Formação Cientista de Dados

Dia 03: Análise de Regressão (Seções 12 e 13)

Vítor Wilher

Cientista de Dados | Mestre em Economia

Plano de Voo

Introdução

Estamos interessados em estimar os parâmetros populacionais β_0 e β_1 de um modelo de regressão simples

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

a partir de uma amostra aleatória de y e x. De acordo com Wooldridge [2013], os estimadores de Mínimos Quadrados Ordinários (MQO) serão

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x} \tag{2a}$$

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Varx}. \tag{2b}$$

Introdução

Baseado nos parâmetros estimados, a reta de regressão será

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{3}$$

Para uma dada amostra, nós precisaremos calcular as quatro estatísticas $\bar{y}$, $\bar{x}$, $Cov(x, y)$ e $Var(x)$ e colocá-las nessas equações. Para ilustrar, vamos considerar o seguinte exemplo.

Salários de CEOs e Retornos sobre o patrimônio

Vamos considerar o exemplo 2.3 de Wooldridge [2013] sobre *Salários de CEOs e Retornos sobre o patrimônio*. Para isso, considere o seguinte modelo

$$salary = \beta_0 + \beta_1 roe + u \tag{4}$$

onde *salary* é o salário anual de CEO em milhares de dólares e *roe* é o retorno médio sobre o patrimônio em percentual. O parâmetro β_1 irá medir a variação no salário anual quando o retorno médio sobre o patrimônio aumentar em um ponto percentual. Para estimar esse modelo, podemos utilizar o conjunto de dados `ceosal1`.

Introdução

```
data(ceosal1, package='wooldridge')
attach(ceosal1)
```

Uma vez que tenhamos carregado o conjunto de dados, podemos calcular manualmente os parâmetros β_0 e β_1, como abaixo.

Introdução

```r
# Cálculo manual dos parâmetros
b1hat = cov(roe,salary)/var(roe)
b1hat
```

```
## [1] 18.50119
```

```r
b0hat = mean(salary) - b1hat*mean(roe)
b0hat
```

```
## [1] 963.1913
```

Introdução

Isto é, a **reta de regressão** será dada por

$$\hat{salary} = 963.1913 + 18.5012 * roe \tag{5}$$

o que pode ser facilmente obtido com o código abaixo:

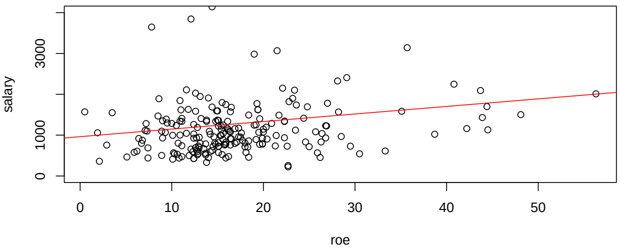
```
lm(salary ~ roe)
```

```
##
## Call:
## lm(formula = salary ~ roe)
##
## Coefficients:
## (Intercept)          roe
##       963.2         18.5
```

Introdução

Implicando que para um $roe = 0$, teremos um salário previsto de US\$ 963.191, que é o intercepto. Ademais, se $\Delta roe = 1$, então $\Delta salary = US\$18.501$. Podemos, por fim, desenhar a reta de regressão com o código abaixo.

```
CEOregress = lm(salary ~ roe)
plot(roe, salary, ylim=c(0,4000))
abline(CEOregress, col='red')
```



Introdução

O **modelo de regressão linear**, tipicamente estimado por *Mínimos Quadrados Ordinários (MQO)*, é a base da estatística aplicada. O modelo é

$$y_i = x_i^T \beta + \varepsilon_i \quad i = 1, ..., n. \tag{6}$$

ou, na forma matricial,

$$y = X\beta + \varepsilon \tag{7}$$

onde y é um vetor $n \times 1$ contendo a variável dependente, $X = (x_1, ..., x_n)$ é uma matriz $n \times k$ de regressores, β é um vetor $k \times 1$ de coeficientes e ε é um vetor $n \times 1$ de termos de erro.

Introdução

Suposições sobre ε dependem do contexto. Para dados **cross section**, $E(\varepsilon|X) = 0$ (exogeneidade) e $Var(\varepsilon|X) = \sigma^2 I$ (homocedasticidade condicional e ausência de autocorrelação) são comuns. Já para **séries temporais**, exogeneidade é algo mais complicado, sendo substituído por algo como $E(\varepsilon_j|x_i) = 0, i \leq j$.

De modo a fixar as notações, temos que $\hat{\beta} = (X^T X)^{-1} X^T y$ denota o estimador de MQO para β. Os valores estimados serão dados por $\hat{y} = X\hat{\beta}$, os resíduos serão dados por $\hat{\varepsilon} = y - \hat{y}$ e a soma dos quadrados dos resíduos por $\hat{\varepsilon}^T \hat{\varepsilon}$.[1]

[1] Para maiores detalhes sobre o modelo de regressão linear, ver Greene [2003], Stock and Watson [2007], Wooldridge [2013] ou Verbeek [2012].

Regressão Linear Simples

Vamos continuar nosso entendimento de **regressões simples** com um pequeno exemplo retirado de Stock and Watson [2007], disponível no pacote **AER**, que pode ser carregado e transformado como abaixo.

```r
library(AER)
data("Journals")
journals = Journals[, c("subs", "price")]
journals$citeprice = Journals$price/Journals$citations
summary(journals)
```

```
##       subs            price          citeprice
##  Min.   :   2.0   Min.   :  20.0   Min.   : 0.005223
##  1st Qu.:  52.0   1st Qu.: 134.5   1st Qu.: 0.464495
##  Median : 122.5   Median : 282.0   Median : 1.320513
##  Mean   : 196.9   Mean   : 417.7   Mean   : 2.548455
##  3rd Qu.: 268.2   3rd Qu.: 540.8   3rd Qu.: 3.440171
##  Max.   :1098.0   Max.   :2120.0   Max.   :24.459459
```

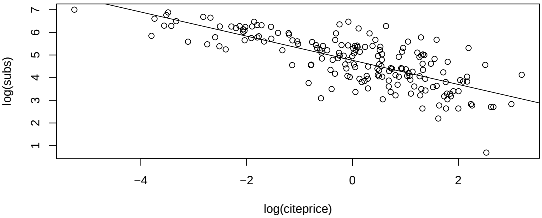
Regressão Linear Simples

Podemos estar interessados em estimar o efeito do preço de uma citação sobre o número de assinantes. Isto é,

$$log(subs)_i = \beta_1 + \beta_2 log(citeprice)_i + \varepsilon_i. \tag{8}$$

A equação 8 pode ser estimada e plotada (a reta de regressão) com o seguinte código.

```
plot(log(subs) ~ log(citeprice), data = journals)
jour_lm <- lm(log(subs) ~ log(citeprice), data = journals)
abline(jour_lm)
```



Regressão Linear Simples

O gráfico pode ficar um pouco mais interessante utilizando o pacote **ggplot2**. . .

```
library(ggplot2)
ggplot(journals, aes(log(citeprice), log(subs)))+
  geom_point(stat='identity')+
  geom_smooth(method='lm')
```

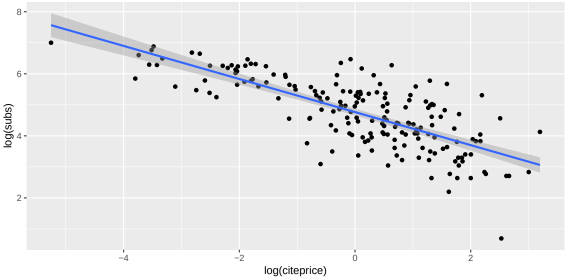


Figure 1: Reta de Regressão

Regressão Linear Simples

A função **lm** estima via MQO nosso modelo de regressão linear...

```r
summary(lm(log(subs) ~ log(citeprice), data = journals))
```

```
##
## Call:
## lm(formula = log(subs) ~ log(citeprice), data = journals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.72478 -0.53609  0.03721  0.46619  1.84808
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.76621    0.05591   85.25   <2e-16 ***
## log(citeprice) -0.53305    0.03561  -14.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7497 on 178 degrees of freedom
## Multiple R-squared:  0.5573, Adjusted R-squared:  0.5548
## F-statistic:   224 on 1 and 178 DF,  p-value: < 2.2e-16
```

Regressão Linear Simples

Rodar uma regressão linear e dela tirar previsões é simples. Primeiro puxamos dados. Nesse exemplo vamos usar os dados da base de dados `emissions`, disponível no pacote `UsingR`. O procedimeto é também simples, carregamos o pacote com a base e usamos a função `data` para especificar qual base queremos - com o nome sempre entre aspas.

A base que usamos relaciona emições de carbono e PIB de 26 países. Se quisermos estimar o efeito que o PIB tem sobre a emissão de poluentes, podemos usar um modelo de regressão linear, com a função `lm`.

Regressão Linear Simples

```r
# importamos dados
library(UsingR)
data("emissions")
# estimamos um modelo linear
modelo = lm(CO2 ~ GDP, data = emissions)
# vemos uma tabela descritiva do modelo estimado
summary(modelo)
```

```
##
## Call:
## lm(formula = CO2 ~ GDP, data = emissions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1107.35   -81.47   -32.69   126.33  1438.79
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.043e+01  9.441e+01   0.216     0.83
## GDP         7.815e-04  5.233e-05  14.933   1.2e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 427.4 on 24 degrees of freedom
## Multiple R-squared:  0.9028, Adjusted R-squared:  0.8988
## F-statistic:   223 on 1 and 24 DF,  p-value: 1.197e-13
```

Regressão Linear Simples

A tabela nos informa as estatística t dos testes t marginais no parâmetro do modelo. Elas são a estatística do teste para a hipótese nula de que o parâmetro é na verdade zero. Ela mede o quanto confiamos que a variável explicativa tem algum efeito sobre a variável de interesse. Observe que PIB (medido pela variável `GDP`) tem um efeito estatisticamente significante sobre nível de poluição. O que acontece se também levarmos em conta o PIB per capita? Será que só tamanho da economia importa ou também seu nível de desenvolvimento?

Estimação pontual e por intervalos

```r
coef(jour_lm)
```

```
##    (Intercept) log(citeprice)
##      4.7662121     -0.5330535
```

```r
confint(jour_lm, level=0.95)
```

```
##                     2.5 %     97.5 %
## (Intercept)     4.6558822  4.8765420
## log(citeprice) -0.6033319 -0.4627751
```

Previsão

Podemos utilizar nosso modelo para fins de previsão. Por exemplo, podemos estar interessados em verificar o número de assinantes para o preço por citação igual a 2.11.[2]

```
predict(jour_lm, newdata = data.frame(citeprice = 2.11),
        interval = "confidence")
```

```
##        fit      lwr     upr
## 1 4.368188 4.247485 4.48889
```

```
predict(jour_lm, newdata = data.frame(citeprice = 2.11),
        interval = "prediction")
```

```
##        fit      lwr      upr
## 1 4.368188 2.883746 5.852629
```

[2] Os intervalos são baseados na distribuição *t*.

Testando uma hipótese linear

Suponha que queremos testar a hipótese de que a elasticidade do número de assinaturas em relação ao preço por citação seja de menos 0.5. Isto é, $H_0 : \beta_2 = -0.5$.

```r
library(car)
linear.hypothesis(jour_lm, "log(citeprice) = -0.5")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(citeprice) = - 0.5
##
## Model 1: restricted model
## Model 2: log(subs) ~ log(citeprice)
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    179 100.54
## 2    178 100.06  1   0.48421 0.8614 0.3546
```

Regressão Múltipla

Na vida real, a maioria das análises feitas através de uma regressão envolve mais de um regressor. Ademais, há regressores especiais, como *dummies*, que são utilizadas para codificar variáveis categóricas. Por fim, também pode ser necessário transformar tanto os regressores quanto a nossa variável de interesse. Para ilustrar como lidar com esse tipo de problema com o R, vamos utilizar o `dataset` **CPS1988**.

```
library(AER)
data("CPS1988")
summary(CPS1988)
```

```
##       wage            education       experience   ethnicity
##  Min.   :   50.05   Min.   : 0.00   Min.   :-4.0   cauc:25923
##  1st Qu.:  308.64   1st Qu.:12.00   1st Qu.: 8.0   afam: 2232
##  Median :  522.32   Median :12.00   Median :16.0
##  Mean   :  603.73   Mean   :13.07   Mean   :18.2
##  3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:27.0
##  Max.   :18777.20   Max.   :18.00   Max.   :63.0
##   smsa            region      parttime
##  no : 7223   northeast:6441   no :25631
##  yes:20932   midwest  :6863   yes: 2524
##              south    :8760
##              west     :6091
##
##
```

Regressão Múltipla

Nesse `dataset` sobre dados *cross-section* envolvendo determinantes de salários para março de 1988 coletados pelo *US Census Bureau*, **wage** representa o salário em dólares por semana, **education** e **experience** são medidos em anos, **ethnicity** é um fator com dois níveis, *Caucasian* e *African-American*. Há outros três fatores: **smsa**, que indica residência em uma região metropolitana padrão; **region** que indica a região dos EUA; e **parttime** que indica indivíduos trabalhando parte do tempo padrão.[3]

[3] A variável **experience** foi construída tendo por base a idade menos o tempo de escolaridade menos seis. Por isso, há observações negativas na amostra.

Regressão Múltipla

Nosso modelo de interesse é

$$
\begin{aligned}
log(wage) = \beta_1 + \beta_2 experience + \beta_3 experience^2 + \\
+ \beta_4 education + \beta_5 ethnicity + \varepsilon
\end{aligned}
\tag{9}
$$

Como aprendemos na aula anterior, ele pode ser facilmente estimado com o código abaixo no R:

```r
cps_lm <- lm(log(wage) ~ experience + I(experience^2) +
               education + ethnicity, data = CPS1988)
```

Regressão Múltipla

Table 1: Determinantes do Salário Semanal

	Dependent variable:
	log(wage)
experience	0.077^{***}
	(0.001)
I(experienceˆ2)	-0.001^{***}
	(0.00002)
education	0.086^{***}
	(0.001)
ethnicityafam	-0.243^{***}
	(0.013)
Constant	4.321^{***}
	(0.019)
Observations	28,155
R^2	0.335
Adjusted R^2	0.335
Residual Std. Error	0.584 (df = 28150)
F Statistic	$3{,}541.036^{***}$ (df = 4; 28150)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Regressão Múltipla

Como o nosso modelo é semilogarítmico, observe que o retorno de um ano a mais de educação é de 8.57% no salário semanal.

Elasticidades

De forma um pouco mais geral, frequentemente, podemos estar interessados em *elasticidades*, isto é, ao invés dos efeitos marginais vistos anteriormente. A elasticidade, por suposto, busca medir a mudança relativa na variável dependente dada uma mudança relativa em uma das x_i variáveis. Em geral, por suposto, elasticidades são estimadas a partir de modelos lineares a partir da utilização de logaritmos, como abaixo:

$$logy_i = (logx_i)\gamma + v_i \tag{10}$$

Regressão Múltipla

onde $logx_i$ é uma notação abreviada para o vetor com elementos $(1, logx_{i2}, ..., logx_{iK})'$ e é assumido que $E(v_i|logx_i) = 0$. Chamamos essa relação de **modelo loglinear**. Nesse caso,

$$\frac{\partial E(y_i|x_i)}{\partial x_{ik}} . \frac{x_{ik}}{E(y_i|x_i)} \approx \frac{\partial E(logy_i|logx_i)}{\partial logx_{ik}} = \gamma_k \qquad (11)$$

onde $\approx$ vem do fato de que $E(logy_i|logx_i) = E(logy_i|x_i) \neq logE(y_i|x_i)$.

Observe, por suposto, que 16 implica que no modelo linear

$$\frac{\partial E(y_i|x_i)}{\partial x_{ik}} . \frac{x_{ik}}{E(y_i|x_i)} = \frac{x_{ik}}{x_i' \beta} \beta_k \tag{12}$$

o que mostra que o modelo linear implica que elasticidades não são constantes e variam com x_i, enquanto modelos loglineares impõem elasticidades constantes.

Regressão Múltipla

Enquanto em muitos casos a escolha da forma funcional é baseada por conveniência na interpretação, outras considerações podem ser importantes. Por exemplo, explicar $log y_i$ ao invés de y_i pode ajudar a reduzir heterocedasticidade, isto é, variância não constante.

Há, ademais, outras possibilidades de se estimar formas funcionais com logaritmos, como

$$log y_i = x_i^{'} \beta + \varepsilon_i \tag{13}$$

Naturalmente, é possível ter um misto entre variáveis explicativas em *log* e outras em nível. Em 13, a interpretação do coeficiente β_k é baseada na mudança relativa em y_i dada uma mudança absoluta de uma unidade em x_{ik}. Isso é referido como **semi-elasticidade**.

A tabela abaixo traz uma interpretação mais geral para as diferentes **formas funcionais**.[4]

Modelo	*Variável Dependente*	*Variável Independente*	*Interpretação de β_1*
nível-nível	y	x	$\Delta y = \beta_1 \Delta x$
nível-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-nível	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Figure 2: Formas funcionais

[4]Ver Wooldridge [2013].

Efeito da poluíção no preço de imóveis

Para ilustrar a aplicação de elasticidades, vamos considerar outra discussão contida em Wooldridge [2013]. Suponha que tenhamos o modelo abaixo para explicar o preço de casas:

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 rooms + e_i$$

onde *nox* significa poluição e *rooms* é o número de quartos. O código abaixo baixa os dados e gera a regressão.

Regressão Múltipla

```r
library(foreign)
hprice2 <- read.dta('http://fmwww.bc.edu/ec-p/data/wooldridge/hprice2.dta')
reg2 <-lm(lprice ~ lnox + rooms, data=hprice2)
reg2
```

```
##
## Call:
## lm(formula = lprice ~ lnox + rooms, data = hprice2)
##
## Coefficients:
## (Intercept)         lnox        rooms
##      9.2337      -0.7177       0.3059
```

Regressão Múltipla

O coeficiente β_1 é, nesse contexto, a elasticidade do preço de casas em relação à poluição (*nox*), enquanto o coeficiente β_2 é a mudança no *log* do preço quando o número de quartos mudar em uma unidade. Ao multiplicarmos por 100, teremos a mudança percentual no preço, de forma aproximada. Assim, quando a poluição aumenta em 1%, o preço de casas se reduz em 0.72%, mantido o número de quartos fixos. Ademais, quando o número de quartos se eleva em uma unidade, os preços aumentam em 30.6%.

Regressão Múltipla

Observe, nesse contexto, que faz sentido colocar a poluição em termos percentuais, mas não faz sentido colocar o número de quartos nessa métrica.

Regressão Múltipla

O modelo linear

$$y_i = x_i^{'}\beta + \varepsilon_i \tag{14}$$

tem pouco significado a não ser que adicionemos algumas suposições a respeito de ε_i. É comum, nesse sentido, estabelecer que ε_i tem um valor esperado nulo e x_i é tomado como dado. Um modo formal de estabelecer isso é assumir que o valor esperado de ε_i dado x_i é zero, isto é,

$$E(\varepsilon_i|x_i) = 0 \tag{15}$$

Regressão Múltipla

Sob 15, a propósito, nós podemos interpretar o modelo linear descrito por 14 como o valor esperado de y_i dados os valores de x_i.[5] Por exemplo, qual o salário esperado para uma mulher aleatória de 40 anos com educação superior e 14 anos de experiência? Ou, qual a taxa de desemprego esperada dadas as taxas de salário, inflação e o produto total de uma economia? A primeira consequência de 15 é a interpretação individual dos coeficientes β.

[5] Seção baseada em Verbeek [2012] e Wooldridge [2013].

Regressão Múltipla

Por exemplo, β_k mede a mudança esperada em y_i se x_{ik} mudar em uma unidade mas todas as demais variáveis contidas em x_i permanecerem constantes.[6] Isto é,

$$\frac{\partial E(y_i|x_i)}{\partial x_{ik}} = \beta_k \tag{16}$$

Assim, se estamos interessados em ver a relação entre y_i e x_{ik}, as demais variáveis em x_i são chamadas de **variáveis de controle**.

[6] Essa última chamada de **condição ceteris paribus**.

Regressão Múltipla

Por exemplo, se estamos interessados em verificação a relação entre preço de imóveis e números de quartos, o tamanho do apartamento e a localização servem como controles para que consigamos verificar de forma mais precisa o que estamos interessados. A depender do nosso interesse, podemos *controlar* para alguns fatores e não para outros. Se, por exemplo, $x_i'\beta$ incluir $idade_i\beta_2 + idade_i^2\beta_3$, o efeito da *idade* sobre y_i será dada por

$$\frac{\partial E(y_i|x_i)}{\partial idade_i} = \beta_2 + 2idade_i\beta_3 \tag{17}$$

Regressão Múltipla

O que significa o impacto da idade em y_i, mantidas as demais variáveis constantes. A interpretação de 14 como esperança condicional, a propósito, não necessariamente implica que podemos interpretar os parâmetros em β como uma medida de efeito causal de x_i sobre y_i. Por exemplo, não é improvável que a taxa de salários esperada varie entre trabalhadores casados ou não casados, mesmo após controlarmos por por outros fatores, mas não é muito provável que casar *cause* maiores salários.

Efeito do cigarro no peso de recém-nascidos

Para ilustrar no **R**, a interpretação de coeficientes, considere o código abaixo, que traz uma discussão contida em Wooldridge [2013].

```r
library(foreign)
bwght = read.dta('http://fmwww.bc.edu/ec-p/data/wooldridge/bwght.dta')
```

Regressão Múltipla

Com efeito, considere o seguinte modelo:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + e_i$$

Onde *bwght* é o peso de recém-nascidos, medido em onças, *cigs* é o número médio de cigarros que a mãe fumou por dia durante a gravidez e *faminc* é a renda anual familiar, em milhares de dólares. Estimamos o modelo com o código abaixo.

Regressão Múltipla

```r
reg = lm(bwght ~ cigs + faminc, data=bwght)
reg
```

```
##
## Call:
## lm(formula = bwght ~ cigs + faminc, data = bwght)
##
## Coefficients:
## (Intercept)         cigs       faminc
##   116.97413     -0.46341      0.09276
```

Regressão Múltipla

Pelo modelo estimado, podemos inferir que se a mãe consumir 10 cigarros por dia, o peso esperado do bebê se reduzirá em 4.63 onças ou 131.33 gramas.

Modelos com funções quadráticas

Formas quadráticas podem ser adicionadas a um modelo para captar aumentos ou decaimentos marginais. Para ilustrar, vamos considerar o exemplo abaixo, utilizando o mesmo conjunto de dados para os preços de casas:

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + e_i$$

onde o número de quartos entra duas vezes agora e há outras variáveis de controle. O modelo é estimado abaixo.

Modelos com funções quadráticas

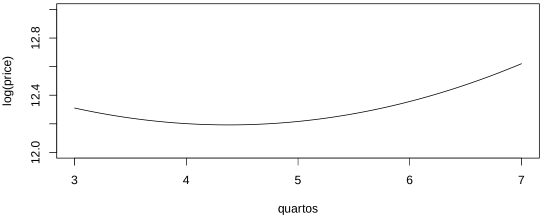
```r
ldist <- log(hprice2$dist)
rooms.sq <- hprice2$rooms^2
reg3 <- lm(lprice ~ lnox + ldist + rooms + rooms.sq + stratio, data=hprice2)
summary(reg3)
```

```
##
## Call:
## lm(formula = lprice ~ lnox + ldist + rooms + rooms.sq + stratio,
##     data = hprice2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04285 -0.12774  0.02038  0.12651  1.25272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.385480   0.566473  23.629  < 2e-16 ***
## lnox        -0.901683   0.114687  -7.862 2.34e-14 ***
## ldist       -0.086782   0.043281  -2.005  0.04549 *
## rooms       -0.545112   0.165454  -3.295  0.00106 **
## rooms.sq     0.062261   0.012805   4.862 1.56e-06 ***
## stratio     -0.047590   0.005854  -8.129 3.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2592 on 500 degrees of freedom
## Multiple R-squared:  0.6028, Adjusted R-squared:  0.5988
## F-statistic: 151.8 on 5 and 500 DF,  p-value: < 2.2e-16
```

Modelos com funções quadráticas

O coeficiente dos quartos é negativo e o coeficiente dos quartos ao quadrado é positivo o que implica que para valores baixos de quartos, um quarto adicional tem efeito negativo sobre o *log* dos preços. A certo ponto, porém, o efeito passa a ser positivo. A figura abaixo ilustra o efeito.

```
curve(reg3$coefficients[1]+reg3$coefficients[4]*x+
        reg3$coefficients[5]*x^2,xlim=c(3,7),
      ylim=c(12, 13), xlab='quartos', ylab="log(price)")
```



Modelos com funções quadráticas

A partir de 4.4 quartos, o efeito passa a ser positivo. Assim, para ilustrar, considere a mudança de cinco para seis quartos. O efeito no preço será dado de forma aproximada por

$$100 * ((-.545 + 2 * .062)\mathit{rooms}) * \Delta \mathit{rooms}$$

Isto é, de cinco para seis quartos, o preço aumenta em 7.5%. Já o aumento de seis para sete quartos é de 19.9%.

Variáveis dummies

Vamos retomar agora o nosso `dataset` **CPS1988**.

```r
library(AER)
data("CPS1988")
summary(CPS1988)
```

```
##       wage            education       experience   ethnicity
##  Min.   :   50.05   Min.   : 0.00   Min.   :-4.0   cauc:25923
##  1st Qu.:  308.64   1st Qu.:12.00   1st Qu.: 8.0   afam: 2232
##  Median :  522.32   Median :12.00   Median :16.0
##  Mean   :  603.73   Mean   :13.07   Mean   :18.2
##  3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:27.0
##  Max.   :18777.20   Max.   :18.00   Max.   :63.0
##   smsa            region     parttime
##  no : 7223   northeast:6441   no :25631
##  yes:20932   midwest  :6863   yes: 2524
##              south    :8760
##              west     :6091
##
##
```

Variáveis dummies

Observe que o nível *cauc* da variável **ethnicity** não aparece no output da regressão. Há apenas um *efeito étnico*, dando a diferença entre os grupos *afam* e *cauc*. Isto é, o quanto os afro-americanos ganham a mais ou a menos do que o grupo de referência.

Como estamos lidando com um modelo semilogarítmico, é automático que se multiplique por 100 o coeficiente β_5. Mas isso não é correto, como pode ser visto em Halvorsen and Palmquist [1980]. A interpretação correta será fazer $(exp(\beta) - 1) * 100$.

Variáveis dummies

Para o nosso caso, temos uma mudança no salário semanal de -21.6% quando consideramos afro-americanos em comparação ao grupo de controle.

A função I

Quando rodamos o nosso modelo, nós emulamos o quadrado daa variável **experience** com a função `I`, isso porque os operadores :, *, /, ^ têm significados especiais quando dentro da função `lm`. Para que tenham, portanto, o significado real, precisamos colocá-los dentro da função `I`.

Exercícios

Salários

```r
library(wooldridge) # abrimos o pacote
data("wage1") # puxamos os dados
str(wage1, max.level=1) # averiguamos a estrutura
```

```
## 'data.frame':    526 obs. of  24 variables:
##  $ wage    : num  3.1 3.24 3 6 5.3 ...
##  $ educ    : int  11 12 11 8 12 16 18 12 12 17 ...
##  $ exper   : int  2 22 2 44 7 9 15 5 26 22 ...
##  $ tenure  : int  0 2 0 28 2 8 7 3 4 21 ...
##  $ nonwhite: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ female  : int  1 1 0 0 0 0 0 1 1 0 ...
##  $ married : int  0 1 0 1 1 1 0 0 0 1 ...
##  $ numdep  : int  2 3 2 0 1 0 0 0 2 0 ...
##  $ smsa    : int  1 1 0 1 0 1 1 1 1 1 ...
##  $ northcen: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ south   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ west    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ construc: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ndurman : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trcommpu: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trade   : int  0 0 1 0 0 0 1 0 1 0 ...
##  $ services: int  0 1 0 0 0 0 0 0 0 0 ...
##  $ profserv: int  0 0 0 0 0 1 0 0 0 0 ...
##  $ profocc : int  0 0 0 0 0 1 1 1 1 1 ...
##  $ clerocc : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ servocc : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ lwage   : num  1.13 1.18 1.1 1.79 1.67 ...
##  $ expersq : int  4 484 4 1936 49 81 225 25 676 484 ...
##  $ tenursq : int  0 4 0 784 4 64 49 9 16 441 ...
##  - attr(*, "datalabel")= chr ""
```

Exercícios

Estimando uma função para o log do salário-hora temos os parâmetros dos retornos percentuais de cada entrada no modelo. Podemos avaliar se, por exemplo, depois de controlar por educação e titularidade, experiência ainda tem um efeito estatisticamente significante no salário-hora.

```r
summary(lm(log(wage) ~ educ + exper + tenure, data=wage1))
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + tenure, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05802 -0.29645 -0.03265  0.28788  1.42809
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.284360   0.104190   2.729  0.00656 **
## educ        0.092029   0.007330  12.555  < 2e-16 ***
## exper       0.004121   0.001723   2.391  0.01714 *
## tenure      0.022067   0.003094   7.133 3.29e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4409 on 522 degrees of freedom
## Multiple R-squared:  0.316,  Adjusted R-squared:  0.3121
## F-statistic: 80.39 on 3 and 522 DF,  p-value: < 2.2e-16
```

Exercícios

E de fato, a 5% de significância existe um efeito para experiência. Mais especificamente, um ano a mais de experiência na média se traduz em $0,41\%$ de aumento salarial.

Exercícios

Notas de alunos e tamanho da escola

Existe um certo debate em economia da educação sobre o efeito do tamanho de uma escola sobre a performance dos alunos. É possível que o maior número de interações ou que o ganho de escala leve a uma educação de mais qualdiade, por exemplo. Há quem argumente que o número maior de alunos impede um certo cuidado especial com cada estudante, diminuindo a performance.

Carregamos uma base de dados com notas de escolas no estado americano do Michigan do ano de 1993. Vamos testar a hipótese nula de que o tamanho da escola tem efeito zero sobre as notas de seus alunos em testes padronizados. Vamos tentar explicar as notas pelos salários dos professores, número de funcionários por mil alunos e número de matrículas.

Exercícios

```r
data("meap93")
summary(lm(math10 ~ salary + staff + enroll, data = meap93))
```

```
##
## Call:
## lm(formula = math10 ~ salary + staff + enroll, data = meap93)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.214  -7.023  -0.863   5.974  41.755
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4259135  6.0562311   0.070    0.944
## salary       0.0005989  0.0001192   5.026 7.53e-07 ***
## staff        0.0529116  0.0396184   1.336    0.182
## enroll      -0.0002530  0.0002152  -1.176    0.240
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.19 on 404 degrees of freedom
## Multiple R-squared:  0.06371,    Adjusted R-squared:  0.05676
## F-statistic: 9.163 on 3 and 404 DF,  p-value: 7.047e-06
```

Exercícios

O parâmetro estimado para a nossa proxy de tamanho da escola é negativo, o que a primeira vista sugere que maiores escolas. No entanto, podemos ter estimado um coeficiente diferente de zero por erro de amostragem.

Queremos testar a hipótese de que $\beta_{enroll} \neq 0$, apesar de que claramente $\hat{\beta} \neq 0$. Para isso usamos a estatística t do parâmetro, que a table nos informa ser $-1,176$. No entanto, o valor crítico da distribuição t com 404 graus de liberdade (que a tabela de regressão nos informa) é $-1,65$. Como a estatística t do parâmetro estimado é *menor* do que o valor crítico, não conseguimos rejeitar a hipótese nula de que o tamanho da escola não afeta as notas. Curiosamente, a razão funcionários para cada mil alunos também não, embora salários de professores tenham um altíssimo nível de significância.

Exercícios

Notas no ensino superior

Podemos sair do ambiente escolar e procurar os determinantes de performance no ensino superior. Será que alunos que faltam mais vão realmente pior? Para isso, vamos construir um modelo que relacione o Coeficiente de Rendimento Acumulado (GPA) ao coeficiente de rendimento do ensino médio, nota no ACT (uma espécie de ENEM americano) e número de aulas faltadas. Por fim, vamos estimar os parâmetros com uma base de dados com 141 alunos.

Exercícios

```r
data("gpa1")
(sumres <- summary(lm(colGPA ~ hsGPA + ACT + skipped, data = gpa1)))
```

```
##
## Call:
## lm(formula = colGPA ~ hsGPA + ACT + skipped, data = gpa1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85698 -0.23200 -0.03935  0.24816  0.81657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.38955    0.33155   4.191 4.95e-05 ***
## hsGPA        0.41182    0.09367   4.396 2.19e-05 ***
## ACT          0.01472    0.01056   1.393  0.16578
## skipped     -0.08311    0.02600  -3.197  0.00173 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3295 on 137 degrees of freedom
## Multiple R-squared:  0.2336, Adjusted R-squared:  0.2168
## F-statistic: 13.92 on 3 and 137 DF,  p-value: 5.653e-08
```

Exercícios

```
## confirmando manualmente
regtabela <- sumres$coefficients
bhat <- regtabela[,1]
se   <- regtabela[,2]
## reproduzindo a estatística t
(tstat <- bhat / se)
```

```
## (Intercept)       hsGPA         ACT     skipped
##    4.191039    4.396260    1.393319   -3.196840
```

```
# reproduzind o p-valor
(pval  <- 2*pt(-abs(tstat), 137))
```

```
##  (Intercept)        hsGPA          ACT      skipped
## 4.950269e-05 2.192050e-05 1.657799e-01 1.725431e-03
```

Aparentemente, existe um efeito estatísticamente significante e negativo entre faltar aulas e notas.

Exercícios

Crimes do Campus e Matrículas

Considere um modelo simples em que o número de crimes em um campus (C) é explicado por uma constante e o número de matrículas (M). Vamos explicita-lo na forma log-log porque isso faz os parâmetros serem interpretados como elasticidades:

$$\log(C) = \beta_0 + \beta_1 \log(M) + u$$

Até agora testamos hipóteses nulas em que o verdadeiro parâmetro, na população, é zero. No entanto, isso talvez não seja interessante num contexto de elasticidade. Talvez queiramos saber se temos elasticidade unitária, maior que 1 ou menor, por exemplo.

Exercícios

```r
data(campus)
summary(lm(lcrime ~ lenroll , data = campus))
```

```
##
## Call:
## lm(formula = lcrime ~ lenroll, data = campus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5136 -0.3858  0.1174  0.4363  2.5782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.6314     1.0335  -6.416 5.44e-09 ***
## lenroll       1.2698     0.1098  11.567  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8946 on 95 degrees of freedom
## Multiple R-squared:  0.5848, Adjusted R-squared:  0.5804
## F-statistic: 133.8 on 1 and 95 DF,  p-value: < 2.2e-16
```

Exercícios

Estimamos um parâmetro maior que 1, mas será que o verdadeiro parâmetro é maior? Podemos ter estimado um parâmetro dessa magnitude por erro de amostragem, por exemplo. Podemos testar isso.

```
# calculando a estatística t para a hipótese nula
# usamos o erro padrão oferecido na tabela de regressão
t = (1.27 - 1 )/ 0.109
t
```

```
## [1] 2.477064
```

Como o valor crítico da distribuição t a 5% de significância com 95 graus de liberdade é cerca de $1,66$, podemos seguramente rejeitar a hipótese nula com 95% de confiança.

Exercícios

Preços de casas e poluição

Com uma amostra de dados imobiliários de Boston, iremos estimar um modelo para explicar preços de casas em função de algumas características locais como distância a centros de empergo, professores por aluno nas escolas próximas, número de cômodos e poluição, medida em partes de óxido nitroso por milhão no ar.

Exercícios

```r
data("hprice2")
summary(lm(lprice ~ lnox + rooms + log(dist) + stratio, data  = hprice2))
```

```
##
## Call:
## lm(formula = lprice ~ lnox + rooms + log(dist) + stratio, data = hprice2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05890 -0.12427  0.02128  0.12882  1.32531
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.083862   0.318111  34.843  < 2e-16 ***
## lnox        -0.953539   0.116742  -8.168 2.57e-15 ***
## rooms        0.254527   0.018530  13.736  < 2e-16 ***
## log(dist)   -0.134339   0.043103  -3.117  0.00193 **
## stratio     -0.052451   0.005897  -8.894  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.265 on 501 degrees of freedom
## Multiple R-squared:  0.584,  Adjusted R-squared:  0.5807
## F-statistic: 175.9 on 4 and 501 DF,  p-value: < 2.2e-16
```

Exercícios

Repetindo o raciocínio do exemplo anterior, talvez queiramos testar se elasticidade agora não é -1, por exemplo. A estatística t do parâmetro `lnox` fica então:

```
t = (-.9535 -(-1))/0.1167
t
```

```
## [1] 0.3984576
```

A estatística t do parâmetro é definitivamente menor que o valor crítico a 5% de significância e 501 graus de liberdade (que é maior do que 1). Não temos evidências estatísticas para apoiar a tese de que a elasticidade é diferente de -1.

Exercícios

Participação em fundos de pensão

Vamos explicar a taxa de participação de funcionários de empresas em fundos de pensão com um modelo que leva em conta, número de empregados, tamanho da empresa, idade média dos funcionários. O modelo segue:

Exercícios

```r
data("k401k")
summary(lm(prate ~ mrate + age + totemp, data = k401k))
```

```
##
## Call:
## lm(formula = prate ~ mrate + age + totemp, data = k401k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.698  -8.074   4.716  12.505  30.307
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.029e+01  7.777e-01 103.242  < 2e-16 ***
## mrate        5.442e+00  5.244e-01  10.378  < 2e-16 ***
## age          2.692e-01  4.514e-02   5.963 3.07e-09 ***
## totemp      -1.291e-04  3.666e-05  -3.521 0.000443 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.88 on 1530 degrees of freedom
## Multiple R-squared:  0.09954,    Adjusted R-squared:  0.09778
## F-statistic: 56.38 on 3 and 1530 DF,  p-value: < 2.2e-16
```

Exercícios

Observe que o tamanho da empresa (medido por `totemp`) é estatisticamente significante. No entanto, o parâmetro estimado não é *relevante*. Sua magnitude é de aproximadamente $0,00013$. Embroa consigamos prover evidências de que é diferente de zero, o parâmetro não é muito relevante.

Exercícios

Treinamentos

A taxa de rejeição de uma firma industrial é a quantidade de produtos descartados a cada 100 produzidos. Podemos avaliar se treinar funcionários ajuda a diminuir esse indicador.

```r
data("jtrain")
summary(lm(scrap ~ hrsemp + lsales + lemploy, data = jtrain))
```

```
##
## Call:
## lm(formula = scrap ~ hrsemp + lsales + lemploy, data = jtrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3763 -3.2215 -1.6099  0.6103 25.4602
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.55314   10.50698   1.290   0.1994
## hrsemp      -0.01279    0.01769  -0.723   0.4709
## lsales      -1.02299    0.83909  -1.219   0.2250
## lemploy      1.58473    0.80867   1.960   0.0522 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.408 on 131 degrees of freedom
##   (336 observations deleted due to missingness)
## Multiple R-squared:  0.03813,    Adjusted R-squared:  0.0161
## F-statistic: 1.731 on 3 and 131 DF,  p-value: 0.1638
```

Exercícios

A variável `hrsemp` é o tempo de treinamento por trabalhador e ela definitivamente não é significante nas margens aceitáveis. Pelo contrário, seu p-valor é maior do que 45%.

Exercícios

Pesquisa e Desenvolvimento

Uma pergunta comum em economia industrial é se existe ligação entre o tamanho de uma firma e seu gasto com pesquisa, e vice-versa. O seguinte modelo pode ajudar a entender isso.

Exercícios

```r
data(rdchem)
# OLS regression:
reg <- lm(log(rd) ~ log(sales)+profmarg, data = rdchem)
# saída da regressão:
summary(reg)
```

```
##
## Call:
## lm(formula = log(rd) ~ log(sales) + profmarg, data = rdchem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97681 -0.31502 -0.05828  0.39020  1.21783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.37827    0.46802  -9.355 2.93e-10 ***
## log(sales)   1.08422    0.06020  18.012  < 2e-16 ***
## profmarg     0.02166    0.01278   1.694    0.101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5136 on 29 degrees of freedom
## Multiple R-squared:  0.918,  Adjusted R-squared:  0.9123
## F-statistic: 162.2 on 2 and 29 DF,  p-value: < 2.2e-16
```

Exercícios

```r
# intervalo de confiança a 95%:
confint(reg)
```

```
##                    2.5 %     97.5 %
## (Intercept) -5.335478450 -3.4210681
## log(sales)   0.961107256  1.2073325
## profmarg    -0.004487722  0.0477991
```

```r
# intervalo de confiança a 99%:
confint(reg, level=0.99)
```

```
##                   0.5 %      99.5 %
## (Intercept) -5.66831270 -3.08823382
## log(sales)   0.91829920  1.25014054
## profmarg    -0.01357817  0.05688955
```

Embora a margem de lucro não seja estatisticamente significante para explicar gasto em pesquisa das firmas da amostra, vemos que a elasticidade vendas-pesquisa é estatisticamente significante e de magnitude relevante.

Exercícios

Usando a base de dados `Cars93`, do pacote `MASS`, estime um modelo de regressão linear para explicar a variável `MPG.highway` em função da variável `Horsepower` e estime o `MPG.highway` de um carro com 150 cavalos de potência.

Exercícios

É normal tentar estimar a altura adulta de uma criança dobrando sua altura aos dois anos de idade. A seguinte tabela relaciona os dois, em polegadas:

Altura aos 2 anos de idade	39	30	32	34	35	36	36	30
Altura Adulta	71	63	63	67	68	68	70	64

Usando esses dados, é possível dizer que a ideia de que dobrar a altura é uma estimativa razoável?

Exercícios

A base de dados `homedata` do pacote `UsingR` contém dados imobiliários. Supõe que com o tempo imóveis tendem a valorizar na maioria dos lugares, a medida que bairros melhoram. Regrida o preço de uma casa no ano 2000 pelo preço nos anos 70 e responda se é verdade que preços de lares tendem a valorizar.

Exercícios

Encontre e gere um gráfico com a linha de regressão do modelo `lm(maxrate ~ age, data = heartrate)`. A base de dados `heartrate` está disponível no pacote `UsingR`.

Exercícios

Abra a base de dados `babies` do pacote `UsingR` e estime um modelo para explicar o peso da uma criança ao nascer com as variáveis `gestation`, `age`, `h`, `wt1`, `dage`, `dht` e `dwt`.

Comparação de Modelos

Podemos, agora, utilizar a função `anova` para comparar o modelo acima com outro que não leva a variável **ethnicity** em consideração.

```
cps_noeth <- lm(log(wage) ~ experience + I(experience^2) +
                  education, data = CPS1988)
anova(cps_noeth, cps_lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ experience + I(experience^2) + education
## Model 2: log(wage) ~ experience + I(experience^2) + education + ethnicity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  28151 9719.6
## 2  28150 9598.6  1    121.02 354.91 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O output da função mostra que o efeito da variável **ethnicity** é de fato significativo.

Modelos Parcialmente Lineares

Considere o seguinte modelo

$$log(wage) = \beta_1 + g(experience) + \beta_2 education + \beta_3 ethnicity + \varepsilon \tag{18}$$

Onde, g é uma função desconhecida a ser estimada a partir do nosso `dataset` a partir de uma regressão *splines*. O código abaixo ilustra.

```
library(splines)
cps_plm <- lm(log(wage) ~ bs(experience, df = 5) +
                education + ethnicity, data = CPS1988)
```

Modelos Parcialmente Lineares

```r
summary(cps_plm)
```

```
##
## Call:
## lm(formula = log(wage) ~ bs(experience, df = 5) + education +
##     ethnicity, data = CPS1988)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9315 -0.3079  0.0565  0.3672  3.9945
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.775582   0.056081   49.49   <2e-16 ***
## bs(experience, df = 5)1  1.891673   0.075814   24.95   <2e-16 ***
## bs(experience, df = 5)2  2.259468   0.046474   48.62   <2e-16 ***
## bs(experience, df = 5)3  2.824582   0.070773   39.91   <2e-16 ***
## bs(experience, df = 5)4  2.373082   0.065205   36.39   <2e-16 ***
## bs(experience, df = 5)5  1.739341   0.119691   14.53   <2e-16 ***
## education                0.088181   0.001258   70.07   <2e-16 ***
## ethnicityafam           -0.248202   0.012725  -19.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5747 on 28147 degrees of freedom
## Multiple R-squared:  0.3557, Adjusted R-squared:  0.3555
## F-statistic:  2220 on 7 and 28147 DF,  p-value: < 2.2e-16
```

A expressão $bs(experience, df = 5)$ irá gerar internamente os regressores pertinentes.

Modelos Parcialmente Lineares

```r
cps <- data.frame(experience = -2:60, education =
                    with(CPS1988, mean(education[ethnicity ==
                                                   "cauc"])),
                  ethnicity = "cauc")
cps$yhat1 <- predict(cps_lm, newdata = cps)
cps$yhat2 <- predict(cps_plm, newdata = cps)
plot(log(wage) ~ jitter(experience, factor = 3), pch = 19,
     col = rgb(0.5, 0.5, 0.5, alpha = 0.02), data = CPS1988)
lines(yhat1 ~ experience, data = cps, lty = 2)
lines(yhat2 ~ experience, data = cps)
legend("topleft", c("quadratic", "spline"), lty = c(2,1),
       bty = "n")
```

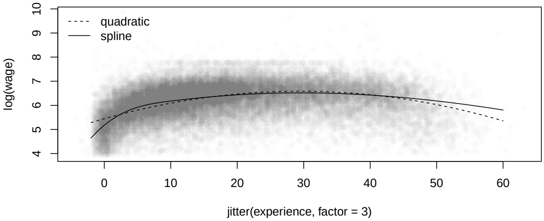


Figure 3: Comparando modelos

Modelos Parcialmente Lineares

Não há muita diferença entre os 20 e 40 anos de experiência entre os modelos. A diferença mais pronunciada fica no início do intervalo da experiência em relação ao seu efeito sobre os salários. O modelo **spline** apresenta *curvaturas*.

Fatores e interações

Em economia do trabalho existem muitos exercícios que tentam identificar algum tipo de discriminação. Por exemplo, de gênero ou etnia. Esse tipo de trabalho, como vimos na equação 9 envolve estimar modelos com variáveis que são fatores ou mesmo interações. Podemos, ademais, construir um modelo mais geral a partir do nosso dataset `CPS1988`.

Podemos estar interessados, por exemplo, na interação da variável binária **ethnicity** com as demais variáveis do nosso dataset. O código abaixo dá um exemplo.

```
cps_int <- lm(log(wage) ~ experience + I(experience^2) +
                education * ethnicity, data = CPS1988)
```

Fatores e interações

```r
coeftest(cps_int)
```

```
##
## t test of coefficients:
##
##                          Estimate  Std. Error  t value Pr(>|t|)
## (Intercept)             4.3131e+00  1.9590e-02 220.1703  < 2e-16 ***
## experience              7.7520e-02  8.8028e-04  88.0625  < 2e-16 ***
## I(experience^2)        -1.3179e-03  1.9006e-05 -69.3388  < 2e-16 ***
## education               8.6312e-02  1.3089e-03  65.9437  < 2e-16 ***
## ethnicityafam          -1.2389e-01  5.9026e-02  -2.0989  0.03584 *
## education:ethnicityafam -9.6481e-03  4.6510e-03  -2.0744  0.03805 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observe que o termo $education * ethnicity$ especifica a inclusão de três termos na nossa regressão.

Mínimos Quadrados Ponderados

Regressões *cross-section* são usualmente contaminadas por problemas de heterocedasticidade. Vamos aprender a diagnoticar esse tipo de problema mais à frente em nosso curso. Por enquanto, vamos ver o método de mínimos quadrados ponderados, de forma a lidar com esse tipo de questão.

Para ilustrar, vamos considerar novamente o dataset `Journals` e o exemplo que fizemos. . .

Mínimos Quadrados Ponderados

```
data("Journals")
journals <- Journals[, c("subs", "price")]
journals$citeprice <- Journals$price/Journals$citations
jour_lm <- lm(log(subs) ~ log(citeprice), data = journals)
plot(resid(jour_lm))
```

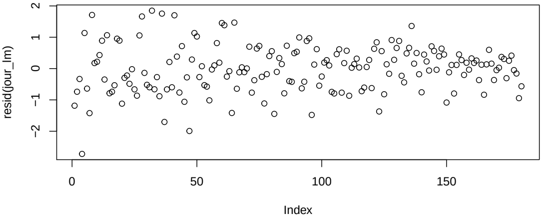


Figure 4: Resíduos do modelo

Mínimos Quadrados Ponderados

Uma possível solução para remediar o problema de heterocedasticidade é especificar um modelo de heterocedasticidade condicional tal qual

$$E(\varepsilon_i^2|x_i, z_i) = g({z_i}^T\gamma),$$

onde g é uma função não-linear que toma apenas valores positivos, z_i é um vetor contendo observações de variáveis exógenas e γ é um vetor de parâmetros.

Mínimos Quadrados Ponderados

Aproveitando o nosso modelo anterior, podemos dizer que nosso **preço por citação** seja nossa variável z_i. Lembre-se, por suposto, que assumir $E(\varepsilon_i^2|x_i, z_i) = \sigma^2 z_i^2$ nos leva a uma regressão de y_i/z_i sobre $1/z_i$ e x_i/z_i. Isso implica que o critério de estimação muda de $\sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2$ para $\sum_{i=1}^n z_i^{-2}(y_i - \beta_1 - \beta_2 x_i)^2$, isto é, cada termo agora é ponderado por z_i^{-2}, de modo que as soluções $\hat{\beta}_1, \hat{\beta}_2$ para o novo problema de minimização são chamadas de estimativas de mínimos quadrados ponderados, um caso especial de mínimos quadrados generalizados.

Mínimos Quadrados Ponderados

```
jour_wls1 <- lm(log(subs) ~ log(citeprice), data = journals,
                weights = 1/citeprice^2)
plot(resid(jour_wls1))
```

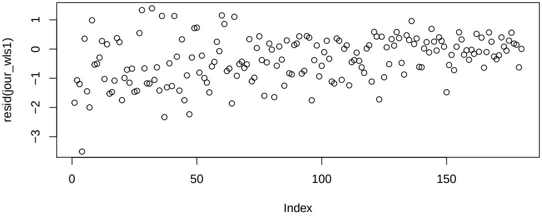


Figure 5: Resíduos do modelo

Mínimos Quadrados Ponderados

```
jour_wls2 <- lm(log(subs) ~ log(citeprice), data = journals,
                weights = 1/citeprice)
plot(log(subs) ~ log(citeprice), data = journals)
abline(jour_lm)
abline(jour_wls1, lwd = 2, lty = 2)
abline(jour_wls2, lwd = 2, lty = 3)
legend("bottomleft", c("OLS", "WLS1", "WLS2"),
       lty = 1:3, lwd = 2, bty = "n")
```

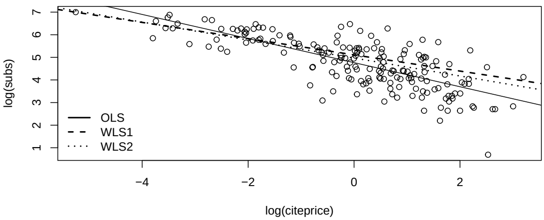


Figure 6: Comparando os métodos

Análise de Variância

Análise de Variância, ou *ANOVA*, é um método de comparar médias através de amostras baseado nas variações das médias.

```
anova(jour_lm)
```

```
## Analysis of Variance Table
##
## Response: log(subs)
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## log(citeprice)   1 125.93 125.934  224.04 < 2.2e-16 ***
## Residuals      178 100.06   0.562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A tabela ANOVA quebra a soma dos quadrados sobre a média (para a variável dependente) em duas partes: uma parte que é contabilizada por uma função linear do $log(citeprice)$ e uma parte atribuída à variação do resíduo.

Análise de Variância de um sentido
Uma análise de variância de um sentido é uma generalização do teste t para duas amostras independentes, nos permitindo comparar médias populacionais de várias amostras independentes. Suponha que temos k populações de interesse e de cada uma destas tirados uma amostra aleatória. Vamos notar que para a i-ésima amostra, x_{in} será o n-ésimo elemento dessa amostra.

Análise de Variância

Suponha que a média da i-ésima populaçõ é μ_i e seu desvio-padrão é σ_i - que será simplesmente σ se o desvio-padrão for consistente entre os grupos. Um modelo estatístico para os dados com um desvio-padrão comum seria:

$$x_{ij} = \mu_i + \epsilon_{ij} \tag{19}$$

Análise de Variância

onde os termos de erro ϵ_{ij} são independentes, normalmente distribuídos com média zero e variância σ^2.

Se quisermos testar se várias amostras têm uma mesma média, podemos considerar o modelo linear apresentado. Ao estima-lo, teremos SQT e SQR, dos quais podemos construir uma estatística que já vimos, a F:

$$F = \frac{SQT/(k-1)}{SQR/(n-k)} \sim F_{(k-1),(n-K)} \tag{20}$$

Análise de Variância

Como conhecemos a distribuição F com $(k-1)$ e $(n-K)$ graus de liberdade, podemos realizar um teste de hipótese para as médias de cada amostra, em que a hipótese nula é de que são todas iguais e a alternativa alguma negativa disso. A função oneway.test implementa esse teste no R.

É conveniente usar fatores para fazer esses testes. Se armazenamos a variável que indica em qual das i amostras está a observação como um fator "f", então podemos especificar o teste como $x \sim f$ que o R interpretará isso corretamente. Uma outra função para implementar análise de variância é "aov".

Análise de Variância

Um exemplo: Pesquisadores da Montana State University realizaram um estudo sobre como os vários tipos de esqui afetam o desempenho no esqui cross-country. Existem três básicos: clássico, moderno e integrado. Suponha que 9 esquiadores sejam designados em aleatório para os três tipos de aderência e para cada um, o esquiador tem sua força no tronco superior medida. Podemos investigar a hipótese nula de que os três tipos produzirão médias iguais com uma análise de variância. Nós assumimos que os erros são todos independentes e que os dados são amostrados a partir de populações normalmente distribuídas com variância comum, mas talvez médias diferentes.

O teste não-paramétrico de Kruskal-Wallis

O teste da soma de postos de Wilcoxon foi discutido como uma alternativa não-paramétrica para o teste t de duas amostras para amostras independentes. Embora não fizéssemos suposições sobre os parâmetros da população, assumimos que elas tinham densidades de mesma forma funcional e talvez centros diferentes. O teste de Kruskal-Wallis, um teste não-paramétrico, é análogo ao de Wilcoxson para comparar as médias populacionais de k amostras independentes. Em particular, se $f(x)$ é uma densidade de uma variável aleatória contínua com com média 0, supomos que x_{ij} são tiradas independentemente dos outros, de uma população com densidade $f(x - \mu_i)$. As hipóteses testadas são $H_0 : \mu_1 = \mu_2 = ... = \mu_i$ e $H_1 : \mu_i \neq \mu_j$ para algum par de amostras i e j.

Análise de Variância

A estatística de teste envolve o posto de todos os dados. Seja r_{ij} o respectivo posto de uma observação quando todos os dados são classificados do menor para o maior, $\bar{r}_i$ a média do posto de cada grupo, e $\bar{r}$ a média total. A estatística de teste é:

$$T = \frac{12}{n(n+1)} \sum_i n_i (\bar{r}_i - \bar{r})^2 \sim \chi^2_{k-1} \tag{21}$$

Como essa estatística tem distribuição conhecida, uma chi-quadrado com $k-1$ graus de liberdade, podemos usa-la para testes de hipótese.

Análise de Variância

Comparando múltiplas diferenças

Quando a análise de variância é executada com a função "lm", a saída do resumo exibe inúmeros testes estatísticos. O teste F realizado é para a hipótese nula de que $\beta_2 = \beta_3 = ůůů = \beta_k = 0$ contra uma alternativa que um ou mais parâmetros diferem de 0. Ou seja, que uma ou mais das variáveis tem efeitos de tratamento em comparação com o nível de referência. Os testes t marginais que são executados são testes de dois lados com uma hipótese nula de que $\beta_i = \beta_1$, cada um é feito para $i = 2, 3, ..., k$. Estes testam se algum dos tratamentos adicionais tem um efeito de tratamento quando controlado pelas outras variáveis.

Análise de Variância

No entanto, podemos querer fazer outras perguntas sobre os vários parâmetros. Por exemplo, comparações que não são informadas por padrão são testes mais específicos como β_2 e β_3 diferem? e β_1 e β_2 são metade de β_3?. Vamos avaliar agora diferentes múltiplas de parâmetros.

Análise de Variância

Se sabemos de antemão que estamos procurando uma diferença entre dois parâmetros, então um teste t simples é apropriado (como no caso em que estamos considerando apenas duas amostras independentes). No entanto, se olharmos para os dados e depois decidirmos para testar se o segundo e terceiro parâmetros diferem, então o nosso teste t é instável. Por quê? Lembre-se de que qualquer teste está correto apenas com alguma probabilidade, mesmo que os modelos estejam corretos. Isso significa que às vezes eles falham e quanto mais testes realizamos, mais provavelmente um ou mais falhará.

Análise de Variância

Podemos, por exemplo, nos perguntar se linhas aéreas diferentes estão sujeitas a tempos diferentes de espera em um mesmo aeroporto. Estaríamos comparando dois parâmetros entre si e com o a *opção nula* de que ambos sejam na verdade 0.

ANCOVA

Análise de Covariância (ANCOVA) é o nome dado aos modelos em que tanto variáveis categóricas quanto numéricas são usadas como preditoras. Também rodamos ANCOVAs com a função "lm". Para comparar a performance de dois modelos dessa maneira, precisamos estimar dois modelos lineares, salva-los como objetos no R e depois alimentá-los à função "anova".

W.H. Greene. *Econometric Analysis*. Pearson Education, 2003.

Robert Halvorsen and Raymond Palmquist. The Interpretation of Dummy Variables in Semilogarithmic Equations. *American Economic Review*, 70(3):474–475, June 1980. URL https://ideas.repec.org/a/aea/aecrev/v70y1980i3p474-75.html.

J. H. Stock and M. W. Watson. *Introduction to Econometrics*. Pearson Education, 2007.

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

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Editora Cengage, 2013.