Modelos de Regressão e Previsão

Gabarito da Lista 1

Prof. Carlos Trucíos
carlos.trucios@facc.ufrj.br
ctruciosm.github.io

Questão 1

Sejam $\hat{\beta}_0$ e $\hat{\beta}_1$ os estimadores MQO da regressão y sobre x. Mostre que os estimadores MQO da regressão cy sobre x são $c\hat{\beta}_0$ e $c\hat{\beta}_1$, respectivamente.

Resposta

Sabemos que os estimadores MQO da regressão y sobre x são:

$$\hat{\beta}_1 = \frac{\text{cov}(x,y)}{\sqrt{V(x)}} \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Denotemos por $\tilde{\beta}_1$ e $\tilde{\beta}_0$ os estimadores MQO da regressão cy sobre x. Então, os estimadores MQO da regressão cy sobre x são:

$$\tilde{\beta}_1 = \frac{\text{cov}(x,cy)}{\sqrt{V(x)}} \quad \text{e} \quad \tilde{\beta}_0 = \overline{cy} - \tilde{\beta}_1\bar{x}$$

Sabemos que:

- $\text{cov}(ax+b, cy+d) = a \ c \ \text{cov}(x,y)$, então

$$\text{cov}(x,cy) = c\,\text{cov}(x,y)$$

- $\overline{cy} = c\bar{y}$

Então:

$$\tilde{\beta}_1 = c \ \underbrace{\frac{\text{cov}(x,y)}{\sqrt{V(x)}}}_{\hat{\beta}_1} \quad \text{e} \quad \tilde{\beta}_0 = c \ \bar{y} - \underbrace{\tilde{\beta}_1}_{c \ \hat{\beta}_1} \bar{x} = c \ \underbrace{(\bar{y} - \hat{\beta}_1\bar{x})}_{\hat{\beta}_0}$$

Assim, os estimadores MQO da regressão cy sobre x são:

$$\tilde{\beta}_1 = c\hat{\beta}_1 \quad \text{e} \quad = \tilde{\beta}_0 = c\hat{\beta}_0$$

Questão 2

Utilizando o software R, rode o seguinte código:

```r
library(MASS)
beta0 = 0.2
beta1 = 0.5
mSigma = matrix(c(1,0,0,1),2)
vMu = c(0,0)
dados_simulados = mvrnorm(2000, mu = vMu, Sigma = mSigma)
x = dados_simulados[,1]
u = dados_simulados[,2]
y = beta0 + beta1*x + u
lm(y~x)
```

Resposta

a. Os valores de $\hat{\beta}$ são próximos dos valores de β?

```r
library(MASS)
beta0 = 0.2
beta1 = 0.5
mSigma = matrix(c(1,0,0,1),2)
vMu = c(0,0)
set.seed(123)
dados_simulados = mvrnorm(2000, mu = vMu, Sigma = mSigma)
x = dados_simulados[,1]
u = dados_simulados[,2]
y = beta0 + beta1*x + u
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      0.2291       0.5132
```

Sim, são próximos

b. E se mSigma = matrix(c(1,0.3,0.3,1),2)?

```r
library(MASS)
beta0 = 0.2
beta1 = 0.5
mSigma = matrix(c(1,0.3,0.3,1),2)
vMu = c(0,0)
set.seed(123)
dados_simulados = mvrnorm(2000, mu = vMu, Sigma = mSigma)
x = dados_simulados[,1]
u = dados_simulados[,2]
y = beta0 + beta1*x + u
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
```

```
##      0.2049       0.8102
```

$\hat{\beta}_1$ ficou afastado de β_1

c. E se mSigma = matrix(c(1,-0.5,-0.5,1),2)?

```r
library(MASS)
beta0 = 0.2
beta1 = 0.5
mSigma = matrix(c(1,-0.5,-0.5,1),2)
vMu = c(0,0)
set.seed(123)
dados_simulados = mvrnorm(2000, mu = vMu, Sigma = mSigma)
x = dados_simulados[,1]
u = dados_simulados[,2]
y = beta0 + beta1*x + u
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.22335     -0.01733
```

$\hat{\beta}_1$ ficou mais afastado de de β_1, inclussive, ficou negativo

d. E se mSigma = matrix(c(1,0.7,0.7,1),2)?

```r
library(MASS)
beta0 = 0.2
beta1 = 0.5
mSigma = matrix(c(1,0.7,0.7,1),2)
vMu = c(0,0)
set.seed(123)
dados_simulados = mvrnorm(2000, mu = vMu, Sigma = mSigma)
x = dados_simulados[,1]
u = dados_simulados[,2]
y = beta0 + beta1*x + u
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      0.1984       1.2014
```

$\hat{\beta}_1$ ficou afastado de β_1

e. Segundo os resultados obtidos, a que conclussões poderia chegar?

Uma das hipóteses do MRL é que $E(u|x) = 0$ o que implica que

$$\text{cov}(x,u) = \underbrace{E(xu)}_{E[E(xu|x)]=E[xE(u|x)]=0} - E(x)\underbrace{E(u)}_{E[E(u|x)]=0} = 0,$$

Ou seja quando $\text{cov}(x, u) \neq 0$ existem consequências na estimação, levando a estimadores viesados, o que consequentemente leva a interpretações erradas e afeta $\hat{y}$.

Questão 3

Mostre que $\displaystyle\sum_{i=1}^{n} \hat{u}_i^2/n$ é um estimador viessado para σ^2

Resposta

Queremos provar que

$$E\Big(\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n}\Big) \neq \sigma^2$$

Na Aula 3 (na demostração da variância estimada do erro), provamos que $E(\displaystyle\sum_{i=1}^{n} \hat{u}_i^2) = (n-2)\sigma^2$. Logo,

$$E\Big(\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n}\Big) = \frac{(n-2)\sigma^2}{n} \neq \sigma^2$$

Questão 4

Seja a regressão linear simples através da origem, *i.e.* $y = \beta_1 x + u$. Derive o estimador MQO $\hat{\beta}_1$

Resposta

O estimador MQO de β_1 é obtido

$$\hat{\beta}_1 = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - bx_i)^2$$

Seja $SQR = \sum_{i=1}^{n} (y_i - bx_i)^2$, derivando w.r.t. b temos que:

$$\frac{\partial SQR}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - bx_i)$$

Igualando a zero temos que $\sum_{i=1}^{n} x_i y_i = b\sum_{i=1}^{n} x_i^2$, então

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Para verificarmos que $\hat{\beta}_1$ é o valor que minimza SQR, precisamos do critério sa segunda derivada. Note que

$$\frac{\partial^2 SQR}{\partial b \partial b} = 2\sum_{i=1}^{n} x_i^2 > 0 \quad \text{ponto de mínimo}$$

Questão 5 (C2 do livro texto, pag 65)

O conjunto de dados **CEOSAL2.txt** contém informações sobre CEOs de empresas dos US. A variável *salary* é a compensação anual (em milhares de USD) e a variavel *ceoten* é o número prévio (em anos) como CEO da empresa.

Resposta

a. Encontre o salário médio e a permanência média dos CEOs

```r
CEOSAL2 = read.table("./DadosMRP/CEOSAL2.txt")
colnames(CEOSAL2) = c("salary", "age",   "college","grad",
                      "comten","ceoten","sales","profits","mktval",
                      "lsalary","lsales","lmktval","comtensq",
                      "ceotensq","profmarg")

# Método 1
summary(CEOSAL2$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   471.0   707.0   865.9  1119.0  5299.0
```

```r
summary(CEOSAL2$ceoten)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   6.000   7.955  11.000  37.000
```

```r
# Método 2
mean(CEOSAL2$salary)
```

```
## [1] 865.8644
```

```r
mean(CEOSAL2$ceoten)
```

```
## [1] 7.954802
```

b. Quantos CEOs estão no seu primeiro ano na empresa? ($ceoten = 0$)

```r
# Metodo 1
table(CEOSAL2$ceoten == 0)
```

```
##
## FALSE  TRUE
##   172     5
```

```r
# Metodo 2
sum(CEOSAL2$ceoten == 0)
```

```
## [1] 5
```

```r
# Metodo 3
table(CEOSAL2$ceoten)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 26 28
##  5 19 10 21 21 10 11  6 11  8  8  4  7  7  5  2  2  2  1  2  4  1  1  3  2  1
## 34 37
##  1  2
```

```r
# Metodo 4
nrow(CEOSAL2[CEOSAL2$ceoten == 0,])
```

```
## [1] 5
```

c. Calcule a regressão $\log(salary) = \beta_0 + \beta_1 ceoten + u$

```r
modelo = lm(log(salary)~ceoten, data = CEOSAL2)
modelo
```

```
##
## Call:
## lm(formula = log(salary) ~ ceoten, data = CEOSAL2)
```

```
##
## Coefficients:
## (Intercept)       ceoten
##    6.505498     0.009724
```

Se preferir, pose escrever o modelo:

$$\widehat{\log(salary)} = 6.505498 + 0.009724 \text{ ceoten}$$

d. Qual o aumento percentual previsto no salário dos CEOs se tem um ano a mais como CEO na empresa?

Para responder este item, precisamos lembrar como interpretar os parêmetros. A seguinte tabela (disponível na Aula 3), apresenta um bom resumo:

Modelo	V. Dep	V. Indep	Interpretação 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	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Nossa equação de regressão é da forma

$$\widehat{\log(salary)} = 6.505498 + 0.009724 \text{ ceoten}$$

Então, quando *ceoten* aumenta em uma unidade, o salário aumenta em $100 \times 0.009724 = 0.9724\%$

Questão 6

O conjunto de dados **WAGE1.txt** contém informações pessoas na força de trabalho em 1996. A variável *wage* é o salário médio por hora (em USD) e a variavel *exper* é o número de anos de experiencia.

Resposta

```r
WAGE1 = read.table("./DadosMRP/WAGE1.txt")[,c(1,3)]
colnames(WAGE1) = c("wage","exper")
```

a. Faça uma análise exporatória de dados de ambas as variáveis

```r
summary(WAGE1)
```

```
##       wage            exper      
##  Min.   : 0.530   Min.   : 1.00  
##  1st Qu.: 3.330   1st Qu.: 5.00  
##  Median : 4.650   Median :13.50  
##  Mean   : 5.896   Mean   :17.02  
##  3rd Qu.: 6.880   3rd Qu.:26.00  
##  Max.   :24.980   Max.   :51.00
```

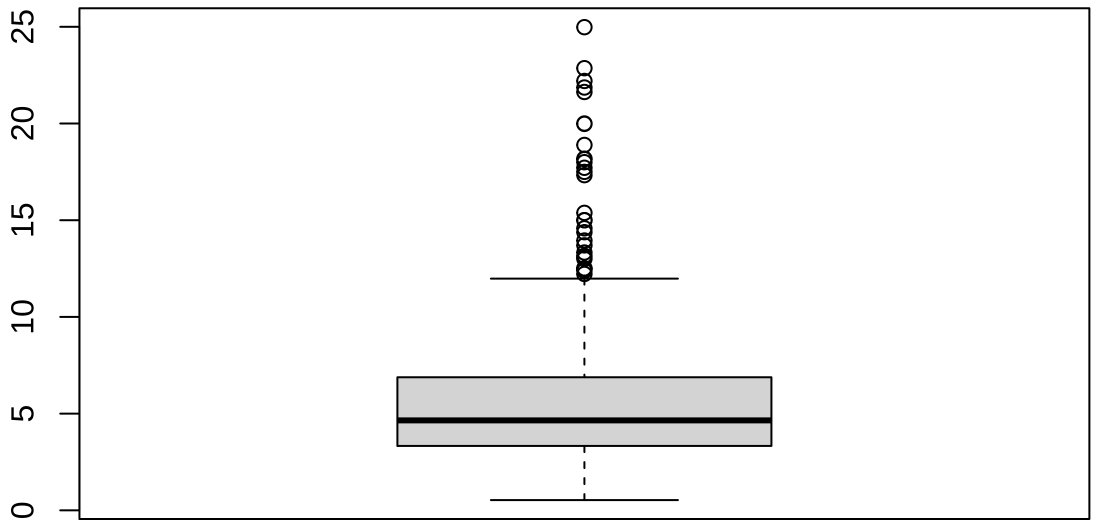
```r
sd(WAGE1$wage)
```

```
## [1] 3.693086
```

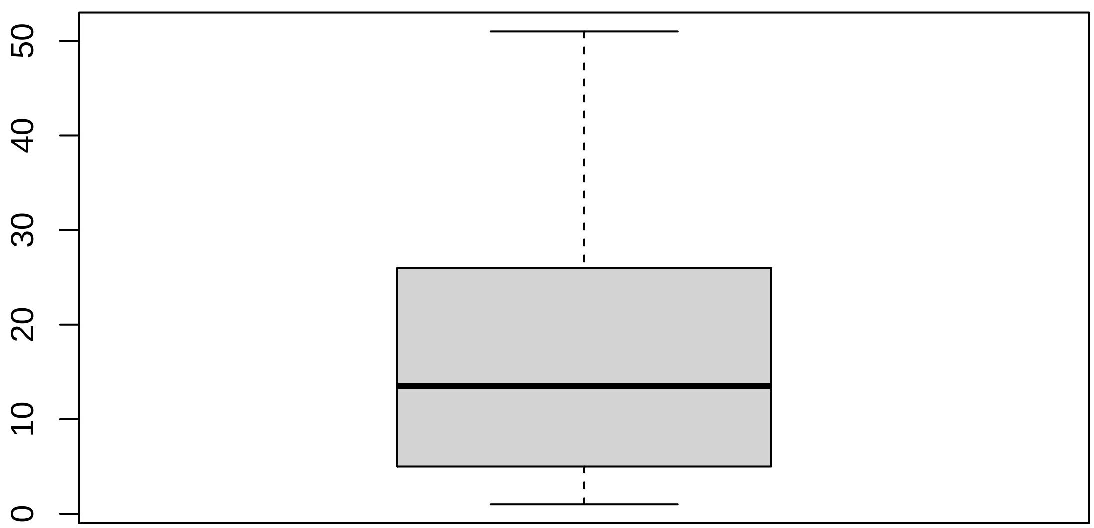
```r
sd(WAGE1$exper)
```

```
## [1] 13.57216
```

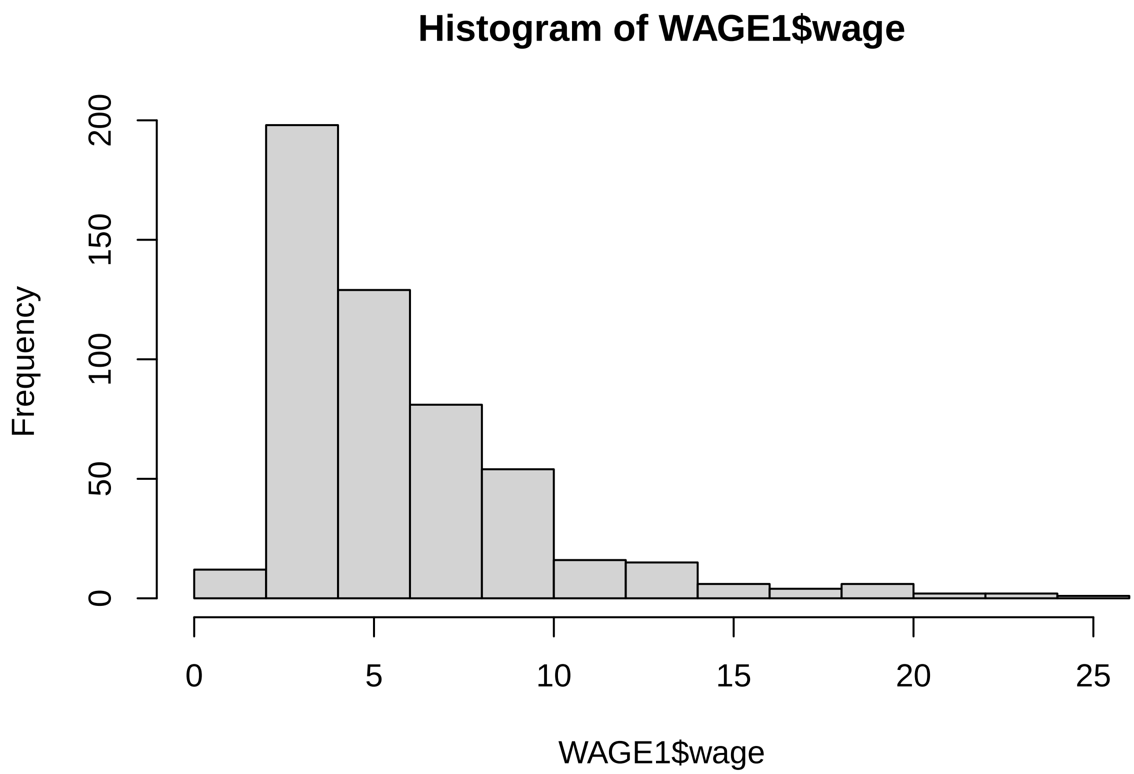
```r
boxplot(WAGE1$wage)
```



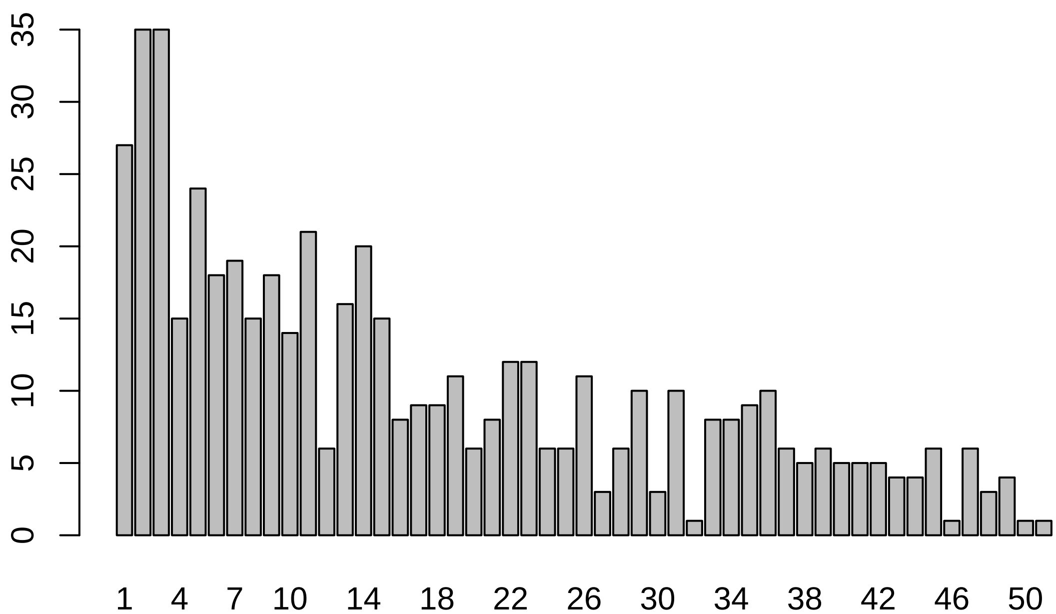

```r
boxplot(WAGE1$exper)
```



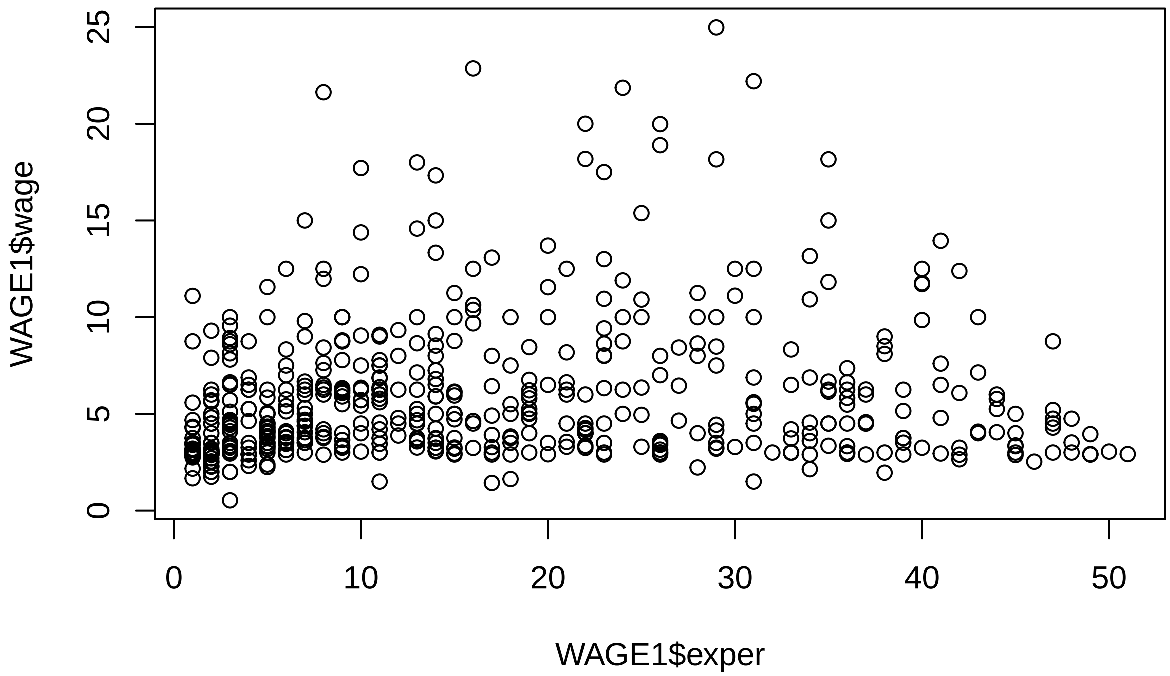

```r
hist(WAGE1$wage)
```




```r
barplot(table(WAGE1$exper))
```

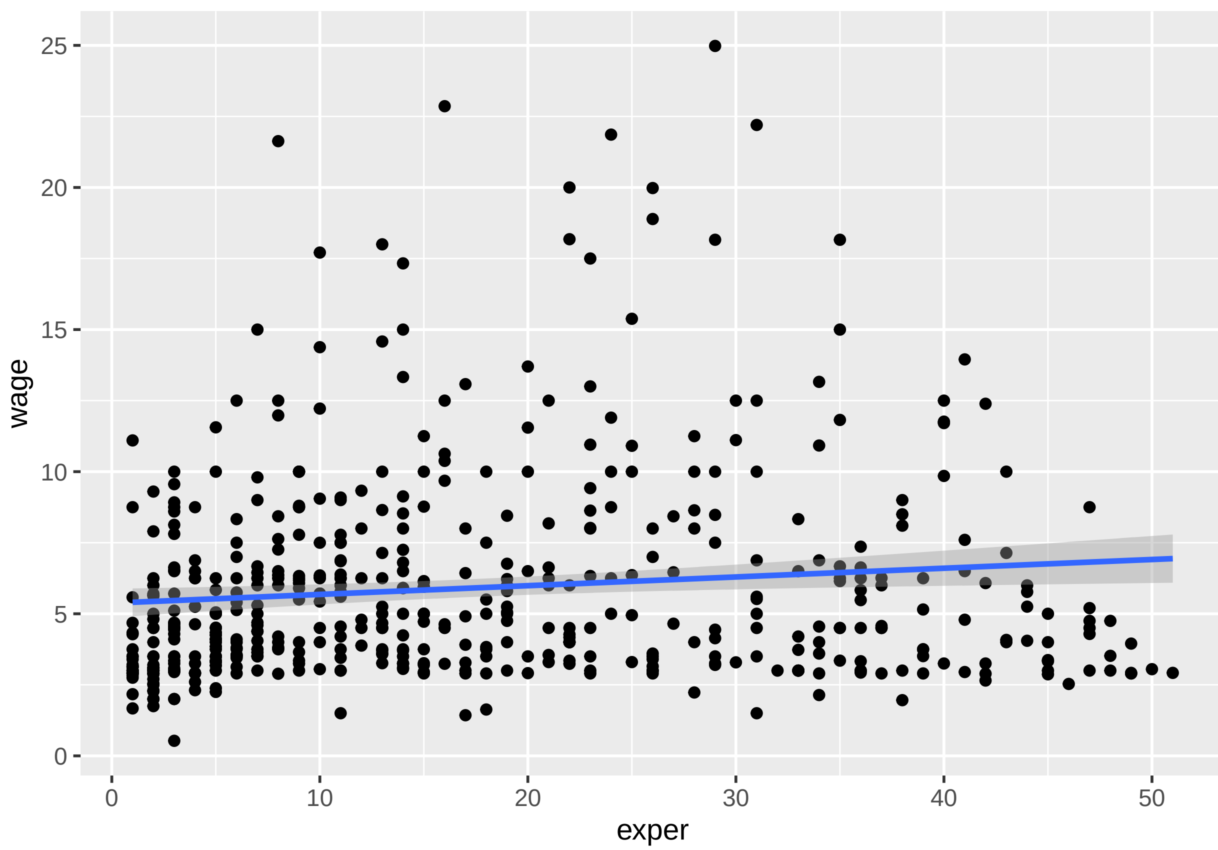



```r
plot(WAGE1$exper,WAGE1$wage)
```



b. Construa um gráfico de dispersão e grafique a reta de regressão

```
library(ggplot2)
ggplot(WAGE1, aes(x = exper, y = wage)) + geom_point() +
  geom_smooth(method = "lm")
```



c. Calcule a regressão $wage = \beta_0 + \beta_1 expert + u$

```
modelo = lm(wage~exper, data = WAGE1)
modelo
```

```
##
## Call:
## lm(formula = wage ~ exper, data = WAGE1)
##
## Coefficients:
## (Intercept)        exper
##     5.37331      0.03072
```

$$\widehat{wage} = 5.37331 + 0.03072 \text{ exper}$$

d. Interprete os resultados

A cada ano de *exper* adicional, o salário médio por hora aumenta em 0.03072 centavos

e. Calcule a regressão $log(wage) = \beta_0 + \beta_1 expert + u$

```
modelo = lm(log(wage)~exper, data = WAGE1)
modelo
```

```
##
## Call:
## lm(formula = log(wage) ~ exper, data = WAGE1)
##
## Coefficients:
## (Intercept)        exper
##    1.549043     0.004362
```

$$\widehat{\log(wage)} = 1.549043 + 0.004362 \text{ exper}$$

f. Interprete os resultados

A cada ano de *exper* adicional, o salário médio por hora aumenta em $(0.004362 \times 100 = 0.4362)$ %

Questão 7

O conjunto de dados **catholic** contém informações de puntuações de testes de estudantes dos U.S. que cursaram a oitava série em um determinado ano. As variáveis *math12* e *read12* são notas padronizadas de matemática e leitura respectivamente.

Resposta

a. Quantos estudantes existem na amostra? Encontre as médias e desvio padrão de cada variavel.

```
library(wooldridge)
```

```
##
## Attaching package: 'wooldridge'

## The following object is masked from 'package:MASS':
##
##     cement
```

```
data(catholic)
dim(catholic)
```

```
## [1] 7430   13
```

```r
round(apply(catholic,2,mean),4)
```

```
##           id       read12       math12       female        asian       hispan
## 4589838.2704      51.7724      52.1336       0.5174       0.0517       0.1035
##        black     motheduc     fatheduc      lfaminc       hsgrad       cathhs
##       0.0707      13.3569      13.6742      10.3533           NA       0.0608
##      parcath
##       0.3459
```

```r
round(apply(catholic,2,sd),4)
```

```
##           id       read12       math12       female        asian       hispan
## 2744467.0466       9.4078       9.4591       0.4997       0.2214       0.3046
##        black     motheduc     fatheduc      lfaminc       hsgrad       cathhs
##       0.2563       2.0060       2.2678       0.7945           NA       0.2390
##      parcath
##       0.4757
```

```r
summary(catholic$hsgrad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  1.0000  1.0000  0.9303  1.0000  1.0000    1460
```

```r
mean(catholic$hsgrad, na.rm = TRUE)
```

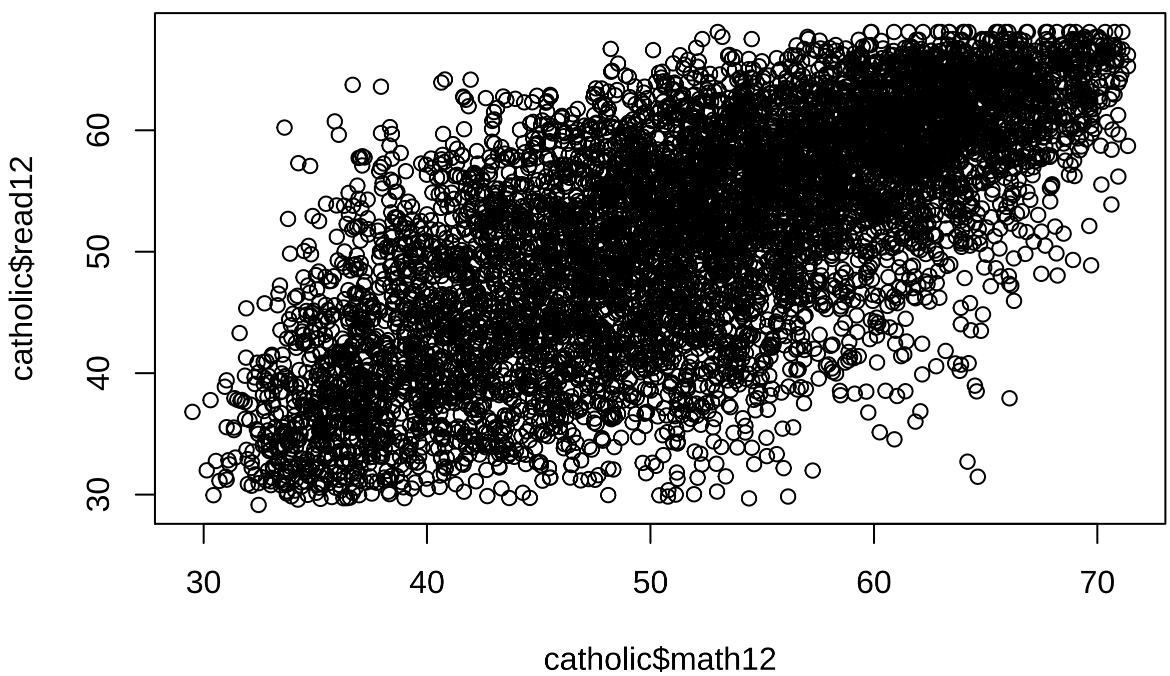
```
## [1] 0.9303183
```

```r
sd(catholic$hsgrad, na.rm = TRUE)
```

```
## [1] 0.2546312
```

b. Construa um boxplot de ambas variáveis. O que poderia dizer ao olhar o gráfico?

```r
plot(catholic$math12, catholic$read12)
```



c. Calcule a regressão *math12* sobre *read12* ($mate12 = \beta_0 + \beta_1 read12 + u$). Como é a qualidade do ajuste do modelo?

```r
modelo = lm(math12~read12, data = catholic)
summary(modelo)
```

```
##
## Call:
## lm(formula = math12 ~ read12, data = catholic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.5477  -4.5934   0.1838   4.6984  27.0182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.15304    0.43204   35.07   <2e-16 ***
## read12       0.71429    0.00821   87.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.658 on 7428 degrees of freedom
## Multiple R-squared:  0.5047, Adjusted R-squared:  0.5046
## F-statistic:  7569 on 1 and 7428 DF,  p-value: < 2.2e-16
```

O modelo

$$\widehat{math}12 = 15.1530 + 0.7143 \text{ read12}$$

explica 50.47% da variabilidade de $math12$, podemos dizer que o modelo se ajusta razoavelmente aos dados.

d. Interprete os resultados

A cada aumento de *read12* em uma unidade, espera-se que *math12* aumente em 0.7143 pontos

e. Faça um graficos de *math12* vs. $\widehat{math}12$

```r
plot(catholic$math12,fitted(modelo))
```