 (df = 523)
F Statistic	103.363^{***} (df = 1; 524)	75.990^{***} (df = 2; 523)
Note:		*p<0.1; **p<0.05; ***p<0.01

O escopo do nosso curso

Nossa tarefa aqui será, em primeiro lugar, identificar o significado dos valores estimados para β_1 e β_2. Sabemos, de antemão, que esses *parâmetros* medem o efeito, respectivamente, da educação e da experiência nos salários. Mas em que medida?

Para responder essa e outras questões, é preciso ter pleno conhecimento sobre a unidade em que salários, educação e experiência estão sendo identificados. A forma como uma variável é apresentada vai modificar a interpretação do parâmetro. Com efeito, é preciso qualificar de forma exata o que o parâmetro β está medindo.

O escopo do nosso curso

Ademais, é preciso verificar se o parâmetro estimado é *estatisticamente significativo*. Para isso, devemos proceder um teste de hipótese, algo não trivial mesmo para alunos que já passaram por cursos formais de econometria.

Em que medida, por fim, o modelo estimado está ajustado aos dados? Como podemos verificar se de fato estamos conseguindo determinar os salários? Ou, de outra forma, se podemos comparar os modelos 1 e 2. Qual deles é o melhor para representar os salários? É possível dizer que um é melhor do que o outro? São questões interessantíssimas, que poderemos explorar de forma plena usando o R.

Regressão linear simples

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.

Regressão linear simples

Para responder essa pergunta, primeiro, considere uma combinação linear arbitrária, incluindo a constante, que pode ser escrita como

$$\tilde{\beta}_1 + \tilde{\beta}_2 x_2 + ... + \tilde{\beta}_k x_k \tag{4}$$

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:[3]

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

Regressão linear simples

$$y_i - [\tilde{\beta}_1 + \tilde{\beta}_2 x_{i2} + ... + \tilde{\beta}_k x_{ik}] \tag{5}$$

Ou,

$$y_i - x_i^{'} \tilde{\beta} \tag{6}$$

Regressão linear simples

Nesses termos, devemos escolher valores para $\tilde{\beta}_1, ..., \tilde{\beta}_k$ de modo que a diferença dada por 6 seja a menor possível. Nesse Curso, a propósito, 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.

Regressão linear simples

Isto é, devemos determinar $\tilde{\beta}$ de modo a minimizar a seguinte função objetivo

$$S(\tilde{\beta}) \equiv \sum_{i=1}^{N} (y_i - x_i^{'} \tilde{\beta})^2 \tag{7}$$

Em resumo, minimizar a soma dos erros da aproximação ao quadrado.

Regressão linear simples

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,

$$-2\sum_{i=1}^{N} x_i(y_i - x_i^{'}\tilde{\beta}) = 0 \tag{8}$$

Ou

$$(\sum_{i=1}^{N} x_i x_i^{'})\tilde{\beta} = \sum_{i=1}^{N} x_i y_i \tag{9}$$

Regressão linear simples

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

$$b = (\sum_{i=1}^{N} x_i x_i^{'})^{-1} \sum_{i=1}^{N} x_i y_i \tag{10}$$

Tomando a condição de segunda ordem, é fácil verificar que b, de fato, corresponde ao mínimo de 7.

Regressão linear simples

A combinação linear resultante de x_i é então dada por

$$\widehat{y}_i = x_i^{'} b \tag{11}$$

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.

Regressão linear simples

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.[4]

[4] 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**.

Regressão linear simples

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

$$y_i = \widehat{y} + e_i = x_i^{'} b + e_i \tag{12}$$

Isso nos permite escrever o valor mínimo para a função objetivo 7 como

$$S(b) = \sum_{i=1}^{N} e_i^2 \tag{13}$$

O que pode ser descrito como o *somatório dos resíduos ao quadrado*.

Regressão linear simples

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 8, substituindo a solução b, nós obtemos

$$\sum_{i=1}^{N} x_i(y_i - x_i^{'}\tilde{\beta}) = \sum_{i=1}^{N} x_i e_i = 0 \qquad (14)$$

Isso significa que o vetor $\varepsilon = (e_1, ..., e_N)$ é ortogonal para cada vetor de observações nas $x - variáveis$.[5]

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

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**.

Regressão linear simples

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 10. De forma equivalente, nós podemos minimizar a soma dos resíduos ao quatro com respeito aos parâmetros desconhecidos de forma direta.

Regressão linear simples

Assim, teremos:

$$S(\tilde{\beta}_1, \tilde{\beta}_2) = \sum_{i=1}^{N} (y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i)^2 \tag{15}$$

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

$$\frac{\partial S(\tilde{\beta}_1, \tilde{\beta}_2)}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^{N} (y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i) = 0 \tag{16}$$

$$\frac{\partial S(\tilde{\beta}_1, \tilde{\beta}_2)}{\partial \tilde{\beta}_2} = -2 \sum_{i=1}^{N} x_i (y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i) = 0 \tag{17}$$

Regressão linear simples

Da equação 16, nós podemos escrever

$$b_1 = \frac{1}{N}\sum_{i=1}^{N} y_i - b_2 \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{y} - b_2 \bar{x}, \tag{18}$$

onde b_2 pode ser resolvida da combinação de 17 e 18.

Regressão linear simples

Primeiro, da equação 17 escrevemos

$$\sum_{i=1}^{N} x_i y_i - b_1 \sum_{i=1}^{N} x_i - \left(\sum_{i=1}^{N} x_i^2 \right) b_2 = 0 \qquad (19)$$

Regressão linear simples

e depois substituímos 18 para obter

$$\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y} - \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right) b_2 = 0 \qquad (20)$$

Regressão linear simples

de tal modo que nós podemos resolver para o coeficiente de inclinação b_2 como

$$b_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{21}$$

Ao dividir tanto o numerador quanto o denominador por $N - 1$, temos que b_2 será a taxa entre a covariância amostral entre x e y e a variância amostral de x. Da equação 18, o intercepto é determinado de modo a tornar o erro médio de aproximação (residual) igual a zero.

Salários versus Experiência no emprego

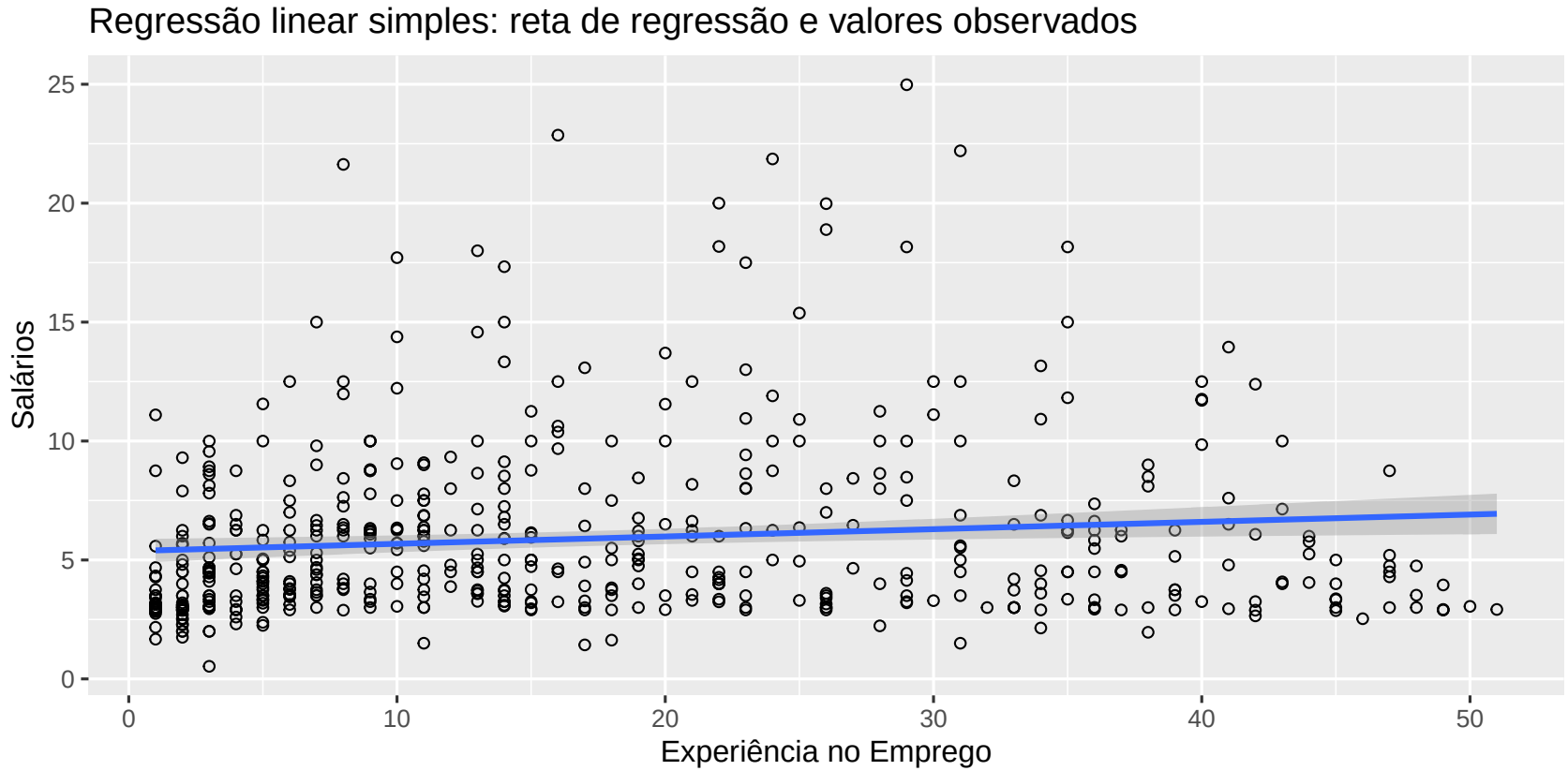
Para ilustrar a **reta de regressão**, vamos pegar aquele mesmo conjunto de dados da seção anterior e tomar o modelo representado por

$$\text{Salário}_i = \beta_0 + \beta_1 \text{Experiência}_i + \varepsilon_i$$

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

Salários versus Experiência no emprego

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



Salários versus Experiência no emprego

```
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 Salários $= 5.37 + 0.03$ Experiência.[6]

[6] Para obter a expressão, basta rodar o código *lm(wage∼exper, data=wage1)*.

Entendendo melhor o que é uma 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.

Entendendo melhor o que é uma aproximação linear

Table 2: 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

Entendendo melhor o que é uma aproximação linear

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

Exemplo 2.5 do Wooldridge

Para terminar, Vamos considerar o `Exemplo 2.5` de Wooldridge [2013]. O arquivo `vote1.dta` contém dados sobre resultados eleitorais e gastos de campanha de 173 disputas entre dois partidos, em 1988, para a Casa dos Representantes dos Estados Unidos. Há dois candidatos em cada disputa: A e B. Seja *voteA* a percentagem de votos recebida pelo candidato A e *shareA* a percentagem das despesas totais de campanha que cabem ao candidato A. Podemos relacionar essas duas variáveis conforme o código abaixo.

Exemplo 2.5 do Wooldridge

```
data("vote1")
# Regressão
lm(voteA ~ shareA, data=vote1)
```

```
##
## Call:
## lm(formula = voteA ~ shareA, data = vote1)
##
## Coefficients:
## (Intercept)       shareA
##     26.8122       0.4638
```

Exemplo 2.5 do Wooldridge

O resultado é positivo. Isto é, se a percentagem de despesas que cabe ao candidato A aumentar em um ponto percentual, os votos dele aumentam em quase meio ponto percentual.

Exemplo 2.6 do Wooldridge

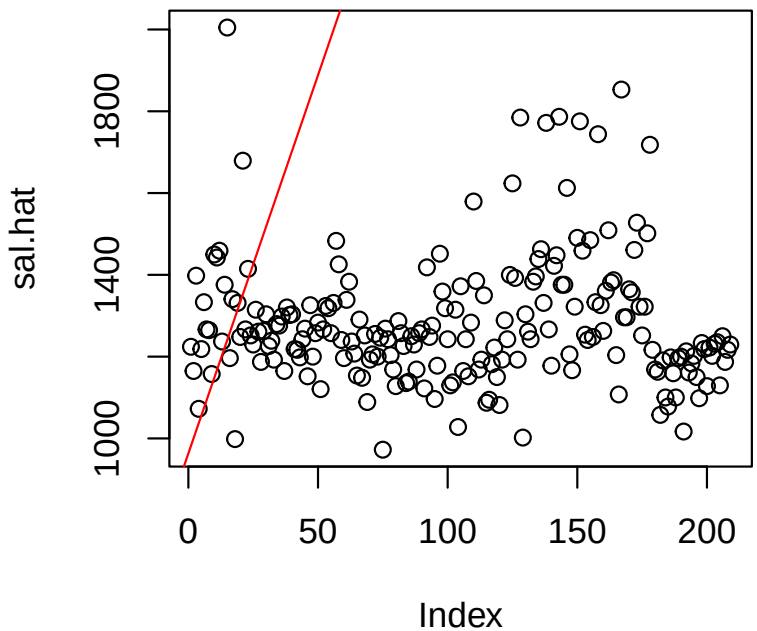
Vamos considerar o `Exemplo 2.6` de Wooldridge [2013] para mostrar algumas coisas que podem ser feitas no **R**.

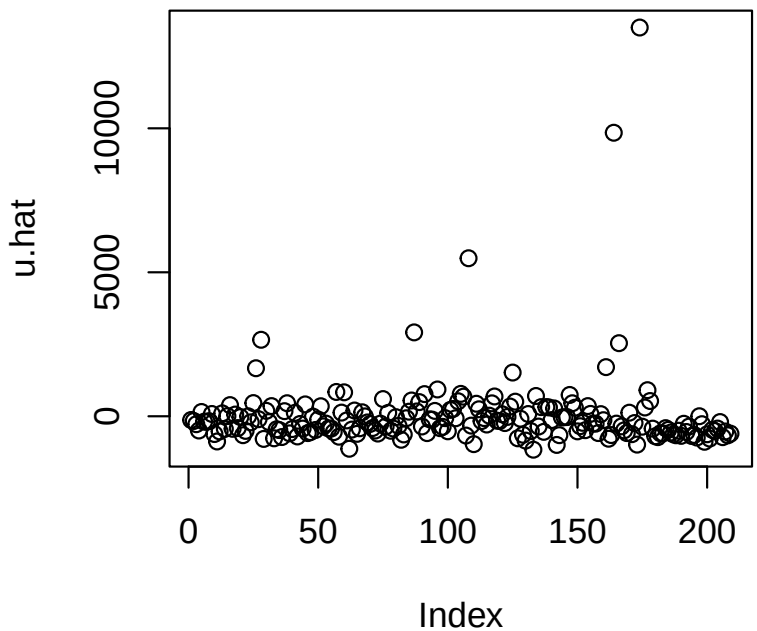
```
data("ceosal1")
# Extrair variáveis como vetores
sal <- ceosal1$salary
roe <- ceosal1$roe
# Regressão com vetores
CEOregres <- lm(sal~roe)
# Obtendo os valores preditos e os resíduos
sal.hat <- fitted(CEOregres)
u.hat <- resid(CEOregres)
```

Exemplo 2.6 do Wooldridge

E agora os gráficos.

```
par(mfrow=c(1,2))
plot(sal.hat)
abline(CEOregres, col='red')
plot(u.hat)
```





D. Gujarati. *Econometria Básica*. Editora Campus, 2006.
M. Verbeek. *A Guide to Modern Econometrics*. Editora Wiley, 2012.
J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Editora Cengage, 2013.