R para Economia

Lucas Mendes

24/03/2020

Modelos Cross - Section

Regressão Linear Simples

Lembra de sua aula de introdução à microeconomia? Tire seu livro do Mankiw do armário!

Agora pense que você irá analisar o mercado de **bananas**. Representando suas curvas de oferta e demanda

curva de demanda: $Y = \beta_d - \alpha_d X$

curva de oferta: $Y = \beta_o + \alpha_o X$

Regressão Linear Simples

Se considerarmos que $\beta_d = 80$ e $\beta_o = 10$ sendo que $\alpha_d = 4$ e $\alpha_o = 6$

- curva de demanda: $Y = 80 - 4X$

Regressão Linear Simples

Se considerarmos que $\beta_d = 80$ e $\beta_o = 10$ sendo que $\alpha_d = 4$ e $\alpha_o = 6$

- curva de demanda: $Y = 80 - 4X$
- curva de oferta: $Y = 10 + 6X$

Regressão Linear Simples

Temos como agora calcular o equilibrio do mercado igualando a curva de demanda a curva de oferta

$$80 - 4X = 10 + 6X \ (1)$$

$$70 = 10X \ (2)$$

$$7 = X \ (3)$$

Quandidade de equilibrio $= 7$

Preço de equilibrio $= 52$

Regressão Linear Simples

Isso foi o que você provavelmente fez em introdução a micro ou algo do tipo

So que nessa época, o seu professor te dava os valores de α e β

Agora você mesmo irá calcula - los!

Disclaimer

Especificação do modelo de regressão linear simples

$$y = \beta_1 + \beta_2 x$$

- O Y pode ser chamado de varios nomes, como variavel regressora, variavel dependente, variavel resposta e por ai vai.

Disclaimer

Especificação do modelo de regressão linear simples

$$y = \beta_1 + \beta_2 x$$

- O Y pode ser chamado de varios nomes, como variavel regressora, variavel dependente, variavel resposta e por ai vai.
- Porém eu irei chama - la de variavel **endógena**, ou seja, que é determinada pelo modelo.

Disclaimer

Especificação do modelo de regressão linear simples

$$y = \beta_1 + \beta_2 x$$

- O Y pode ser chamado de varios nomes, como variavel regressora, variavel dependente, variavel resposta e por ai vai.
- Porém eu irei chama - la de variavel **endógena**, ou seja, que é determinada pelo modelo.
- A mesma coisa vale para X, que tem varios nomes, mas eu chamarei de varável **exógena**.

Disclaimer

Especificação do modelo de regressão linear simples

$$y = \beta_1 + \beta_2 x$$

- O Y pode ser chamado de varios nomes, como variavel regressora, variavel dependente, variavel resposta e por ai vai.
- Porém eu irei chama - la de variavel **endógena**, ou seja, que é determinada pelo modelo.
- A mesma coisa vale para X, que tem varios nomes, mas eu chamarei de varável **exógena**.
- O que estiver no lado esquerdo da equação = **endógena**

Disclaimer

Especificação do modelo de regressão linear simples

$$y = \beta_1 + \beta_2 x$$

- O Y pode ser chamado de varios nomes, como variavel regressora, variavel dependente, variavel resposta e por ai vai.
- Porém eu irei chama - la de variavel **endógena**, ou seja, que é determinada pelo modelo.
- A mesma coisa vale para X, que tem varios nomes, mas eu chamarei de varável **exógena**.
- O que estiver no lado esquerdo da equação = **endógena**
- O que estiver no lado direito da equação = **exógena**

Diferença entre da teoria para a vida real

Modelo Determinístico

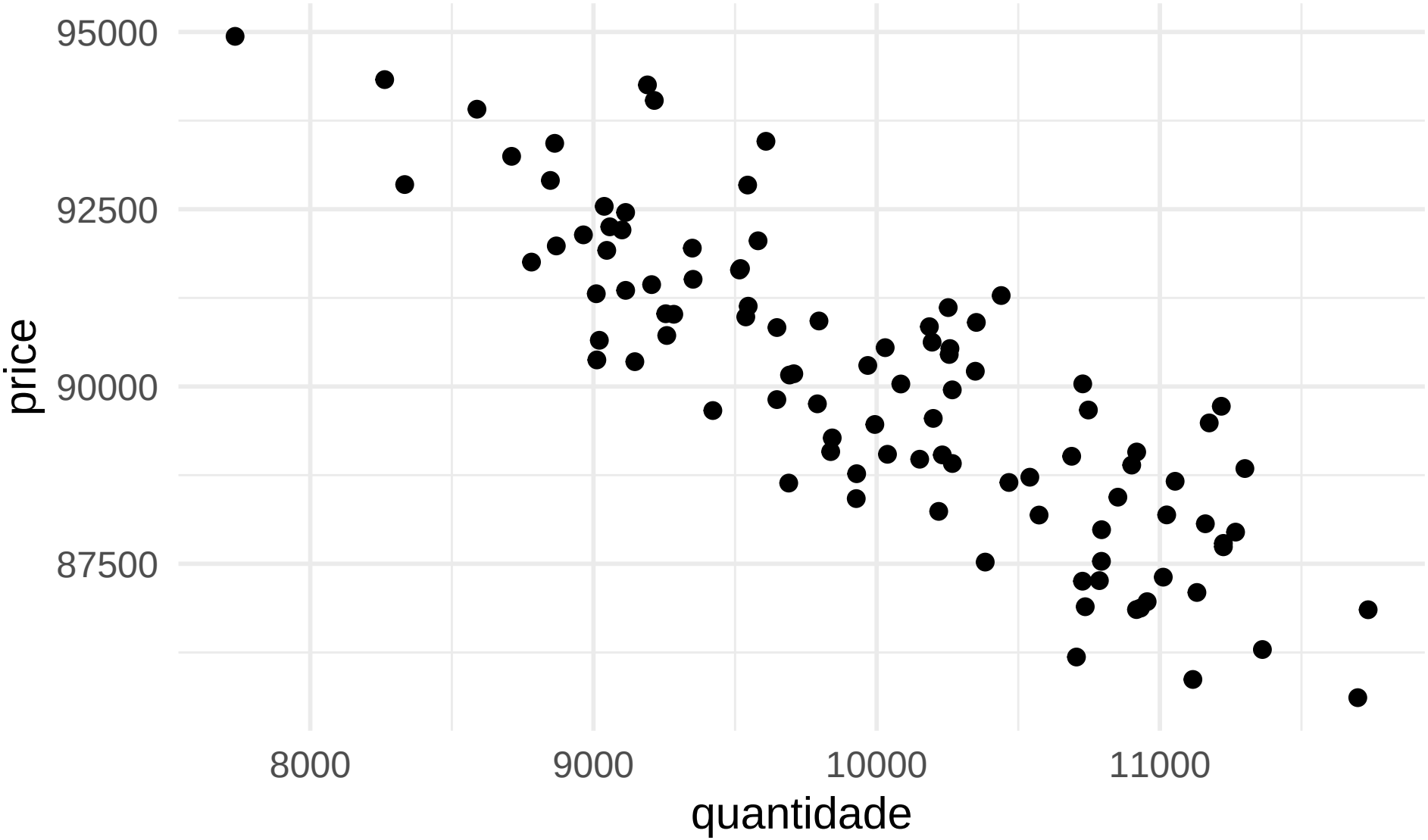
$$y = \beta_1 + \beta_2 x$$

Modelo estocástico

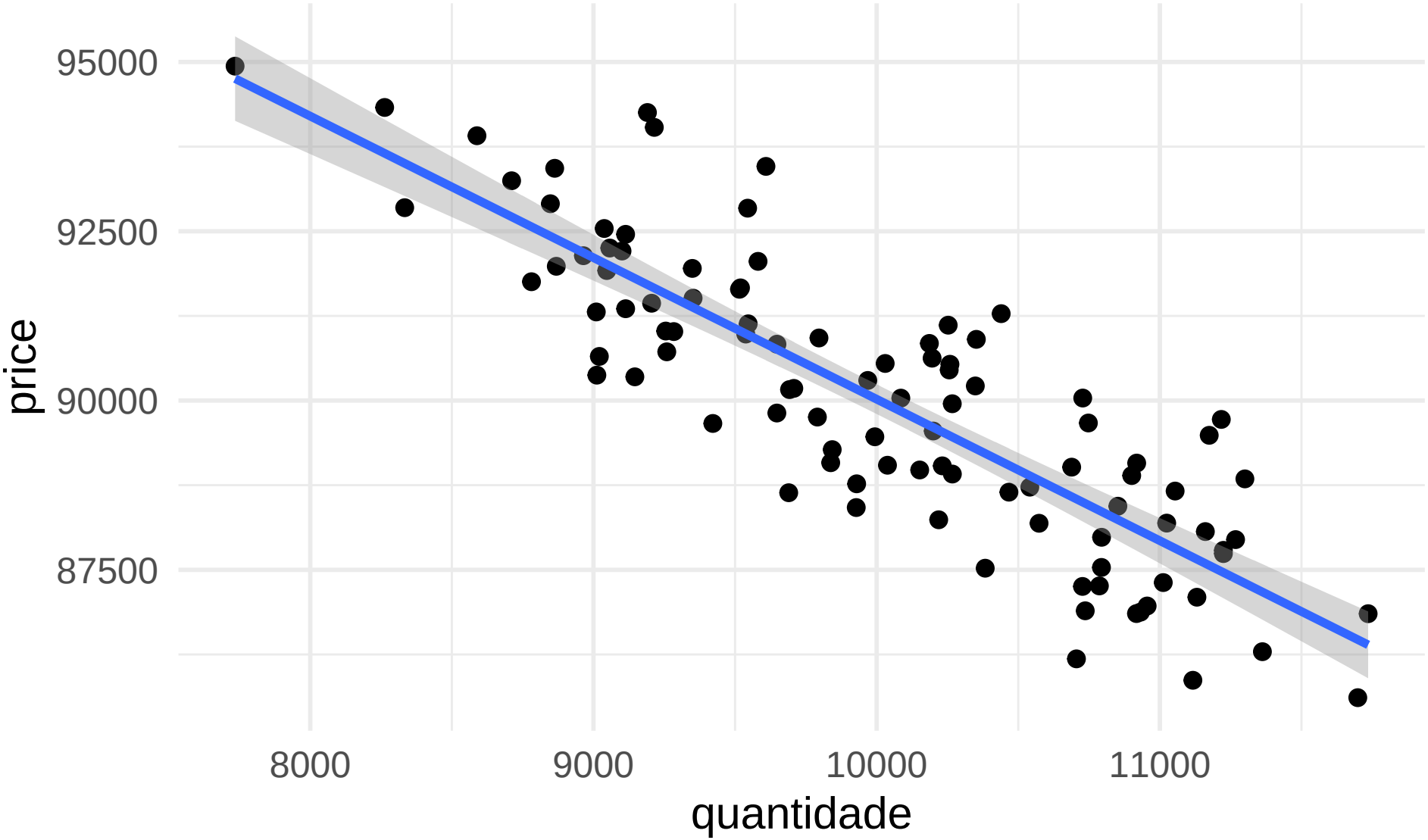
$$y = \beta_1 + \beta_2 x + \epsilon$$

- O que é o ϵ? O erro do nosso modelo, você não achava que ele ia acertar sempre né?
- O nosso algoritmo de regressão vai tentar achar β_1 e β_2 que esse erro.

Graficamente



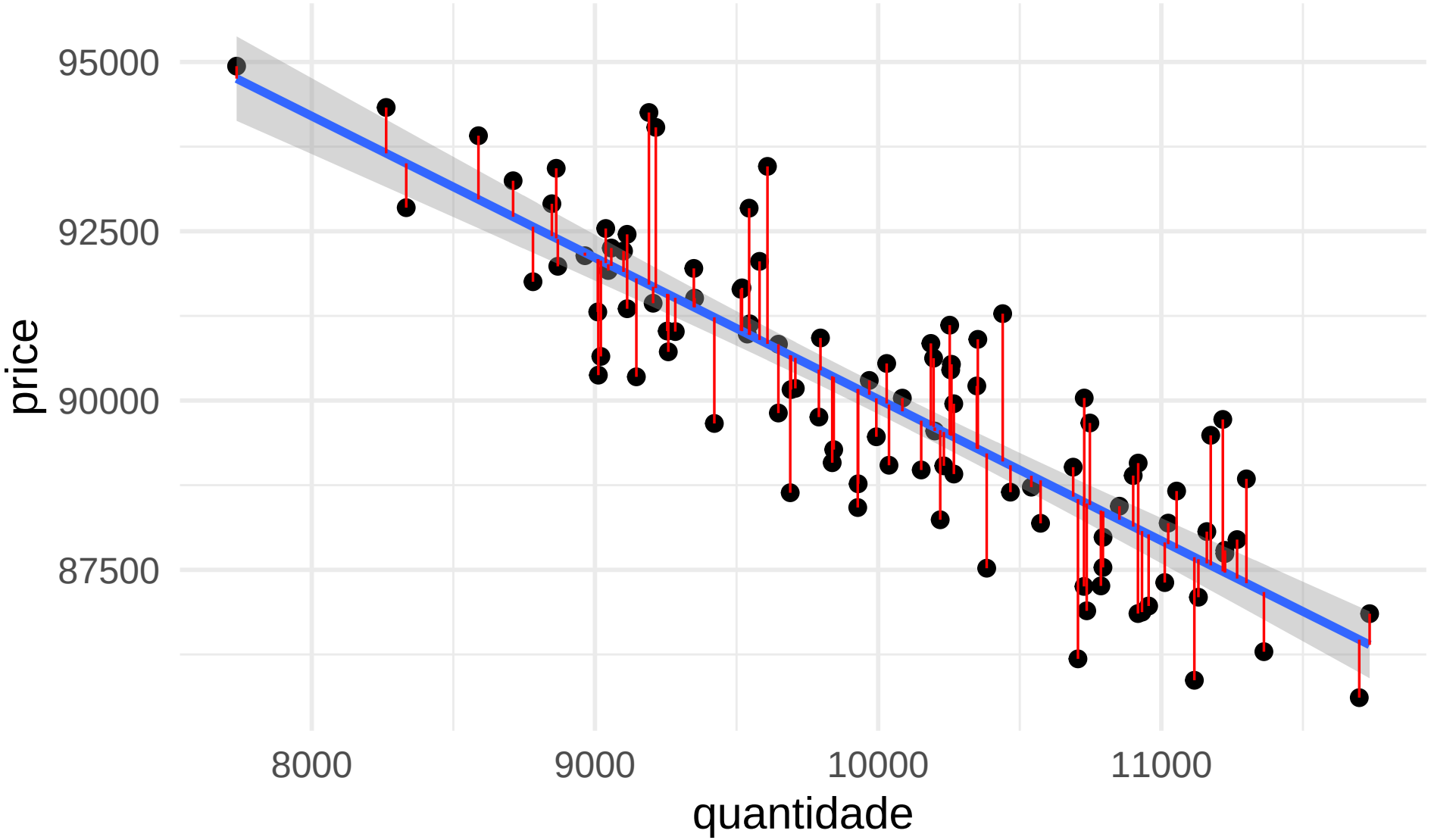
Achando a reta melhor reta



Achando a reta melhor reta

- Nós queremos achar a melhor reta que descreve o relacionamento entre as variáveis, ou seja, aquela que minimiza o erro
- Note porém que não é qualquer erro, estamos falando do erro quadrado

Onde está o erro?



Onde está o erro?

O erro é o quadrado das retas em vermelho (Distância do ponto até a reta).

$$\epsilon = y - \beta_1 + \beta_2 x$$

Elevando tudo ao quadrado...

$$\sum_{i=1}^{n} (y - \beta_1 + \beta_2 x)^2$$

- Desse jeito, derivamos para achar o melhor α e β
- Mas isso o algoritmo faz por nós!

Regressão Linear Simples

Nesse capitulo iremos usar o pacote `AER` (Applied Econometrics with R) e o pacote caret (Machine Learning)

Cole no console e rode

```
# install.packages('AER')
# install.packages('caret')
```

```
library(AER)
library(caret)
```

Regressão Linear Simples

Iremos analisar agora a base de dados CPS1985, referente a pesquisa de determinação salarial feita em 1985 nos EUA.

Queremos verificar qual o impacto do total de anos de educação sobre o salario/hora de um indivíduo

Carregando o pacote

```
data('CPS1985')
```

Regressão Linear Simples

	wage	education	experience	age	ethnicity
1	5.10	8	21	35	hispanic
1100	4.95	9	42	57	cauc
2	6.67	12	1	19	cauc
3	4.00	12	4	22	cauc
4	7.50	12	17	35	cauc
5	13.07	13	9	28	cauc

Regressão Linear Simples

Especificando nossa regressão

$$wage = \beta_1 + \beta_2 educ$$

Essa é nossa especificação da regressão, o código seguinte irá calcular os parâmetros β_1 e β_2

Regressão Linear Simples

Iremos agora treinar um modelo de regressão linear usando a função `train()` do pacote **caret**

```
modelo <- train(wage ~ # Variavel Exógena
                  education, # Variavel endógena
                method = "lm", # Linear Model
                data = CPS1985) # Base de dados
```

Regressão Linear Simples

Com o modelo criado,podemos observar as estatisticas usando o comando `summary()`.

```
summary(modelo)
```

Regressão Linear Simples

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.911 -3.260 -0.760  2.240 34.740
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.74598    1.04545  -0.714    0.476
## education    0.75046    0.07873   9.532   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 4.754 on 532 degrees of freedom
```

Regressão Linear Simples

Eu particularmente não gosto muito do formato que o summary nos retorna.

Como eu sigo a filosofia do tidyverse, eu transformo isso para um dataframe com a função `tidy()` do pacote `broom` (Já instalado com tidyverse)

```
library(broom)
summary(modelo) %>% tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.7459797	1.0454541	-0.7135461	0.4758208
education	0.7504608	0.0787337	9.5316300	0.0000000

Regressão Linear Simples

Nosso modelo estimado

$$wage = -0.74 + educ0.75$$

O que podemos retirar do modelo e das estatísticas?
Normalmente olhamos para:

- O coefieciente das variaveis

Regressão Linear Simples

Nosso modelo estimado

$$wage = -0.74 + educ0.75$$

O que podemos retirar do modelo e das estatísticas?
Normalmente olhamos para:

- O coefieciente das variaveis
- O valor t dessas variaveis

Regressão Linear Simples

Nosso modelo estimado

$$wage = -0.74 + educ0.75$$

O que podemos retirar do modelo e das estatísticas?
Normalmente olhamos para:

- O coefieciente das variaveis
- O valor t dessas variaveis
- O R^2

Regressão Linear Simples

Coeficiente

- Quando analisamos o coeficiente de uma regressão, normalmente nós esperamos o seu sinal devido a uma teoria prévia.

Regressão Linear Simples

Coeficiente

- Quando analisamos o coeficiente de uma regressão, normalmente nós esperamos o seu sinal devido a uma teoria prévia.
- No nosso exemplo esperamos que seja positivo ja que é um consenso que mais anos de estudo impactam positivamente no salario.

Regressão Linear Simples

Coeficiente

- Quando analisamos o coeficiente de uma regressão, normalmente nós esperamos o seu sinal devido a uma teoria prévia.
- No nosso exemplo esperamos que seja positivo ja que é um consenso que mais anos de estudo impactam positivamente no salario.
- O que normalmente queremos testar é a magnitude do efeito de uma variavel sobre a outra.

Regressão Linear Simples

Coeficiente

- Quando analisamos o coeficiente de uma regressão, normalmente nós esperamos o seu sinal devido a uma teoria prévia.
- No nosso exemplo esperamos que seja positivo ja que é um consenso que mais anos de estudo impactam positivamente no salario.
- O que normalmente queremos testar é a magnitude do efeito de uma variavel sobre a outra.
- O nosso modelo nos forneceu que β_2 era 0.75

Regressão Linear Simples

Coeficiente

- Quando analisamos o coeficiente de uma regressão, normalmente nós esperamos o seu sinal devido a uma teoria prévia.
- No nosso exemplo esperamos que seja positivo ja que é um consenso que mais anos de estudo impactam positivamente no salario.
- O que normalmente queremos testar é a magnitude do efeito de uma variavel sobre a outra.
- O nosso modelo nos forneceu que β_2 era 0.75
- A interpretação portanto é: Se um indivíduo estuda 1 ano a mais, ele ganha em média 0.75 centavos/hora a mais de salário

Regressão Linear Simples

Coeficiente

- Supondo que um indivíduo **A** estudou 10 anos

Regressão Linear Simples

Coeficiente

- Supondo que um indivíduo **A** estudou 10 anos
- Salario/hora $= -0.74 + 10 * 0.75 = 6.76$

Regressão Linear Simples

Coeficiente

- Supondo que um indivíduo **A** estudou 10 anos
- Salario/hora $= -0.74 + 10 * 0.75 = 6.76$
- Supondo que um individuo **B** estudou 11 anos

Regressão Linear Simples

Coeficiente

- Supondo que um indivíduo **A** estudou 10 anos
- Salario/hora $= -0.74 + 10 * 0.75 = 6.76$
- Supondo que um individuo **B** estudou 11 anos
- Salario/hora $= -0.74 + 11 * 0.75 = 7.51$

Regressão Linear Simples

Valor T

- O valor t é um valor que vem da formula $t = \frac{\beta}{EP(\beta)}$

Regressão Linear Simples

Valor T

- O valor t é um valor que vem da formula $t = \frac{\beta}{EP(\beta)}$
- Essa pequena conta é um teste estatistico que avalia se o nosso coeficiente β_i é diferente de zero.

Regressão Linear Simples

Valor T

- O valor t é um valor que vem da formula $t = \frac{\beta}{EP(\beta)}$
- Essa pequena conta é um teste estatistico que avalia se o nosso coeficiente β_i é diferente de zero.
- A regra de bolso que levamos é que se $t > |2|$, podemos rejeitar que o coeficiente é igual a zero.

Regressão Linear Simples

Valor T

- O valor t é um valor que vem da formula $t = \frac{\beta}{EP(\beta)}$
- Essa pequena conta é um teste estatistico que avalia se o nosso coeficiente β_i é diferente de zero.
- A regra de bolso que levamos é que se $t > |2|$, podemos rejeitar que o coeficiente é igual a zero.
- Vamos fazer a conta

Regressão Linear Simples

Valor T

term	estimate	std.error	statistic	p.value
(Intercept)	-0.7459797	1.0454541	-0.7135461	0.4758208
education	0.7504608	0.0787337	9.5316300	0.0000000

Regressão Linear Simples

R^2

- O R^2 mede o poder de explicação de uma regressão

Regressão Linear Simples

R^2

- O R^2 mede o poder de explicação de uma regressão
- Seus valores variam de 0 a 1.

Regressão Linear Simples

R^2

- O R^2 mede o poder de explicação de uma regressão
- Seus valores variam de 0 a 1.
- No nosso exemplo ele é 0.14, ou 14%

Regressão Linear Simples

R^2

- O R^2 mede o poder de explicação de uma regressão
- Seus valores variam de 0 a 1.
- No nosso exemplo ele é 0.14, ou 14%
- Muitos podem se enganar olhando apenas esse indicador, use o com cuidado.

Regressão Linear Simples

R^2

$$R^2 = \frac{SQE}{SQT}$$

$$SQR = SQT - SQE$$

- SQT = Soma dos quadrados Totais
- SQE = Soma dos quadrados Explicados
- SQR = Soma dos quadrados dos Resíduos

Regressão Linear Simples

R^2

```
summary(modelo) %>% glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df
0.1458645	0.1442589	4.753987	90.85197	0	2

Exercicios

Elasticidades

Regressão Linear Simples (Elasticidades)

Talvez você ja tenha ouvido falar sobre elasticidades, talvez até calculado na forma discreta.

Para calcular elasticidades, precisamos deixar as variaveis logarizadas usando a função `log()`

No nosso exemplo sobre educação, ficaria da seguinte maneira

```
modelo <- train(log(wage) ~ log(education),
      method = "lm",
      data = CPS1985)
```

Regressão Linear Simples (Elasticidades)

Observando as estatisticas

```
summary(modelo) %>% broom::tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0701300	0.2371804	0.2956821	0.7675883
log(education)	0.7822175	0.0928906	8.4208482	0.0000000

Regressão Linear Simples (Elasticidades)

$$log(wage) = 0.07 + 0.78log(educ)$$

- Agora a interpretação dos coeficientes muda um pouco.

Regressão Linear Simples (Elasticidades)

$$log(wage) = 0.07 + 0.78log(educ)$$

- Agora a interpretação dos coeficientes muda um pouco.
- Nós lemos da seguinte maneira:

Regressão Linear Simples (Elasticidades)

$$log(wage) = 0.07 + 0.78log(educ)$$

- Agora a interpretação dos coeficientes muda um pouco.
- Nós lemos da seguinte maneira:
- Se eu aumentar meus anos de estudo em 1%

Regressão Linear Simples (Elasticidades)

$$log(wage) = 0.07 + 0.78log(educ)$$

- Agora a interpretação dos coeficientes muda um pouco.
- Nós lemos da seguinte maneira:
- Se eu aumentar meus anos de estudo em 1%
- Meu salario/hora irá aumentar em média 0.78%

Regressão Linear Simples (Elasticidades)

$$log(wage) = 0.07 + 0.78log(educ)$$

- Agora a interpretação dos coeficientes muda um pouco.
- Nós lemos da seguinte maneira:
- Se eu aumentar meus anos de estudo em 1%
- Meu salario/hora irá aumentar em média 0.78%
- Todas as estatisticas seguem o mesmo procedimento de análise

Um pouco da sua aula de micro

Micro

Logarizar também serve para:

- Deixar relações exponenciais em lineares

Supondo que temos uma função de produção F(L) = Y

$$Y = \beta_1 + L^{\beta_2}$$

Se aplicarmos log

$$log(Y) = \beta_1 + \beta_2 log(L)$$

Micro

Rodando essa regressão...

```
##
## Call:
## lm(formula = log(Y) ~ log(L), data = df)
##
## Coefficients:
## (Intercept)       log(L)
##      3.0120       0.7423
```

Não sei se vocês já estudaram, mas o nosso β_2 é a produtividade marginal do trabalho e como sabemos, ela segue a lei dos **rendimentos decrescentes**.

Vamos checar!

Micro

Primeiro iremos criar uma sequencia de trabalhadores que uma firma pode cotratar, essa é uma sequencia que vai de 1 até 50.

```
L <- 1:50
```

Agora iremos calcular o produto dessa firma para cada quantidade de trabalhadores que ela pode ter de acordo a nossa função de produção

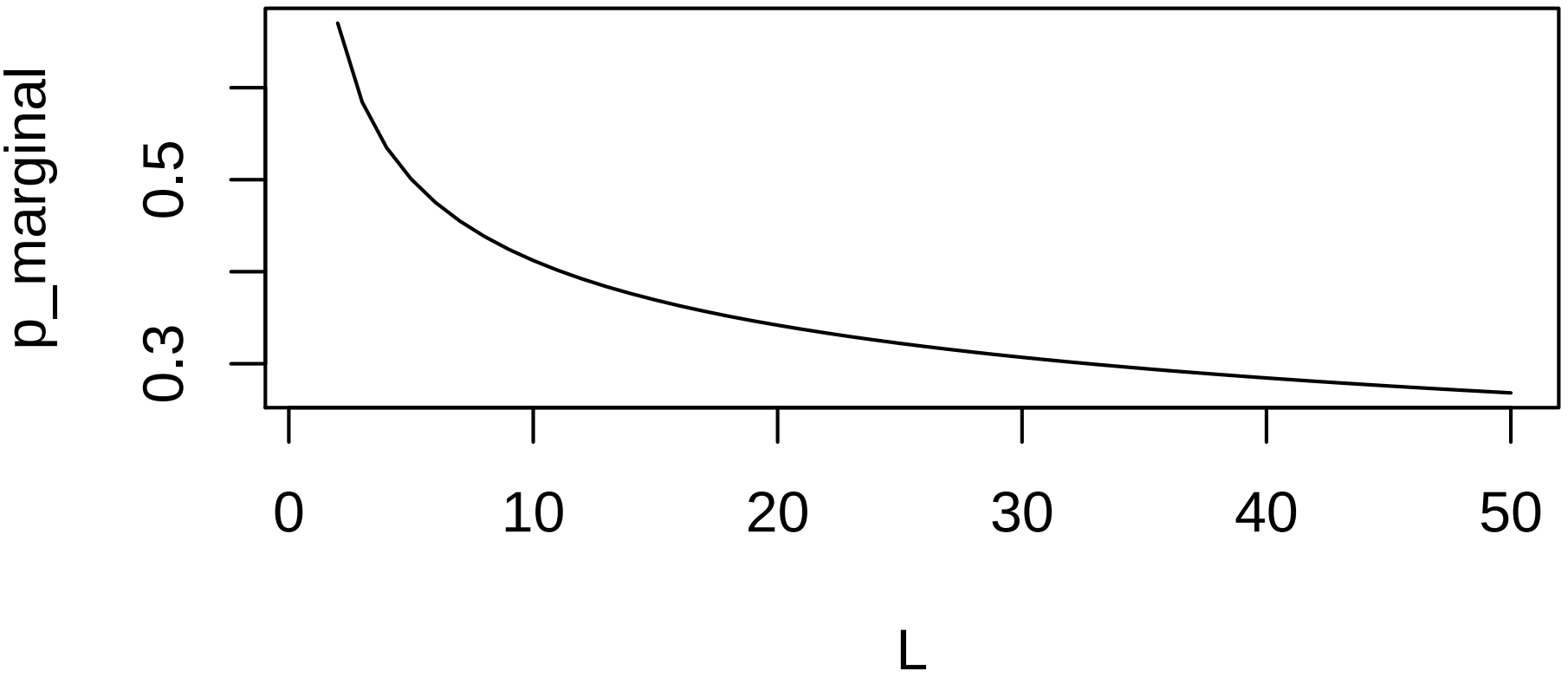
```
produto <- 3.01 + L^0.74
```

Agora iremos calcular o produto marginal por trabalhador

```
p_marginal <- produto - lag(produto)
```

Plotando

```
plot(L,p_marginal,type = "l")
```



Exercicios

Regressão Linear Multipla

Regressão Linear Multipla

A regressão linear multipla é quando estamos usando mais de uma varivel **exógena**.

Exemplo

$$wage = \beta_1 + \beta_2 educ + \beta_3 experience$$

Regressão Linear Multipla

Nós usamos a mesma função no R

```
modelo <- train(wage ~ education + experience,
      method = "lm",
      data = CPS1985)
```

Regressão Linear Multipla

```
summary(modelo) %>% broom::tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-4.9044823	1.2189240	-4.023616	6.56e-05
education	0.9259646	0.0814035	11.374999	0.00e+00
experience	0.1051316	0.0171975	6.113181	0.00e+00

Regressão Linear Multipla

Estimação do modelo

$$wage = -4.9 + 0.92educ + 0.1exp$$

Regressão Linear Multipla

F-statistic, Valor - P e R^2 ajustado

- Além de todas as estatisticas que estudamos, agora temos mais três para analisar

Regressão Linear Multipla

F-statistic, Valor - P e R^2 ajustado

- Além de todas as estatisticas que estudamos, agora temos mais três para analisar
- A **estatistica F** é uma continha que testa se conjuntamente, há pelo menos um coeficiente diferente de zero.

Regressão Linear Multipla

F-statistic, Valor - P e R^2 ajustado

- Além de todas as estatisticas que estudamos, agora temos mais três para analisar
- A **estatistica F** é uma continha que testa se conjuntamente, há pelo menos um coeficiente diferente de zero.
- Porém não há uma regra de bolso pois ele depende do graus de liberdade da regressão, então olhamos o valor - p por facilidade.

Regressão Linear Multipla

F-statistic, Valor - P e R^2 ajustado

- Além de todas as estatisticas que estudamos, agora temos mais três para analisar
- A **estatistica F** é uma continha que testa se conjuntamente, há pelo menos um coeficiente diferente de zero.
- Porém não há uma regra de bolso pois ele depende do graus de liberdade da regressão, então olhamos o valor - p por facilidade.
- A regra de bolso do **valor - p** é, caso seja menor que 5%(0.05), sua regressão tem pelo menos um coeficiente diferente de zero.

Regressão Linear Multipla

F-statistic, Valor - P e R^2 ajustado

- Além de todas as estatisticas que estudamos, agora temos mais três para analisar
- A **estatistica F** é uma continha que testa se conjuntamente, há pelo menos um coeficiente diferente de zero.
- Porém não há uma regra de bolso pois ele depende do graus de liberdade da regressão, então olhamos o valor - p por facilidade.
- A regra de bolso do **valor - p** é, caso seja menor que 5%(0.05), sua regressão tem pelo menos um coeficiente diferente de zero.
- Outra coisa que sabemos é que o R^2 de uma regressão sempre irá crescer ou pelo menos ficar constante caso você acrescente uma variavel enxógena

Regressão Linear Multipla

F-statistic, Valor - P e R^2 ajustado

- Além de todas as estatisticas que estudamos, agora temos mais três para analisar
- A **estatistica F** é uma continha que testa se conjuntamente, há pelo menos um coeficiente diferente de zero.
- Porém não há uma regra de bolso pois ele depende do graus de liberdade da regressão, então olhamos o valor - p por facilidade.
- A regra de bolso do **valor - p** é, caso seja menor que 5%(0.05), sua regressão tem pelo menos um coeficiente diferente de zero.
- Outra coisa que sabemos é que o R^2 de uma regressão sempre irá crescer ou pelo menos ficar constante caso você acrescente uma variavel enxógena
- Por isso, para compararmos regressões multiplas, usamos o **R^2 ajustado**, que penaliza o incremento de variáveis que não ajudem o modelo a explicar melhor

Regressão Linear Multipla

F-statistic, Valor - P e R^2 ajustado

```
summary(modelo) %>% broom::glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df
0.2020248	0.1990192	4.599365	67.21709	0	3

Test F

O teste F, chamado de teste F de significancia global da regressão, é calculado pela seguinte forma:

$$F = \frac{SQR}{SQT}$$

Ele testa se ,conjuntamente, alguma das variáveis enxógenas do modelo é **diferente** de zero.

H0 = Conjuntamente, os coeficientes são iguais a zero H1 = Há pelo menos 1 coeficiente diferente de zero

Calculado a estatistica F, comparamos com o F crítico que está na tabela de distribuição F de acordo com os graus de liberdade.

Para isso não ser preciso, olhamos então para o valor-P

Valor P

O **valor P**, resumidamente é:

O menor nível de significância com que se rejeitaria a hipótese nula.

Vocês podem não saber, mas estamos testando hipóteses com um nível de significância, respectivamente 10%,5% e 1%.

Normalmente, 5% de significancia é o valor padrão. Então, caso o valor p esteja abaixo disso, nós rejeitamos H0.

R^2 Ajustado

$$R^2_{ajustado} = 1 - \frac{n-1}{n-(k+1)}(1-R^2)$$

Quando quisermos comparar modelos com diferentes variáveis exógenas, usamos o **R^2 ajustado**.

n = Tamanho da amostra

k = Total de variáveis exógenas (Intercepto não conta)

Regressão com variáveis categóricas

Regressão com variáveis categóricas

- Até agora vimos regressões somente com variáveis exógenas **contínuas** e/ou **discretas**.

Regressão com variáveis categóricas

- Até agora vimos regressões somente com variáveis exógenas **contínuas** e/ou **discretas**.
- Agora iremos ver como aplicar regressões com variaveis exógenas **categóricas**, na qual representam classes.

Regressão com variáveis categóricas

- Até agora vimos regressões somente com variáveis exógenas **contínuas** e/ou **discretas**.
- Agora iremos ver como aplicar regressões com variaveis exógenas **categóricas**, na qual representam classes.
- Para quem não sabe o que são as categóricas aqui em baixo vão dois exemplos:

Regressão com variáveis categóricas

- Até agora vimos regressões somente com variáveis exógenas **contínuas** e/ou **discretas**.
- Agora iremos ver como aplicar regressões com variaveis exógenas **categóricas**, na qual representam classes.
- Para quem não sabe o que são as categóricas aqui em baixo vão dois exemplos:
- Categóricas **cardinais**: Quando não há um ordenamento. Sexo (H,M)

Regressão com variáveis categóricas

- Até agora vimos regressões somente com variáveis exógenas **contínuas** e/ou **discretas**.
- Agora iremos ver como aplicar regressões com variaveis exógenas **categóricas**, na qual representam classes.
- Para quem não sabe o que são as categóricas aqui em baixo vão dois exemplos:
- Categóricas **cardinais**: Quando não há um ordenamento. Sexo (H,M)
- Categóricas **ordinais**: Quando há um ordenamento. Educação (Doutorado $>$ Mestrado $>$ Graduação)

Regressão com variáveis categóricas

- Até agora vimos regressões somente com variáveis exógenas **contínuas** e/ou **discretas**.
- Agora iremos ver como aplicar regressões com variaveis exógenas **categóricas**, na qual representam classes.
- Para quem não sabe o que são as categóricas aqui em baixo vão dois exemplos:
- Categóricas **cardinais**: Quando não há um ordenamento. Sexo (H,M)
- Categóricas **ordinais**: Quando há um ordenamento. Educação (Doutorado $>$ Mestrado $>$ Graduação)
- No R essas variáveis são da classe `factor`

Regressão com variáveis categóricas

```
# Gender é uma variavel categorica
modelo <- train(wage ~ education + gender,
      method = "lm",
      data = CPS1985)
```

Regressão com variáveis categóricas

```
summary(modelo) %>% broom::tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.2178312	1.0363216	0.2101966	0.8335949
education	0.7512834	0.0768225	9.7794725	0.0000000
genderfemale	-2.1240567	0.4028322	-5.2728076	0.0000002

Regressão com variáveis categóricas

Modelo estimado

$$wage = 0.21 + educ0.75 - genderfemale2.12$$

- O coeficiente da variavel `gender` é -2.12.
  - Quando gender é **female**, o coeficiente é multiplicado por 1
  - Quando gender é **male**, o coeficiente é multiplicado por 0
  - O resultado dessa multiplicação é somado ao intercepto, que tem valor de 0.21

Regressão com variáveis categóricas

Fazendo a conta

- O indivíduo **A**, estudou por 10 anos e é do gênero masculino, logo:

Regressão com variáveis categóricas

Fazendo a conta

- O indivíduo **A**, estudou por 10 anos e é do gênero masculino, logo:
- $wage = 0.21 + 10 * 0.75 - 0 * 2.12 = 7.71$

Regressão com variáveis categóricas

Fazendo a conta

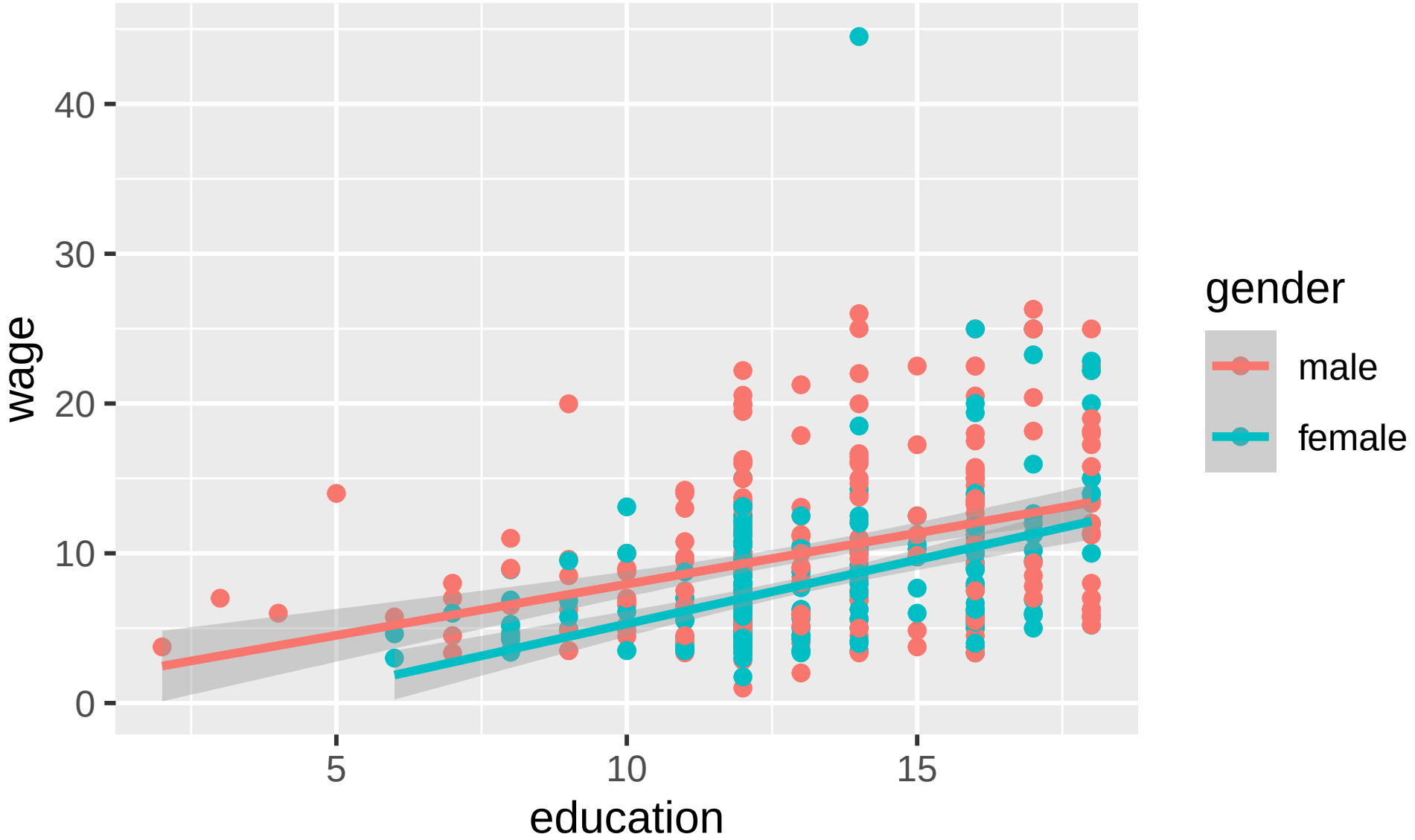
- O indivíduo **A**, estudou por 10 anos e é do gênero masculino, logo:
- $wage = 0.21 + 10 * 0.75 - 0 * 2.12 = 7.71$
- O indivíduo **B**, estudou por 10 anos e é do gênero feminino, logo:

Regressão com variáveis categóricas

Fazendo a conta

- O indivíduo **A**, estudou por 10 anos e é do gênero masculino, logo:
- $wage = 0.21 + 10 * 0.75 - 0 * 2.12 = 7.71$
- O indivíduo **B**, estudou por 10 anos e é do gênero feminino, logo:
- $wage = 0.21 + 10 * 0.75 - 1 * 2.12 = 5.59$

Graficamente



Modelo Logístico

Modelo Logístico

O modelo de regressão logística também usa variáveis categóricas, so que agora endógenamente.

Ou seja, não queremos agora prever um possível número médio, mas sim uma classe como **sim** ou **não**.

Vamos pensar o contrário na nossa base de dados agora. Dado o salário/hora,anos de educação e experiencia conseguimos descobrir se a pessoa é do sexo masculino ou feminino?

Modelo Logístico

```
modelo <- train(gender ~ wage + education + experience,
      method = "glm",
      family = "binomial",
      data = CPS1985)
```

Modelo Logístico

```
summary(modelo)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.742  -1.085  -0.673   1.143   3.264
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.440471   0.572072  -2.518 0.011803 *
## wage        -0.134674   0.023724  -5.677 1.37e-08 ***
## education    0.148640   0.043049   3.453 0.000555 ***
## experience   0.029629   0.008334   3.555 0.000378 ***
##
```

Modelo Logístico

Especificação

$$gender = -1.44 - 0.13wage + 0.14educ + 0.02exp$$

Modelo Logístico

Considerações sobre o modelo logístico

- A interpretação dos coeficientes são feitas em forma de probablidade, e temos que passar a fórmula e^{β} para calcula-los

Modelo Logístico

Considerações sobre o modelo logístico

- A interpretação dos coeficientes são feitas em forma de probablidade, e temos que passar a fórmula e^{β} para calcula-los
- O mesmo ocorre com a variavel `gender`, que temos que passar a função $\frac{1}{1+e^{y}}$

Modelo Logístico

Considerações sobre o modelo logístico

- A interpretação dos coeficientes são feitas em forma de probablidade, e temos que passar a fórmula e^{β} para calcula-los
- O mesmo ocorre com a variavel `gender`, que temos que passar a função $\frac{1}{1+e^{y}}$
- Vamos calcular um exemplo

Modelo Logístico

Caso tívessemos uma observação com as seguintes varíaveis

wage $= 5.1$, educ $= 8$, exp $= 21$, qual seria a probabilidade dessa pessoa ser do gênero feminino

Jogando na fórmula

$$-1.44 - 0.13 * 5.1 + 0.14 * 8 + 0.02 * 21 = 0.76$$

Jogando agora na formula $\frac{1}{1+e^{(}0.76)} = 0.31$

A chance de ser do gênero feminino seria de 31%

Modelo Logístico

Agora para isso voltar como uma variavel categórica nós precisamos definir um valor de decisão que varia entre 0 a 1.

- Na maioria dos casos esse valor é 0.5, ou seja.

Modelo Logístico

Agora para isso voltar como uma variavel categórica nós precisamos definir um valor de decisão que varia entre 0 a 1.

- Na maioria dos casos esse valor é 0.5, ou seja.
- Se gender $>$ 0.5, o indivíduo é do genero feminino

Modelo Logístico

Agora para isso voltar como uma variavel categórica nós precisamos definir um valor de decisão que varia entre 0 a 1.

- Na maioria dos casos esse valor é 0.5, ou seja.
- Se gender $>$ 0.5, o indivíduo é do genero feminino
- Se gender $<$ 0.5, o indivíduo é do genero masculino

Modelo Logístico

Podemos automatizar todo esse processo no R com a função `predict`, na qual nos retorna um vetor com a previsão de classificação do nosso data frame.

```
previsao <-predict(modelo,newdata = CPS1985)
```

Modelo Logístico

Verificando a acurácia do modelo, usando uma matriz de confusão

```
table(previsao,CPS1985[,"gender"])
```

```
##
## previsao male female
##   male    218    115
##   female   71    130
```

Essa matriz de confusão nos retorna diversos indicadores de acurácia do nosso modelo.

Para calcular a acurácia geral fazemos o seguinte: $(218 + 130)/\ 534 = 65\%$

Ou seja, nosso modelo acertou no geral 65% das classificações.

Para saber mais sobre essa matriz, clique aqui

Testando premissas do modelo

Padrão dos erros

Todo modelo de linear tem premissas quanto ao seu erro (resíduo), destacando se :

- Indepêndencia
- Homocedasticidade (Variancia constante)
- Segue uma distribuição normal

Caso essas premissas não forem atendidas, podemos ter um modelo ineficiente e/ou tendencioso.

Vamos aos testes

Nosso modelo a ser testado

Especificação

$$wage = \beta_1 + educ\beta_2$$

Usaremos a função `lm` ao invés de `train`

```
modelo <- lm(wage ~ education,
             method = "lm",
             data = CPS1985)
library(lmtest) # Pacote de testes
```

Testando indepêndencia

Iremos usar o pacote `lmtest` para usar a função `bgtest` que aplica o teste de Breusch-Godfrey

H0 = Há independencia

H1 = Não há independencia

Se o valor p $<$ 0.05, rejeitamos H0, caso o contrario, não rejeitamos

```
bgtest(modelo)
```

Testando indepêndencia

```
##
##  Breusch-Godfrey test for serial correlation of order up to
##
## data:  modelo
## LM test = 3.6841, df = 1, p-value = 0.05493
```

Testando Homocedasticidade

Iremos usar o pacote `lmtest` para usar a função `ncvTest`

H0 = Homocedástico H1 = Não Homocedástico (Héterocedástico)

Se o valor p $<$ 0.05, rejeitamos H0, caso o contrario, não rejeitamos

```
ncvTest(modelo)
```

Testando Homocedasticidade

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 40.72868, Df = 1, p = 1.749e-10
```

Testando Normalidade

Iremos a função `shapiro.test` para realizar o teste Shapiro-Wilk

H0 = Normal H1 = Não normal

Se o valor p $<$ 0.05, rejeitamos H0, caso o contrario, não rejeitamos

```
shapiro.test(modelo$residuals)
```

Testando Normalidade

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.90539, p-value < 2.2e-16
```

Exercicios