Exercícios em computador

C1 Os dados do arquivo `401K` são um subconjunto de dados analisados por Papke (1995) para estudar a relação entre a participação em um plano de pensão `401k` e a generosidade do plano. A variável `prate` é a porcentagem de trabalhadores aptos e com uma conta ativa; esta é a variável que gostaríamos de explicar. A medida da generosidade é a taxa de contribuição do plano, `mrate`. Esta variável mostra a quantia média com que a empresa contribui para o fundo trabalhista a cada US$ 1 de contribuição do trabalhador. Por exemplo, se a `mrate` = 0,50, então uma contribuição de US$ 1 do trabalhador corresponde a uma contribuição de US$ 0,50 da empresa.

```
library(wooldridge)
data("k401k")
```

(i) Encontre a taxa de participação e a taxa de contribuição médias na amostra de planos.

```
mean(k401k$prate) # media taxa de participacao
```

```
## [1] 87.36291
```

```
mean(k401k$mrate) # media taxa de contribuicao
```

```
## [1] 0.7315124
```

(ii) Agora, estime a equação de regressão simples, e relate os resultados ao lado do tamanho da amostra e do R-quadrado.

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate$$

```
model <- lm(prate ~ mrate, data = k401k)
summary(model) # resultados
```

```
##
## Call:
## lm(formula = prate ~ mrate, data = k401k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.303  -8.184   5.178  12.712  16.807
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  83.0755     0.5633  147.48   <2e-16 ***
## mrate         5.8611     0.5270   11.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 16.09 on 1532 degrees of freedom
## Multiple R-squared:  0.0747, Adjusted R-squared:  0.0741
## F-statistic: 123.7 on 1 and 1532 DF,  p-value: < 2.2e-16
```

(iii) Interprete o intercepto de sua equação. Interprete o coeficiente de `mrate`.

O intercepto $\beta_0 = 83,07$ indica o valor esperado ou estimado de `prate` quando `mrate` é igual a zero. Já o coeficiente β_1 indica que existe uma relação positiva entre `mrate` e `prate`, e que também a cada unidade acrescidade de `mrate` estima-se um aumento em `prate` de 5,86 unidades.

(iv) Encontre a prate prevista quando `mrate` = 3,5. Esta é uma previsão razoável? Explique o que está ocorrendo aqui.

$$\widehat{prate} = 83,07 + 5,86 * 3,5$$

```
83.07 + (5.86 * 3.5)
```

```
## [1] 103.58
```

Este valor é impossível, dado que a taxa máxima de participação é de 100%. Isto ilustra que, especialmente quando as variáveis dependentes são limitadas, um modelo de regressão simples pode fornecer previsões estranhas para valores extremos da variável independente. Na amostra, existem apenas 34 valores de `mrate` maiores ou iguais a 3,5.

(v) Quanto da variação da prate é explicada pela `mrate`? Na sua opinião, isso é bastante?

Aproximadamento 7,4% da variação de `prate` é explicada por `mrate`. Este não é um valor alto, indicando que provavelmente existem outros fatores que influenciam a taxa de participação.

C2 O conjunto de dados do arquivo `CEOSAL2` contém informações sobre CEOs de corporações norte-americanas. A variável `salary` é a compensação anual, em milhares de dólares, e `ceoten` é o número prévio de anos como CEO da empresa.

```
library(wooldridge)
data("ceosal2")
```

(i) Encontre o salário médio e a permanência média na amostra.

```
mean(ceosal2$salary) # salario medio
```

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

```
mean(ceosal2$ceoten) # permanencia media
```

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

(ii) Quantos CEOs estão em seu primeiro ano no cargo (isto é, `ceoten` = 0)? Qual é a permanência mais longa como CEO?

```
max(subset(ceosal2, ceoten == 0)$comten) # permanencia mais longa quando ceoten igual a 0
```

```
## [1] 33
```

(iii) Estime o modelo de regressão simples e registre seus resultados da forma usual. Qual é o aumento percentual previsto (aproximado) no salário quando se tem um ano a mais como CEO?

$$log(salary) = \beta_0 + \beta_0 ceoten + \mu$$

```
model <- lm(lsalary ~ ceoten, data = ceosal2) # estimando modelo
summary(model) # resultados
```

```
##
## Call:
## lm(formula = lsalary ~ ceoten, data = ceosal2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15314 -0.38319 -0.02251  0.44439  1.94337
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.505498   0.067991  95.682   <2e-16 ***
## ceoten      0.009724   0.006364   1.528    0.128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6038 on 175 degrees of freedom
## Multiple R-squared:  0.01316,    Adjusted R-squared:  0.007523
## F-statistic: 2.334 on 1 and 175 DF,  p-value: 0.1284
```

$$\widehat{log(salary)} = 6,5055 + 0,0097 * ceoten$$
$$n = 177, \qquad R^2 = 0,01316$$

Um aumento de um ano a mais como CEO gera um aumento no salário em 0,97%.

C3 Use os dados do arquivo `SLEEP75`, de Biddle e Hamermesh (1990), para estudar se há uma compensação entre o tempo gasto dormindo por semana e o tempo gasto em um trabalho remunerado. Podemos usar qualquer variável como a variável dependente. Para materializar, estime o modelo

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

em que `sleep` são os minutos dormidos à noite por semana e `totwrk` é o total de minutos trabalhados durante a semana.

```r
library(wooldridge)
data("sleep75")
```

```r
model <- lm(sleep ~ totwrk, data = sleep75) # estimando modelo
summary(model) # resultados
```

```
##
## Call:
## lm(formula = sleep ~ totwrk, data = sleep75)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2429.94  -240.25     4.91   250.53  1339.72
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3586.37695   38.91243  92.165   <2e-16 ***
## totwrk        -0.15075    0.01674  -9.005   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 421.1 on 704 degrees of freedom
## Multiple R-squared:  0.1033, Adjusted R-squared:  0.102
## F-statistic: 81.09 on 1 and 704 DF,  p-value: < 2.2e-16
```

(i) Registre seus resultados em uma equação junto com o número de observações e o R^2. O que o intercepto desta equação significa?

$$\widehat{sleep} = 3586,38 - 0,15 * totwrk$$
$$n = 706, \qquad R^2 = 0,1033$$

O intercepto da equação indica os minutos dormidos por semana quando a quantidade de minutos trabalhados na semana, `totwrk` for igual a zero.

(ii) Se `totwrk` aumentar 2 horas, quanto você estima que `sleep` cairá? Você acha que este é um efeito grande?

```r
model$coefficients[2] * 2 * 60 # totwrk é medido em horas, portanto precisa fazer a conversão
```

```
##   totwrk
## -18.0895
```

Um aumento de duas horas de trabalho na semana, reduzirá em 18 minutos dormidos durante a semana. Isso não parece ser um valor muito alto.

C4 Use os dados do arquivo `WAGE2` para estimar uma regressão simples que explique o salário mensal (`wage`) em termos da pontuação do QI (`IQ`).

```r
library(wooldridge)
data("wage2")
```

(i) Encontre o salário médio e o IQ médio da amostra. Qual é o desvio padrão amostral do IQ? (Pontuações de IQ são padronizadas, por isso, a média na população é 100 com um desvio padrão igual a 15.)

```r
mean(wage2$wage) # salário médio
```

```
## [1] 957.9455
```

```r
mean(wage2$IQ) # QI médio
```

```
## [1] 101.2824
```

```r
sd(wage2$IQ) # desvio-padrão de QI
```

```
## [1] 15.05264
```

(ii) Estime um modelo de regressão simples em que um aumento de um ponto em IQ altere wage em uma quantia constante de dólares. Use este modelo para encontrar o aumento previsto do salário para o caso de um acréscimo de 15 pontos de IQ. O IQ explica a maior parte da variação em wage?

```r
model <- lm(wage ~ IQ, data = wage2) # estimando o modelo
summary(model) # resultados
```

```
##
## Call:
## lm(formula = wage ~ IQ, data = wage2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -898.7 -256.5  -47.3  201.1 2072.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.9916    85.6415   1.366    0.172
## IQ            8.3031     0.8364   9.927   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 384.8 on 933 degrees of freedom
## Multiple R-squared:  0.09554,    Adjusted R-squared:  0.09457
## F-statistic: 98.55 on 1 and 933 DF,  p-value: < 2.2e-16
```

$$\widehat{salary} = 116,9 + 8,30(IQ)$$
$$n = 935, \qquad R^2 = 0,09554$$

Um aumento de 15 no `IQ` aumenta o salário mensal previsto em $8,30 * (15) = \$124,50$ (em dólares de 1980). A pontuação de `IQ` não explica nem 10% da variação salarial.

(iii) Agora, estime um modelo em que cada acréscimo de um ponto em `IQ` tenha o mesmo efeito percentual em `wage`. Se `IQ` aumentar 15 pontos, qual será o aumento percentual previsto aproximado em `wage`?

```
model <- lm(lwage ~ IQ, data = wage2) # estimando modelo
summary(model)
```

```
##
## Call:
## lm(formula = lwage ~ IQ, data = wage2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09324 -0.25547  0.02261  0.27544  1.21487
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.8869944  0.0890206   66.13   <2e-16 ***
## IQ          0.0088072  0.0008694   10.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3999 on 933 degrees of freedom
## Multiple R-squared:  0.09909,    Adjusted R-squared:  0.09813
## F-statistic: 102.6 on 1 and 933 DF,  p-value: < 2.2e-16
```

$$\widehat{log(salary)} = 5,89 + 0,008(IQ)$$
$$n = 935, \qquad R^2 = 0,09$$

Se variar em 15 então $\Delta log(salário) = 0,0088(15) = 0,132$, que é a mudança proporcional (aproximada) no salário previsto. O aumento percentual é, portanto, de aproximadamente 13,2.

C5 Para a população de empresas do setor químico, defina `rd` como os gastos anuais em pesquisa e desenvolvimento, e `sales` como as vendas anuais (ambos em milhões de dólares).

(i) Escreva um modelo (não uma equação estimada) que implique uma elasticidade constante entre `rd` e `sales`. Qual é o parâmetro da elasticidade?

$$log(rd) = \beta_0 + \beta_1(sales) + \mu$$

- β_1 é a elasticidade de `rd` em relação a `sales`

(ii) Agora, estime o modelo usando os dados do arquivo `RDCHEM`. Monte a equação estimada da forma usual. Qual é a elasticidade estimada de `rd` em relação a `sales`? Explique o que essa elasticidade significa.

```
library(wooldridge)
data("rdchem")
```

```
model <- lm(lrd ~ lsales, data = rdchem) # estimando modelo
summary(model)
```

```
##
## Call:
## lm(formula = lrd ~ lsales, data = rdchem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90406 -0.40086 -0.02178  0.40562  1.10439
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.10472    0.45277  -9.066 4.27e-10 ***
## lsales       1.07573    0.06183  17.399  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5294 on 30 degrees of freedom
## Multiple R-squared:  0.9098, Adjusted R-squared:  0.9068
## F-statistic: 302.7 on 1 and 30 DF,  p-value: < 2.2e-16
```

A elasticidade estimada de `rd` em relação à `sales` é de 1,076. Estima-se que um aumento de um por cento nas vendas aumente os gastos em pesquisa e desenvolvimento em cerca de 1,08%.

C6 Usamos os dados do arquivo `MEAP93` no Exemplo 2.12. Agora, queremos explorar a relação entre a taxa de aprovação em matemática (`math10`) e os gastos por estudante (`expend`).

```
library(wooldridge)
data("meap93")
```

(i) Você acha que cada dólar adicional gasto tem o mesmo efeito sobre a taxa de aprovação ou um efeito decrescente seria mais razoável? Explique.

A taxa de aprovação deve crescer a taxas decrescentes, ou seja, quanto maior a taxa de aprovação, maior será o gasto necessário para aumentar a taxa de aprovação na mesma unidade.

(ii) No modelo populacional, argumente que $\beta_1/10$ é a porcentagem de alteração em `math10` dado um aumento de 10% em gasto.

$$math10 = \beta_0 + \beta_1 log(expend) + \mu$$

Ceteris paribus, a variação de `math10` é dada por

$$\Delta math10 = \beta_1 \Delta log(expend)$$
$$\Delta math10 \approx \frac{\beta_1}{100}(\%\Delta expend)$$

Portanto, para uma variação de 10% em gasto, $\%\Delta expend = 10$

$$\Delta math10 \approx \frac{\beta_1}{100} \times 10 \approx \frac{\beta_1}{10}$$

(iii) Use os dados do arquivo `MEAP93` para estimar o modelo (ii). Descreva a equação estimada da forma usual, incluindo o tamanho da amostra e o R-quadrado.

```
model <- lm(math10 ~ lexpend, data = meap93) # estimando modelo
summary(model)
```

```
##
## Call:
## lm(formula = math10 ~ lexpend, data = meap93)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.343  -7.100  -0.914   6.148  39.093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -69.341     26.530  -2.614 0.009290 **
## lexpend       11.164      3.169   3.523 0.000475 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.35 on 406 degrees of freedom
## Multiple R-squared:  0.02966,    Adjusted R-squared:  0.02727
## F-statistic: 12.41 on 1 and 406 DF,  p-value: 0.0004752
```

$$\widehat{math10} = -69,341 + 11,164\ log(expend)$$
$$n = 608, \qquad R^2 = 0,02966$$

(iv) Quão grande é o efeito de gastos estimado? Em outras palavras, se os gastos aumentarem 10%, qual será o aumento percentual estimado em `math10`?

Se aumentarmos os gastos em 10%, `math10` aumentará em 1,1% aproximadamente. Este efeito é maior quando consideramos escolas com um gasto baixo.

(v) Alguns podem se preocupar com o fato de que a análise de regressão pode produzir valores ajustados para `math10` maiores do que 100. Por que isso não é tão preocupante neste conjunto de dados?

```r
subset(meap93, math10 > 100) # quantiade de observações maiores do que 100
```

```
##  [1] lnchprg  enroll   staff    expend   salary   benefits droprate gradrate
##  [9] math10   sci11    totcomp  ltotcomp lexpend  lenroll  lstaff   bensal
## [17] lsalary
## <0 rows> (or 0-length row.names)
```

```r
max(meap93$math10) # valor máximo de math10 na base de dados
```

```
## [1] 66.7
```

```r
max(model$fitted.values) # valor máximo de math ajustado no modelo
```

```
## [1] 30.15375
```

Isso é preocupante pois não existem observações no conjunto de dados com valores para `math10` maiores do que 100.

C7 Use os dados do arquivo `CHARITY` [retirado de Franses e Paap (2001)] para responder às seguintes questões:

```r
library(wooldridge)
data("charity")
```

(i) Qual é a doação (`gift`) média da amostra de 4.268 pessoas (em florins holandeses)? Qual é a porcentagem de pessoas com nenhuma doação?

```r
mean(charity$gift) # doação média
```

```
## [1] 7.44447
```

```r
mean(charity$gift == 0) # porcentagem de pessoas com nenhuma doação
```

```
## [1] 0.6000469
```

(ii) Qual é a média de envios por ano? Quais são os valores mínimos e máximos?

```r
mean(charity$mailsyear) # media de envios por ano
```

```
## [1] 2.049555
```

```r
min(charity$mailsyear) # valor mínimo
```

```
## [1] 0.25
```

```r
max(charity$mailsyear) # valor máximo
```

```
## [1] 3.5
```

(iii) Estime o modelo por MQO e registre os resultados da forma usual, incluindo o tamanho da amostra e o R-quadrado.

$$gift = \beta_0 + \beta_1 mailsyear + \mu$$

```r
model <- lm(gift ~ mailsyear, data = charity) # estimando modelo
summary(model)
```

```
##
## Call:
## lm(formula = gift ~ mailsyear, data = charity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.287  -7.976  -5.976   2.687 245.999
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0141     0.7395   2.724  0.00648 **
## mailsyear     2.6495     0.3431   7.723  1.4e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.96 on 4266 degrees of freedom
## Multiple R-squared:  0.01379,    Adjusted R-squared:  0.01356
## F-statistic: 59.65 on 1 and 4266 DF,  p-value: 1.404e-14
```

$$\widehat{gift} = 2,0141 + 2,6495(mailsyear)$$
$$n = 4268, \qquad R^2 = 0,01379$$

(iv) Interprete o coeficiente de inclinação. Se cada envio custa um florim, a instituição de caridade espera obter um lucro líquido em cada um dos envios? Isso quer dizer que a instituição obtém um lucro líquido em todos os envios? Explique.

O coeficiente de inclinação indica que cada envio por ano está associado a cerca de 2,65 florins adicionais, em média. Portanto, se cada envio custa um florim, o lucro esperado de cada envio é estimado em 1,65 florins. Esta é apenas a média, no entanto. Algumas correspondências não geram contribuições ou geram uma contribuição inferior ao custo da correspondência, outras correspondências geraram muito mais do que o custo da correspondência.

(v) Qual é a menor contribuição à instituição prevista na amostra? Usando essa análise de regressão simples, você pode prever zero de `gift`?

Como o menor ano de correspondência na amostra é 0.25, o menor valor previsto de presentes é $2,01 + 2,65(0,25) \approx 2,67$. Mesmo se olharmos para a população em geral, onde algumas pessoas não receberam qualquer correspondência, o menor valor previsto é cerca de dois. Portanto, com esta equação estimada, nunca prevemos zero doações de caridade.

C8 Para completar este exercício, você precisará de um programa que lhe permita gerar dados das distribuições uniforme e normal.

(i) Comece gerando 500 observações em x_i – a variável explicativa – a partir da distribuição uniforme com variação [0,10]. (A maioria dos programas estatísticos tem um comando para distribuição Uniforme(0,1); só multiplique essas observações por 10.) Qual é a média da amostra e o desvio padrão da amostra de x_i?

```
x <- runif(500, min = 0, max = 10)
mean(x) # media
```

```
## [1] 4.784269
```

```
sd(x) # desvio padrão
```

```
## [1] 2.893138
```

(ii) Gere, de forma aleatória, 500 erros, u_i, a partir da distribuição Normal(0,36). Se você gerar uma Normal(0,1), como geralmente está disponível, simplesmente multiplique os resultados por seis.) A média amostral de u_i é exatamente zero? Por que sim ou por que não? Qual é o desvio padrão amostral de u_i?

```
u <- rnorm(500, mean = 0, sd = 6)
mean(u) # media
```

```
## [1] 0.05204305
```

```
sd(u) # desvio padrão
```

```
## [1] 6.235285
```

A média amostral não é exatamente zero. Isso ocorre porque uma distribuição normal possui um distribuição de probabilidades que contenha 95% dos valores dentro de um intervalo de 2 desvios-padrões da média. Portanto, dificilmente os valores serão exatamente iguais a zero.

(iii) Agora gere y_i como a equação abaixo, isto é, o intercepto da população é um e a inclinação populacional é dois. Use os dados para executar a regressão de y_i em x_i. Quais são suas estimativas de intercepto e inclinação? Elas são iguais aos valores populacionais da equação abaixo? Explique.

$$y_i = 1 + 2x_i + \mu_i \equiv \beta_0 + \beta_1 x_1 + \mu_i$$

```
y <- 1 + 2*x + u
model <- lm(y ~ x)
model
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      0.9308       2.0253
```

(iv) Obtenha os resíduos MQO, $\hat{u}_i$, e verifique se a equação (2.60) se mantém (sujeita a erros de arredondamento).

```
sum(model$residuals)
```

```
## [1] 1.434408e-13
```

```
sum(model$residuals * x)
```

```
## [1] 9.585666e-13
```

(v) Calcule as mesmas quantidades da equação (2.60), mas use os erros u_i no lugar dos resíduos. Agora, o que você conclui?

(vi) Repita os itens (i), (ii) e (iii) com uma nova amostra de dados, começando com a geração de x_i. Agora, o que você obtém de $\hat{\beta}_0$ e $\hat{\beta}_1$? Por que isto é diferente do que você obteve no item (iii)?

C9 Use os dados do arquivo COUNTYMURDERS para responder a essas questões. Utilize somente os dados de 1996.

```
library(wooldridge)
data("countymurders")
dados <- subset(countymurders, year == 1996)
```

(i) Quantos condados tiveram zero assassinatos em 1996? Quantos condados tiveram pelo menos uma execução? Qual é o maior número de execuções?

```r
sum(dados$arrests == 0, na.rm = T) # condados com zero assassinatos em 1996
```

```
## [1] 1043
```

```r
sum(dados$execs > 1, na.rm = T) # condados com pelo menos uma execução em 1996
```

```
## [1] 3
```

(ii) Estime a equação abaixo, em que murders corresponde ao número de assassinatos, por MQO e relate os resultados da forma usual, incluindo o tamanho da amostra e o R-quadrado.

$$murders = \beta_0 + \beta_1 execs + \mu$$

```r
model <- lm(murders ~ execs, data = dados) # estimando o modelo
summary(model)
```

```
##
## Call:
## lm(formula = murders ~ execs, data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -149.12   -5.46   -4.46   -2.46 1338.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4572     0.8348   6.537 7.79e-11 ***
## execs        58.5555     5.8333  10.038  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.89 on 2195 degrees of freedom
## Multiple R-squared:  0.04389,    Adjusted R-squared:  0.04346
## F-statistic: 100.8 on 1 and 2195 DF,  p-value: < 2.2e-16
```

$$\widehat{murders} = 5,45 + 58,55(execs)$$
$$n = 2197, \qquad R^2 = 0,04389$$

(iii) Interprete o coeficiente de inclinação registrado no item (ii). A equação estimada sugere um efeito dissuasor da pena capital?

O coeficiente de inclinação indica que para o aumento de um `execs` estima-se um aumento em 58 assassinatos (`murders`). Portanto, a pena capital não possui um efeito dissuasor.

(iv) Qual é o menor número de assassinatos que pode ser previsto pela equação? Qual é o resíduo de um condado com zero execuções e zero assassinatos?

O menor número de assassinatos que pode ser previsto pelo modelo é 5.46. Este valor é também o erro para os casos em que não existem assassinatos nem pena capitais.

(v) Explique por que uma análise de regressão simples não é adequada para determinar se a pena capital tem um efeito dissuasor sobre os assassinatos.

Uma análise de regressão simples pode não ser adequada para determinar se a pena capital tem um efeito dissuasor sobre os assassinatos por várias razões:

1. **Fatores Confusos (Confounding Variables):**Existem muitos fatores que podem influenciar as taxas de homicídio além da presença da pena capital. Fatores como:

- Nível de policiamento e eficácia das forças policiais
- Condições socioeconômicas (desemprego, pobreza, educação)
- Políticas públicas e programas de prevenção ao crime
- Presença de drogas e gangues
- Se esses fatores não forem controlados na análise, a regressão simples pode fornecer uma relação - espúria entre a pena capital e as taxas de homicídio.

2. **Causalidade vs. Correlação:**Uma regressão simples pode mostrar uma correlação entre a pena capital e as taxas de homicídio, mas isso não implica causalidade. A relação observada pode ser influenciada por outros fatores ou pode ser puramente coincidental.

3. **Viés de Seleção:**As jurisdições que adotam a pena capital podem ser diferentes das que não adotam em aspectos fundamentais que também afetam as taxas de homicídio. Por exemplo, estados com taxas de homicídio mais altas podem ser mais propensos a adotar a pena capital, o que pode introduzir viés na análise.

4. **Endogeneidade:**Há um problema de endogeneidade se a variável independente (pena capital) estiver correlacionada com o termo de erro na regressão. Isso pode ocorrer se, por exemplo, as taxas de homicídio elevadas levarem à adoção da pena capital, e não o contrário. Este problema torna as estimativas de regressão simples viesadas e inconsistentes.

5. **Mudanças ao Longo do Tempo (Dinâmica Temporal):**As taxas de homicídio podem variar ao longo do tempo devido a uma série de fatores. Uma análise de regressão simples que não leve em conta essas variações temporais pode não capturar corretamente a relação entre pena capital e homicídios.

6. **Heterogeneidade entre Jurisdições:** Os estados ou países podem diferir significativamente em termos de aplicação da pena capital, eficácia do sistema judicial, cultura e normas sociais. Uma análise de regressão simples que não leve em consideração essas heterogeneidades pode ser inadequada.

C10 O conjunto de dados do arquivo `CATHOLIC` inclui informações de pontuações de testes de mais de 7.000 estudantes dos Estados Unidos que cursaram a oitava série em 1988. As variáveis `mate12` e `leitu12` são notas padronizadas de matemática e leitura, respectivamente.

```
library(wooldridge)
data("catholic")
```

(i) Quantos estudantes existem na amostra? Encontre as médias e desvios padrão de `mate12` e `leitu12`.

```r
nrow(catholic) # quantidade de estudantes na amostra
```

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

```r
sapply(catholic[,2:3], mean) # medias
```

```
##   read12   math12
## 51.77240 52.13362
```

```r
sapply(catholic[,2:3], sd) # desvios-padrão
```

```
##   read12   math12
## 9.407761 9.459117
```

(ii) Compute a regressão simples de `mate12` sobre `leitu12` para obter o intercepto MQO e as estimativas de inclinação. Reporte os resultados na forma usual.

```r
model <- lm(math12 ~ read12, data = catholic) # estimando o modelo
summary(model)
```

```
##
## 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
```

$$\widehat{mate12} = 15,15 + 0,71(leitu12)$$
$$n = 7430, \qquad R^2 = 0,5047$$

(iii) O intercepto registrado na parte (ii) tem uma interpretação significativa? Explique.

O intercepto significa neste caso, a nota esperada em matemática independentemente da nota de leitura.

(iv) Você está surpreso pelo β_1 encontrado? E quanto ao R^2?

Existe uma forte assosiação entre as notas de leitura e matemática. O modelo indica que, para um aumento na pontuação de leitura em uma unidade, espera-se um aumento de 0,7 na pontuação de matemática. Isso pode ser explicado pelo fato de que bons alunos, em geral, apresentam boas notas em diversas matérias, sendo o contrário também verdadeiro, maus alunos, apresentam notas ruins em várias matérias.

(v) Suponha que você apresente suas descobertas ao superintendente distrital de educação e ele diga: "Suas descobertas mostram que, para aumentar as notas de matemática, precisamos somente melhorar as notas de leitura; portanto, devemos contratar mais professores de leitura". Como você responderia a este comentário? (Dica: Se você calculasse a regressão de `leitu12` sobre `mate12`, ao invés do contrário, o que esperaria descobrir?)

```
model <- lm(read12 ~ math12, data = catholic)
summary(model)
```

```
##
## Call:
## lm(formula = read12 ~ math12, data = catholic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.1459  -4.2021   0.4885   4.4920  22.8935
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.937062   0.430318   34.71   <2e-16 ***
## math12       0.706556   0.008122   87.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.621 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
```

Esse raciocínio não está correto, pelos fatores apontados na resposta anterior.