AULA 5: MODELOS LINEARES GERAIS II

Análise Estatística e Modelagem de Dados Ecológicos

Thiago S. F. Silva - tsfsilva@rc.unesp.br

30 de Março de 2015

Programa de Pós Graduação em Ecologia e Biodiversidade - UNESP

Inferências sobre o modelo

Relembrando: partição de variância

Tabela ANOVA e teste F para regressão

Regressão Múltipla

INFERÊNCIAS SOBRE O MODELO

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i$$

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i$$

Estimador para o valor médio ($E[Y_i]$) da variável **dependente** Y_i

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0$$

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0$$

Estimador do coeficiente **intercepto** (β_0), nos dá valor de $\hat{Y}_i$ quando $X_i = 0$

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0 + b_1$$

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0 + b_1$$

Estimador do coeficiente de **inclinação** (β_1), nos diz o quanto $\hat{Y}$ cresce/diminui quando $\Delta X = 1$

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0 + b_1 X_i$$

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0 + b_1 X_i$$

Variável **independente** ou **preditora**

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0 + b_1 X_i + e_i$$

Quais os componentes de um modelo de regressão simples, estimado a partir de uma amostra de duas variáveis X e Y?

$$\hat{Y}_i = b_0 + b_1 X_i + e_i$$

Resíduo de $\hat{Y}_i$, dado por $Y_i - \hat{Y}_i$, é um estimador dos erros da regressão, ε

O que nós sabemos sobre os resíduos e?

O que nós sabemos sobre os resíduos e?

- Possuem média zero, $E(e) = 0$

O que nós sabemos sobre os resíduos e?

- Possuem média zero, $E(e) = 0$
- Possuem variância constante, $Var(e) = s^2$

O que nós sabemos sobre os resíduos e?

- Possuem média zero, $E(e) = 0$
- Possuem variância constante, $Var(e) = s^2$

O que nós sabemos sobre $\hat{Y}$?

O que nós sabemos sobre os resíduos e?

- Possuem média zero, $E(e) = 0$
- Possuem variância constante, $Var(e) = s^2$

O que nós sabemos sobre $\hat{Y}$?

- Sua média é dada pela equação da reta, $E(\hat{Y}) = b_0 + b_1 X$

O que nós sabemos sobre os resíduos e?

- Possuem média zero, $E(e) = 0$
- Possuem variância constante, $Var(e) = s^2$

O que nós sabemos sobre $\hat{Y}$?

- Sua média é dada pela equação da reta, $E(\hat{Y}) = b_0 + b_1 X$
- Também possuem variância constante, já que $Var(\hat{Y}) = Var(e) = s^2$

O que nós sabemos sobre b_0 e b_1?

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)
- São os melhores estmadores para β_0 e β_1:

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)
- São os melhores estmadores para β_0 e β_1:
    - São estimadores não-tendenciosos

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)
- São os melhores estmadores para β_0 e β_1:
  - São estimadores não-tendenciosos
  - Possuem a menor variância possível dentre todos os estimadores

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)
- São os melhores estmadores para β_0 e β_1:
  - São estimadores não-tendenciosos
  - Possuem a menor variância possível dentre todos os estimadores

O que nós **não** sabemos sobre $\hat{Y}$,b_0, b_1 e e?

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)
- São os melhores estmadores para β_0 e β_1:
  - São estimadores não-tendenciosos
  - Possuem a menor variância possível dentre todos os estimadores

O que nós **não** sabemos sobre $\hat{Y}$,b_0, b_1 e e?

- $\hat{Y} \sim ?(b_0 + b_1 X, s^2)$

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)
- São os melhores estmadores para β_0 e β_1:
  - São estimadores não-tendenciosos
  - Possuem a menor variância possível dentre todos os estimadores

O que nós **não** sabemos sobre $\hat{Y}$,b_0, b_1 e e?

- $\hat{Y} \sim ?(b_0 + b_1 X, s^2)$
- $b_0 \sim ?(b_0, ?)$; $b_1 \sim ?(b_1, ?)$

O que nós sabemos sobre b_0 e b_1?

- São estimados pelo método dos Mínimos Quadrados (*Least Squares*)
- São os melhores estmadores para β_0 e β_1:
  - São estimadores não-tendenciosos
  - Possuem a menor variância possível dentre todos os estimadores

O que nós não sabemos sobre $\hat{Y}$, b_0, b_1 e e?

- $\hat{Y} \sim ?(b_0 + b_1 X, s^2)$
- $b_0 \sim ?(b_0, ?)$; $b_1 \sim ?(b_1, ?)$
- $e \sim ?(0, s^2)$

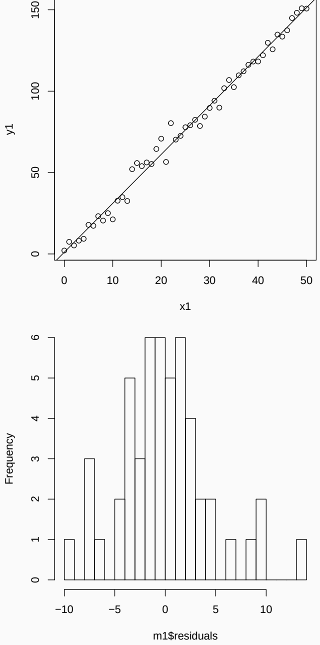
Fato interessante (e importante):

O método dos Mínimos Quadrados **garante** que b_0 e b_1 são os **melhores** estimadores de β_0 e β_1, **qualquer que seja** a distribuição de ε

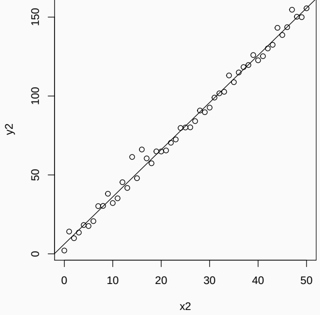
Fato interessante (e importante):

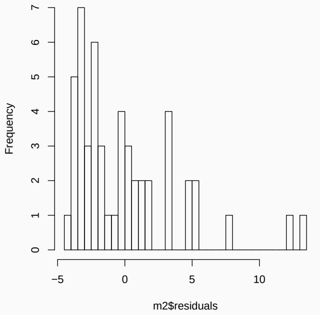
O método dos Mínimos Quadrados **garante** que b_0 e b_1 são os **melhores** estimadores de β_0 e β_1, **qualquer que seja** a distribuição de ε

Isso significa que, independentemente da distribuição dos resíduos (e de Y), a reta de regressão é o melhor estimador para $E(Y)$




```
set.seed(511)
x1 <- c(0:50)
y1 <- 2 + 3*x1 + rnorm(51,0,5)
m1 <- lm(y1 ~ x1)
plot(x1,y1)
abline(m1)
hist(m1$residuals,breaks=30,main=NA)
```






```r
set.seed(511)
x2 <- c(0:50)
y2 <- 2 + 3*x2 + rexp(51,0.2)
m2 <- lm(y2 ~ x2)
plot(x2,y2)
abline(m2)
hist(m2$residuals,breaks=30,main=NA)
```

Contudo, geralmente queremos saber mais do que $\hat{Y}$:

- Qual o grau de certeza sobre $\hat{Y}$?

Contudo, geralmente queremos saber mais do que $\hat{Y}$:

- Qual o grau de certeza sobre $\hat{Y}$?
- Qual o grau de certeza sobre β_0 e β_1?

Contudo, geralmente queremos saber mais do que $\hat{Y}$:

- Qual o grau de certeza sobre $\hat{Y}$?
- Qual o grau de certeza sobre β_0 e β_1?
- Qual o grau de certeza sobre a relação entre X e Y?

Contudo, geralmente queremos saber mais do que $\hat{Y}$:

- Qual o grau de certeza sobre $\hat{Y}$?
- Qual o grau de certeza sobre β_0 e β_1?
- Qual o grau de certeza sobre a relação entre X e Y?
- Com qual grau de certeza posso prever o valor esperado de Y?

Contudo, geralmente queremos saber mais do que $\hat{Y}$:

- Qual o grau de certeza sobre $\hat{Y}$?
- Qual o grau de certeza sobre β_0 e β_1?
- Qual o grau de certeza sobre a relação entre X e Y?
- Com qual grau de certeza posso prever o valor esperado de Y?
- Com qual grau de certeza posso prever um valor aleatório de Y?

Contudo, geralmente queremos saber mais do que $\hat{Y}$:

- Qual o grau de certeza sobre $\hat{Y}$?
- Qual o grau de certeza sobre β_0 e β_1?
- Qual o grau de certeza sobre a relação entre X e Y?
- Com qual grau de certeza posso prever o valor esperado de Y?
- Com qual grau de certeza posso prever um valor aleatório de Y?

Para respondermos à essas perguntas, precisamos especificar uma distribuição de probabilidade para ε

No modelo de regressão clássico, assumimos que esta distribuição é...

Contudo, geralmente queremos saber mais do que $\hat{Y}$:

- Qual o grau de certeza sobre $\hat{Y}$?
- Qual o grau de certeza sobre β_0 e β_1?
- Qual o grau de certeza sobre a relação entre X e Y?
- Com qual grau de certeza posso prever o valor esperado de Y?
- Com qual grau de certeza posso prever um valor aleatório de Y?

Para respondermos à essas perguntas, precisamos especificar uma distribuição de probabilidade para ε

No modelo de regressão clássico, assumimos que esta distribuição é...normal: $\varepsilon \sim N(0, \sigma^2)$

PRESSUPOSIÇÕES DA REGRESSÃO ...*REDUX*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

1. Existe uma relação linear entre X e Y: $E(Y) = \beta_0 + \beta_1 X$.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

1. Existe uma relação linear entre X e Y: $E(Y) = \beta_0 + \beta_1 X$.
2. Os erros ε (e por consequência, Y) tem variância constante (σ^2).

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

1. Existe uma relação linear entre X e Y: $E(Y) = \beta_0 + \beta_1 X$.
2. Os erros ε (e por consequência, Y) tem variância constante (σ^2).
3. Os erros ε (e por consequência, Y) são independentes.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

1. Existe uma relação linear entre X e Y: $E(Y) = \beta_0 + \beta_1 X$.
2. Os erros ε (e por consequência, Y) tem variância constante (σ^2).
3. Os erros ε (e por consequência, Y) são independentes.
4. Os valores de X são medidos sem erro (cada X_i pode ser tratado como constante).

PRESSUPOSIÇÕES DA REGRESSÃO ...*REDUX*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

1. Existe uma relação linear entre X e Y: $E(Y) = \beta_0 + \beta_1 X$.
2. Os erros ε (e por consequência, Y) tem variância constante (σ^2).
3. Os erros ε (e por consequência, Y) são independentes.
4. Os valores de X são medidos sem erro (cada X_i pode ser tratado como constante).
5. Os erros ε (e por consequência, Y) são normalmente distribuídos:
    - $\varepsilon \sim N(0, \sigma^2)$
    - $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

O parâmetro b_1 é um dos mais importantes em um modelo de regressão, porque...

DISTRIBUIÇÃO DE PROBABILIDADE PARA b_1

O parâmetro b_1 é um dos mais importantes em um modelo de regressão, porque...nos diz qual o grau de relação entre X e Y (estima β_1).

O parâmetro b_1 é um dos mais importantes em um modelo de regressão, porque...nos diz qual o grau de relação entre X e Y (estima β_1).

Se $\beta_1 = 0$?

O parâmetro b_1 é um dos mais importantes em um modelo de regressão, porque...nos diz qual o grau de relação entre X e Y (estima β_1).

Se $\beta_1 = 0$?

$$E(Y) = \beta_0 + \varepsilon; \quad E(Y) \sim N(\beta_0, s^2); \quad E(Y) = \beta_0 = \bar{Y}$$

O parâmetro b_1 é um dos mais importantes em um modelo de regressão, porque...nos diz qual o grau de relação entre X e Y (estima β_1).

Se $\beta_1 = 0$?

$E(Y) = \beta_0 + \varepsilon; \quad E(Y) \sim N(\beta_0, s^2); \quad E(Y) = \beta_0 = \bar{Y}$

$E(Y)$ é constante para qualquer nível de X, ou seja, não existe relação.

Analiticamente, nós havíamos determinado que:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Analiticamente, nós havíamos determinado que:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Podemos re-escrever essa equação como:

$$b_1 = \sum k_i Y_i; \quad k_i = \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$$

Analiticamente, nós havíamos determinado que:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Podemos re-escrever essa equação como:

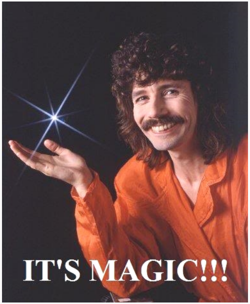
$$b_1 = \sum k_i Y_i; \quad k_i = \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$$

Ou seja, b_1 é uma combinação linear de valores de Y_i

Então, se $Y_i \sim N(\hat{Y}, s^2)$...

Então, se $Y_i \sim N(\hat{Y}, s^2)$...

$$b_1 \sim N\left(\beta_1, \frac{s^2}{\sum(X_i - \bar{X})^2}\right)$$

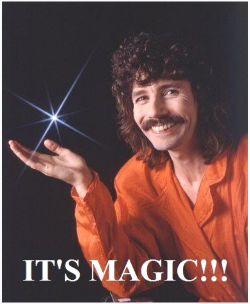


Então, se $Y_i \sim N(\hat{Y}, s^2)$...

$$b_1 \sim N\left(\beta_1, \frac{s^2}{\sum(X_i - \bar{X})^2}\right)$$

Então, se $Y_i \sim N(\hat{Y}, s^2)$...

$$b_1 \sim N\left(\beta_1, \frac{s^2}{\sum(X_i - \bar{X})^2}\right)$$



← Isso também quer dizer que quanto maior a dispersão de X, mais certeza eu tenho sobre b_1

Quem é s^2?

Quem é s^2?

- o nosso estimador de σ^2

Quem é s^2?

- o nosso estimador de σ^2
- A variância dos resíduos

Quem é s^2?

- o nosso estimador de σ^2
- A variância dos resíduos

$$SQ_{res} = \sum_{i-1}^{n} (Y_i - \hat{Y})^2 = \sum_{i-1}^{n} e_i^2$$

Quem é s^2?

- o nosso estimador de σ^2
- A variância dos resíduos

$$SQ_{res} = \sum_{i-1}^{n} (Y_i - \hat{Y})^2 = \sum_{i-1}^{n} e_i^2$$

$$s^2 = MQ_{res} = \frac{SQ_{res}}{n-2} = \frac{\sum (Y_i - \hat{Y})^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Quem é s^2?

- o nosso estimador de σ^2
- A variância dos resíduos

$$SQ_{res} = \sum_{i-1}^{n} (Y_i - \hat{Y})^2 = \sum_{i-1}^{n} e_i^2$$

$$s^2 = MQ_{res} = \frac{SQ_{res}}{n-2} = \frac{\sum (Y_i - \hat{Y})^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Então, calculamos a variância de b_1 como:

$$s_{b_1}^2 = \frac{s^2}{\sum (X_i - \bar{X})^2} = \frac{MQ_{res}}{\sum (X_i - \bar{X})^2}$$

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$\mu =$ média populacional;

$\bar{X} =$ estimador de μ;

$\sigma/\sqrt{n} =$ erro padrão da média populacional (estima $\sqrt{Var(\mu)}$)

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$\mu =$ média populacional;

$\bar{X} =$ estimador de μ;

$\sigma/\sqrt{n} =$ erro padrão da média populacional (estima $\sqrt{Var(\mu)}$)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = Z$$

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$\mu =$ média populacional;

$\bar{X} =$ estimador de μ;

$\sigma/\sqrt{n} =$ erro padrão da média populacional (estima $\sqrt{Var(\mu)}$)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = Z$$

$$Z \sim N(0, 1)$$

Mas nós não conhecemos σ, só conhecemos s (amostra).

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Mas nós não conhecemos σ, só conhecemos s (amostra).

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

$\mu =$ média populacional;

$\bar{X} =$ estimador de μ;

$s/\sqrt{n} =$ erro padrão da média amostral (estima $\sqrt{Var(\bar{X})}$)

Qual a distribuição desta nova variável?

Mas nós não conhecemos σ, só conhecemos s (amostra).

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

$\mu =$ média populacional;

$\bar{X} =$ estimador de μ;

$s/\sqrt{n} =$ erro padrão da média amostral (estima $\sqrt{Var(\bar{X})}$)

Qual a distribuição desta nova variável?

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} = T$$

Mas nós não conhecemos σ, só conhecemos s (amostra).

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

$\mu =$ média populacional;

$\bar{X} =$ estimador de μ;

$s/\sqrt{n} =$ erro padrão da média amostral (estima $\sqrt{Var(\bar{X})}$)

Qual a distribuição desta nova variável?

$$\frac{\bar{Y}-\mu}{s/\sqrt{n}} = T$$

E se ...usarmos b_1 e β_1 em vez de $\bar{X}$ e μ?

E se ...usarmos b_1 e β_1 em vez de $\bar{X}$ e μ?

$$\frac{b_1 - \beta_1}{?} = t$$

E se ...usarmos b_1 e β_1 em vez de $\bar{X}$ e μ?

$$\frac{b_1 - \beta_1}{?} = t$$

$s/\sqrt{n}$ era o estimador de $\sqrt{Var(\bar{X})}$ Quem é o nosso estimador de $\sqrt{s^2_{b_1}}$?

E se ...usarmos b_1 e β_1 em vez de $\bar{X}$ e μ?

$$\frac{b_1 - \beta_1}{?} = t$$

$s/\sqrt{n}$ era o estimador de $\sqrt{Var(\bar{X})}$ Quem é o nosso estimador de $\sqrt{s^2_{b_1}}$?

$$\frac{b_1 - \beta_1}{\sqrt{s^2_{b_1}}} = \frac{b_1 - \beta_1}{\sqrt{\frac{MQ_{res}}{\sum(X_i - \bar{X})^2}}} = t \sim t_{n-2}$$

E se ...usarmos b_1 e β_1 em vez de $\bar{X}$ e μ?

$$\frac{b_1 - \beta_1}{?} = t$$

$s/\sqrt{n}$ era o estimador de $\sqrt{Var(\bar{X})}$ Quem é o nosso estimador de $\sqrt{s^2_{b_1}}$?

$$\frac{b_1 - \beta_1}{\sqrt{s^2_{b_1}}} = \frac{b_1 - \beta_1}{\sqrt{\frac{MQ_{res}}{\sum(X_i - \bar{X})^2}}} = t \sim t_{n-2}$$

Agora podemos fazer inferências sobre β_1!

Intervalos de confiança são uma maneira diferente de expressar o seu grau de certeza sobre as suas estimativas. A sua construção é paramétrica e segue o mesmo procedimento do valor p.

Exemplo: Coletei 20 observações de uma amostra, a qual teve média $\bar{X} = 10$ e desvio padrão $s = 5$. Qual o intervalo de confiança para essa média?

Intervalos de confiança são uma maneira diferente de expressar o seu grau de certeza sobre as suas estimativas. A sua construção é paramétrica e segue o mesmo procedimento do valor p.

Exemplo: Coletei 20 observações de uma amostra, a qual teve média $\bar{X} = 10$ e desvio padrão $s = 5$. Qual o intervalo de confiança para essa média?

Poderíamos assumir uma distribuição Normal: $\bar{X} \sim N(\bar{X}, s/\sqrt{n})$. A partir daí, só precisaríamos calcular os valores correspondentes a $\alpha = 0.05$, bi-caudal.

```r
# Se o teste é bi-caudal, dividimos a probabilidade de 0.05
# entre as duas caudas da distribuição

x_bar <- 10
s <- 5
n <- 20

qnorm(0.025,mean=x_bar,sd=s/sqrt(n))

## [1] 7.808694

qnorm(0.975,mean=x_bar,sd=s/sqrt(n))

## [1] 12.19131
```

Nosso I.C. para $\bar{X}$ é $P(7.8 \leq \bar{X} \leq 12.2) = 0.95$.

Ou seja, se repetíssemos a mesma amostragem infinitamente, a média estaria entre esses valores 95% das vezes.

Poderíamos também assumir uma distribuição t:

$\frac{\bar{X}}{s/\sqrt{n}} \sim t(\nu = n - 1).$

Nesse caso, primeiro calculamos os valores de t que correspondem a $\alpha = 0.05$, e depois multiplicamos esse valor por $s/\sqrt{n}$

```
se <- s/sqrt(n)

x_bar + qt(0.025,n-1)* se

## [1] 7.659928

x_bar + qt(0.975,n-1)* se

## [1] 12.34007
```

O intervalo de confiança para β_1 é:

$$P(b_1 - t^* \times s_{b_1} \leq \beta_1 \leq b_1 + t^* \times s_{b_1}) = 1 - \alpha$$

O intervalo de confiança para β_1 é:

$$P(b_1 - t^* \times s_{b_1} \leq \beta_1 \leq b_1 + t^* \times s_{b_1}) = 1 - \alpha$$

Onde $t^* = t(1 - \alpha/2; n - 2)$

I.C. E TESTE DE HIPÓTESES PARA β_1

O intervalo de confiança para β_1 é:

$$P(b_1 - t^* \times s_{b_1} \leq \beta_1 \leq b_1 + t^* \times s_{b_1}) = 1 - \alpha$$

Onde $t^* = t(1 - \alpha/2; n - 2)$

E podemos testar a hipótese $H_0 : \beta_1 = 0$ usando

$$t^* = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

Quanto menor o valor de p, mais temos certeza sobre o valor estimado de β_1

Assim, como b_1, b_0 é uma combinação linear de valores de Y_i (acreditem)

$$b_0 = \bar{Y} - b_1 X$$

Assim, como b_1, b_0 é uma combinação linear de valores de Y_i (acreditem)

$$b_0 = \bar{Y} - b_1 X$$

Por esse motivo, a distribuição de b_0 também é normal, com:

$$E(b_0) = \beta_0$$

Assim, como b_1, b_0 é uma combinação linear de valores de Y_i (acreditem)

$$b_0 = \bar{Y} - b_1 X$$

Por esse motivo, a distribuição de b_0 também é normal, com:

$$E(b_0) = \beta_0$$

$$s^2_{b_0} = MQ_{res}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)$$

A construção de um I.C. para β_0 é:

$$P(b_0 - t^* \times s_{b_0} \leq \beta_0 \leq b_0 + t^* \times s_{b_0}) = 1 - \alpha c$$

Onde $t^* = t(1 - \alpha/2; n - 2)$

I.C. E TESTE DE HIPÓTESES PARA β_0

A construção de um I.C. para β_0 é:

$$P(b_0 - t^* \times s_{b_0} \leq \beta_0 \leq b_0 + t^* \times s_{b_0}) = 1 - \alpha c$$

Onde $t^* = t(1 - \alpha/2; n - 2)$

E podemos testar a hipótese $H_0 : \beta_0 = 0$ usando

$$t^* = \frac{b_0 - \beta_0}{s_{b_0}} = \frac{b_0 - 0}{s_{b_0}} = \frac{b_0}{s_{b_0}}$$

Quanto menor o valor de p, mais temos certeza sobre o valor estimado de β_1

Adivinhem: $\hat{Y}_h$ (h significando um nível qualquer de X) é…

Adivinhem: $\hat{Y}_h$ (h significando um nível qualquer de X) é...uma combinação linear dos valores de Y_i

Por esse motivo, a distribuição de $\hat{Y}_h$ também é normal, com:

$$E(\hat{Y}_h) = E(Y_h)$$

Adivinhem: $\hat{Y}_h$ (h significando um nível qualquer de X) é...uma combinação linear dos valores de Y_i

Por esse motivo, a distribuição de $\hat{Y}_h$ também é normal, com:

$$E(\hat{Y}_h) = E(Y_h)$$

$$s^2_{\hat{Y}_h} = MQ_{res}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)$$

Adivinhem: $\hat{Y}_h$ (h significando um nível qualquer de X) é…uma combinação linear dos valores de Y_i

Por esse motivo, a distribuição de $\hat{Y}_h$ também é normal, com:

$$E(\hat{Y}_h) = E(Y_h)$$

$$s^2_{\hat{Y}_h} = MQ_{res}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)$$

Atenção!

Quanto mais longe estivermos de $\bar{X}$ ao longo da reta, maior a incerteza sobre $\hat{Y}$!

A construção de um I.C. para $E(Y_h)$ é:

$$P(\hat{Y}_h - t^* \times s_{\hat{Y}_h} \leq E(Y_h) \leq \hat{Y}_h + t^* \times s_{\hat{Y}_h}) = 1 - \alpha$$

Onde $t^* = t(1 - \alpha/2; n - 2)$

A construção de um I.C. para $E(Y_h)$ é:

$$P(\hat{Y}_h - t^* \times s_{\hat{Y}_h} \leq E(Y_h) \leq \hat{Y}_h + t^* \times s_{\hat{Y}_h}) = 1 - \alpha$$

Onde $t^* = t(1 - \alpha/2; n - 2)$

Dificilmente fará sentido testar se $E(Y_h) = 0$, mas podemos testar a hipótese $H_0 : E(Y_h) = H$ para um valor H qualquer, usando

$$t^* = \frac{\hat{Y}_h - H}{s_{\hat{Y}_h}}$$

PREDIÇÃO DE UM NOVO VALOR: $y_{h(novo)}$

Um dos principais usos de um modelo de regressão é a predição de novos valores de Y, dado um certo X_h. Chamaremos esse novo valor de $Y_{h(novo)}$.

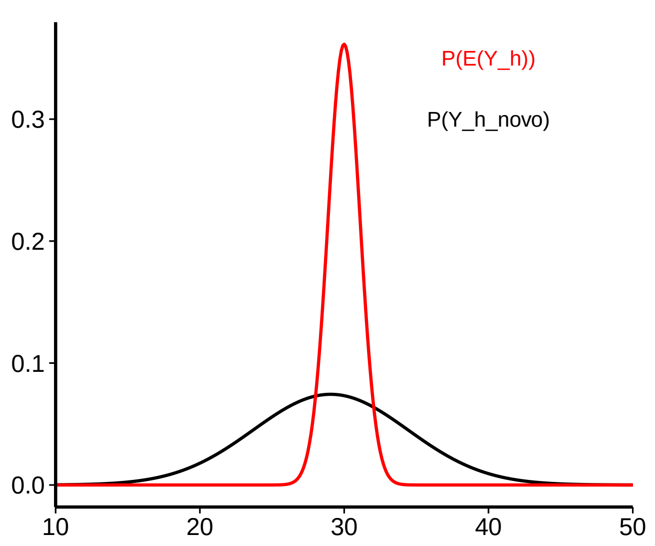
Um dos principais usos de um modelo de regressão é a predição de novos valores de Y, dado um certo X_h. Chamaremos esse novo valor de $Y_{h(novo)}$.

Mas já sabemos calcular o I.C. para $E(Y_h)$...não é a mesma coisa?

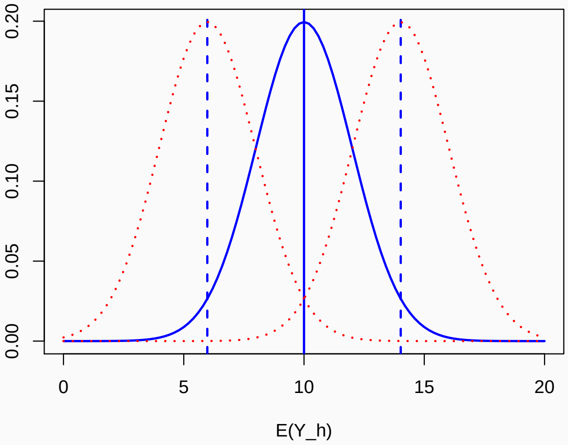
Um dos principais usos de um modelo de regressão é a predição de novos valores de Y, dado um certo X_h. Chamaremos esse novo valor de $Y_{h(novo)}$.

Mas já sabemos calcular o I.C. para $E(Y_h)$...não é a mesma coisa?

No caso de $E(Y_h)$, estamos considerando a distribuição da esperança (média). Quando falamos de $Y_{h(novo)}$, estamos considerando a distribuição **de todos os valores possíveis** de Y para $X = X_h$.


P(E(Y_h))
P(Y_h_novo)
0.3
0.2
0.1
0.0
10
20
30
40
50

DIFERENÇA ENTRE I.C. DE $e(y_h)$ E I.C. DE $y_{h(novo)}$



Utilizando novamente a distribuição t de Student, podemos usar:

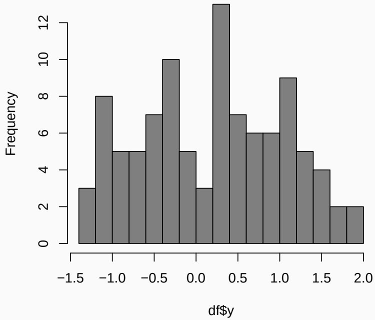
$$t = \frac{Y_{h(novo)} - \hat{Y}_h}{s_{Y_{h(novo)}}}$$

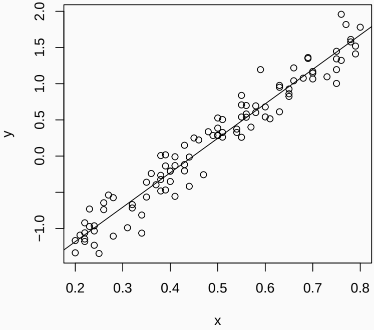
E através das mágicas propriedades da esperança e variância, podemos dizer que:

$$\sigma^2_{Y_{h(novo)}} = \sigma^2 + \sigma^2_{\hat{Y}_h}$$

$$s^2_{Y_{h(novo)}} = MQ_{res}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)$$

Decidimos investigar a relação entre duas variáveis. Primeiro fazemos uma inspeção dos dados:





Os dados aparentam ter uma relação linear, e a variável dependente parece ser normalmente distribuída, então podemos ajustar um modelo de regressão:

Os dados aparentam ter uma relação linear, e a variável dependente parece ser normalmente distribuída, então podemos ajustar um modelo de regressão:

```
m1 <- lm(y ~ x,data=df)
m1$coefficients
```

```
## (Intercept)           x
##   -2.134360    4.760905
```

```
anova(m1)[1:3]
```

```
##           Df Sum Sq Mean Sq
## x          1 69.185  69.185
## Residuals 98  4.040   0.041
```

```
var(df$y)*99
```

```
## [1] 73.22441
```

$$\bar{Y}_i = -2.13 + 4.76 \times X$$

Os dados aparentam ter uma relação linear, e a variável dependente parece ser normalmente distribuída, então podemos ajustar um modelo de regressão:

```
m1 <- lm(y ~ x,data=df)
m1$coefficients
```

```
## (Intercept)           x
##   -2.134360    4.760905
```

```
anova(m1)[1:3]
```

```
##           Df Sum Sq Mean Sq
## x          1 69.185  69.185
## Residuals 98  4.040   0.041
```

```
var(df$y)*99
```

```
## [1] 73.22441
```

$$\bar{Y}_i = -2.13 + 4.76 \times X$$

$$\sum(e_i)^2 = SQ_{res} = 4$$

Os dados aparentam ter uma relação linear, e a variável dependente parece ser normalmente distribuída, então podemos ajustar um modelo de regressão:

```
m1 <- lm(y ~ x,data=df)
m1$coefficients
## (Intercept)           x
##   -2.134360    4.760905
anova(m1)[1:3]
##           Df Sum Sq Mean Sq
## x          1 69.185  69.185
## Residuals 98  4.040   0.041
var(df$y)*99
## [1] 73.22441
```

$$\bar{Y}_i = -2.13 + 4.76 \times X$$
$$\sum(e_i)^2 = SQ_{res} = 4$$
$$\sum(Y_i - \bar{Y})^2 = SQ_{tot} = 73.2$$

Os dados aparentam ter uma relação linear, e a variável dependente parece ser normalmente distribuída, então podemos ajustar um modelo de regressão:

```
m1 <- lm(y ~ x,data=df)
m1$coefficients

## (Intercept)           x
##   -2.134360    4.760905

anova(m1)[1:3]

##           Df Sum Sq Mean Sq
## x          1 69.185  69.185
## Residuals 98  4.040   0.041

var(df$y)*99

## [1] 73.22441
```

$$\bar{Y}_i = -2.13 + 4.76 \times X$$
$$\sum(e_i)^2 = SQ_{res} = 4$$
$$\sum(Y_i - \bar{Y})^2 = SQ_{tot} = 73.2$$
$$SQ_{reg} = SQ_{tot} - SQ_{res} = 69.2$$

Os dados aparentam ter uma relação linear, e a variável dependente parece ser normalmente distribuída, então podemos ajustar um modelo de regressão:

```r
m1 <- lm(y ~ x,data=df)
m1$coefficients
```

```
## (Intercept)           x
##   -2.134360    4.760905
```

```r
anova(m1)[1:3]
```

```
##           Df Sum Sq Mean Sq
## x          1 69.185  69.185
## Residuals 98  4.040   0.041
```

```r
var(df$y)*99
```

```
## [1] 73.22441
```

$$\bar{Y}_i = -2.13 + 4.76 \times X$$
$$\sum(e_i)^2 = SQ_{res} = 4$$
$$\sum(Y_i - \bar{Y})^2 = SQ_{tot} = 73.2$$
$$SQ_{reg} = SQ_{tot} - SQ_{res} = 69.2$$
$$\sum(X_i - \bar{X})^2 = 3.05$$
$$\bar{X} = 0.489$$

1) Qual o r^2 desse modelo?

1) Qual o r^2 desse modelo?

$$r^2 = \frac{SQ_{reg}}{SQ_{tot}} = \frac{69.2}{73.2} = 0.945$$

1) Qual o r^2 desse modelo?

$$r^2 = \frac{SQ_{reg}}{SQ_{tot}} = \frac{69.2}{73.2} = 0.945$$

```
summary(miaf)$r.squared

## [1] 0.9448324
```

2) Qual o I.C. para β_1, quando $\alpha = 0.5$?

2) Qual o I.C. para β_1, quando $\alpha = 0.5$?

$\frac{b_1 - \beta_1}{s_{b_1}} \sim t(n-2); \quad H_0 : \beta_1 = 0; \quad t^*_{(\alpha/2, n-2)} =?$

2) Qual o I.C. para β_1, quando $\alpha = 0.5$?

$$\frac{b_1 - \beta_1}{s_{b_1}} \sim t(n-2); \quad H_0 : \beta_1 = 0; \quad t^*_{(\alpha/2, n-2)} = ?$$

$$s_{b_1} = \sqrt{\frac{MQ_{res}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{SQ_{res}}{(n-2)}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{4}{98}}{3.05}} = \sqrt{0.013} = 0.116$$

2) Qual o I.C. para β_1, quando $\alpha = 0.5$?

$\frac{b_1-\beta_1}{s_{b_1}} \sim t(n-2); \quad H_0 : \beta_1 = 0; \quad t^*_{(\alpha/2, n-2)} =?$

$$s_{b_1} = \sqrt{\frac{MQ_{res}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{SQ_{res}}{(n-2)}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{4}{98}}{3.05}} = \sqrt{0.013} = 0.116$$

$$P(b_1 - t^* \times s_{b_1} \leq \beta_1 \leq b_1 + t^* \times s_{b_1}) = 1 - \alpha$$

2) Qual o I.C. para β_1, quando $\alpha = 0.5$?

$\frac{b_1 - \beta_1}{s_{b_1}} \sim t(n-2); \quad H_0 : \beta_1 = 0; \quad t^*_{(\alpha/2, n-2)} = ?$

$$s_{b_1} = \sqrt{\frac{MQ_{res}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{SQ_{res}}{(n-2)}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{4}{98}}{3.05}} = \sqrt{0.013} = 0.116$$

$$P(b_1 - t^* \times s_{b_1} \leq \beta_1 \leq b_1 + t^* \times s_{b_1}) = 1 - \alpha$$

$$P(4.76 - 1.98 \times 0.116 \leq \beta_1 \leq 4.76 + 1.98 \times 0.116) = 0.95$$

2) Qual o I.C. para β_1, quando $\alpha = 0.5$?

$\frac{b_1 - \beta_1}{s_{b_1}} \sim t(n-2); \quad H_0 : \beta_1 = 0; \quad t^*_{(\alpha/2, n-2)} = ?$

$$s_{b_1} = \sqrt{\frac{MQ_{res}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{SQ_{res}}{(n-2)}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\frac{4}{98}}{3.05}} = \sqrt{0.013} = 0.116$$

$$P(b_1 - t^* \times s_{b_1} \leq \beta_1 \leq b_1 + t^* \times s_{b_1}) = 1 - \alpha$$

$$P(4.76 - 1.98 \times 0.116 \leq \beta_1 \leq 4.76 + 1.98 \times 0.116) = 0.95$$

$$P(4.53 \leq \beta_1 \leq 4.99) = 0.95$$

EXEMPLO

```r
summary(m1)$coefficients

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.134360 0.06031152 -35.38893 1.256108e-57
## x            4.760905 0.11620932  40.96836 1.818830e-63

4.76 - 1.98*0.11621

## [1] 4.529904

4.76 + 1.98*0.11621

## [1] 4.990096

confint(m1)

##                  2.5 %    97.5 %
## (Intercept) -2.254047 -2.014674
## x            4.530292  4.991519
```

3) Qual o I.C. para β_0, quando $\alpha = 0.5$?

3) Qual o I.C. para β_0, quando $\alpha = 0.5$?

$\frac{b_0 - \beta_0}{s_{b_0}} \sim t(n-2); \quad H_0 : \beta_0 = 0; \quad t^*_{(\alpha/2, n-2)} = ?$

3) Qual o I.C. para β_0, quando $\alpha = 0.5$?

$$\frac{b_0-\beta_0}{s_{b_0}} \sim t(n-2); \quad H_0 : \beta_0 = 0; \quad t^*_{(\alpha/2,n-2)} =?$$

$$s_{b_0} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{0.24}{3.05}\right)} = 0.06$$

3) Qual o I.C. para β_0, quando $\alpha = 0.5$?

$\frac{b_0-\beta_0}{s_{b_0}} \sim t(n-2); \quad H_0 : \beta_0 = 0; \quad t^*_{(\alpha/2, n-2)} =?$

$$s_{b_0} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{0.24}{3.05}\right)} = 0.06$$

$$P(b_0 - t^* \times s_{b_0} \leq \beta_0 \leq b_0 + t^* \times s_{b_0}) = 1 - \alpha$$

3) Qual o I.C. para β_0, quando $\alpha = 0.5$?

$$\frac{b_0-\beta_0}{s_{b_0}} \sim t(n-2); \quad H_0: \beta_0 = 0; \quad t^*_{(\alpha/2, n-2)} = ?$$

$$s_{b_0} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{0.24}{3.05}\right)} = 0.06$$

$$P(b_0 - t^* \times s_{b_0} \leq \beta_0 \leq b_0 + t^* \times s_{b_0}) = 1 - \alpha$$

$$P(-2.13 - 1.98 \times 0.06 \leq \beta_1 \leq -2.13 + 1.98 \times 0.06) = 0.95$$

3) Qual o I.C. para β_0, quando $\alpha = 0.5$?

$$\frac{b_0-\beta_0}{s_{b_0}} \sim t(n-2); \quad H_0 : \beta_0 = 0; \quad t^*_{(\alpha/2, n-2)} = ?$$

$$s_{b_0} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{0.24}{3.05}\right)} = 0.06$$

$$P(b_0 - t^* \times s_{b_0} \leq \beta_0 \leq b_0 + t^* \times s_{b_0}) = 1 - \alpha$$

$$P(-2.13 - 1.98 \times 0.06 \leq \beta_1 \leq -2.13 + 1.98 \times 0.06) = 0.95$$

$$P(-2.25 \leq \beta_0 \leq -2.01) = 0.95$$

EXEMPLO

```r
summary(miaf)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.134360 0.06031152 -35.38893 1.256108e-57
## x            4.760905 0.11620932  40.96836 1.818830e-63
```

```r
-2.13 - 1.98*0.06031
```

```
## [1] -2.249414
```

```r
-2.13 + 1.98*0.06031
```

```
## [1] -2.010586
```

```r
confint(m1)
```

```
##                 2.5 %    97.5 %
## (Intercept) -2.254047 -2.014674
## x            4.530292  4.991519
```

4) Qual o valor esperado de $\hat{Y}_h$ e seu I.C. quando $X_h = 0.5$?

4) Qual o valor esperado de $\hat{Y}_h$ e seu I.C. quando $X_h = 0.5$?

$\frac{\hat{Y}_h - E(Y_h)}{s_{\hat{Y}_h}} \sim t(n-2); \quad H_0 : E(Y_h) = 0; \quad t^*_{(\alpha/2, n-2)} =?$

4) Qual o valor esperado de $\hat{Y}_h$ e seu I.C. quando $X_h = 0.5$?

$\frac{\hat{Y}_h - E(Y_h)}{s_{\hat{Y}_h}} \sim t(n-2); \quad H_0 : E(Y_h) = 0; \quad t^*_{(\alpha/2, n-2)} = ?$

$s_{\hat{Y}_h} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{(0.5-0.489)^2}{3.05}\right)} = 0.02$

4) Qual o valor esperado de $\hat{Y}_h$ e seu I.C. quando $X_h = 0.5$?

$\frac{\hat{Y}_h - E(Y_h)}{s_{\hat{Y}_h}} \sim t(n-2); \quad H_0 : E(Y_h) = 0; \quad t^*_{(\alpha/2, n-2)} = ?$

$s_{\hat{Y}_h} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{(0.5 - 0.489)^2}{3.05}\right)} = 0.02$

$P(\hat{Y}_h - t^* \times s_{\hat{Y}_h} \leq \hat{E}(Y_h) \leq \hat{Y}_h + t^* \times s_{\hat{Y}_h}) = 1 - \alpha$

4) Qual o valor esperado de $\hat{Y}_h$ e seu I.C. quando $X_h = 0.5$?

$\frac{\hat{Y}_h - E(Y_h)}{s_{\hat{Y}_h}} \sim t(n-2); \quad H_0 : E(Y_h) = 0; \quad t^*_{(\alpha/2, n-2)} =?$

$s_{\hat{Y}_h} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{(0.5-0.489)^2}{3.05}\right)} = 0.02$

$P(\hat{Y}_h - t^* \times s_{\hat{Y}_h} \leq \hat{E}(Y_h) \leq \hat{Y}_h + t^* \times s_{\hat{Y}_h}) = 1 - \alpha$

$P(0.25 - 1.98 \times 0.02 \leq E(Y_h) \leq 0.25 + 1.98 \times 0.02) = 0.95$

4) Qual o valor esperado de $\hat{Y}_h$ e seu I.C. quando $X_h = 0.5$?

$\frac{\hat{Y}_h - E(Y_h)}{s_{\hat{Y}_h}} \sim t(n-2); \quad H_0 : E(Y_h) = 0; \quad t^*_{(\alpha/2, n-2)} =?$

$s_{\hat{Y}_h} = \sqrt{MQ_{res}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} = \sqrt{\frac{4}{98}\left(\frac{1}{100} + \frac{(0.5-0.489)^2}{3.05}\right)} = 0.02$

$P(\hat{Y}_h - t^* \times s_{\hat{Y}_h} \leq \hat{E}(Y_h) \leq \hat{Y}_h + t^* \times s_{\hat{Y}_h}) = 1 - \alpha$

$P(0.25 - 1.98 \times 0.02 \leq E(Y_h) \leq 0.25 + 1.98 \times 0.02) = 0.95$

$P(0.21 \leq E(Y_h) \leq 0.28) = 0.95$

5) Qual o intervalo de predição de $\hat{Y}_{h(novo)}$ quando $X_h = 0.5$?

5) Qual o intervalo de predição de $\hat{Y}_{h(novo)}$ quando $X_h = 0.5$?

$$s_{\hat{Y}_{h(novo)}} = \sqrt{MQ_{res}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} =$$

$$\sqrt{\frac{4}{98}\left(1 + \frac{1}{100} + \frac{(0.5 - 0.489)^2}{3.05}\right)} = 0.203$$

5) Qual o intervalo de predição de $\hat{Y}_{h(novo)}$ quando $X_h = 0.5$?

s$_{\hat{Y}_{h(novo)}} = \sqrt{MQ_{res}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} =$
$\sqrt{\frac{4}{98}\left(1 + \frac{1}{100} + \frac{(0.5 - 0.489)^2}{3.05}\right)} = 0.203$

$$P(\hat{Y}_h - t^* \times s_{\hat{Y}_{h(novo)}} \leq Y_{h(novo)} \leq \hat{Y}_h + t^* \times s_{\hat{Y}_{h(novo)}}) = 1 - \alpha$$

5) Qual o intervalo de predição de $\hat{Y}_{h(novo)}$ quando $X_h = 0.5$?

s $_{\hat{Y}_{h(novo)}} = \sqrt{MQ_{res}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} =$
$\sqrt{\frac{4}{98}\left(1 + \frac{1}{100} + \frac{(0.5-0.489)^2}{3.05}\right)} = 0.203$

$$P(\hat{Y}_h - t^* \times s_{\hat{Y}_{h(novo)}} \leq Y_{h(novo)} \leq \hat{Y}_h + t^* \times s_{\hat{Y}_{h(novo)}}) = 1 - \alpha$$

$$P(0.25 - 1.98 \times 0.203 \leq Y_{h(novo)} \leq 0.25 + 1.98 \times 0.203) = 0.95$$

5) Qual o intervalo de predição de $\hat{Y}_{h(novo)}$ quando $X_h = 0.5$?

$$s_{\hat{Y}_{h(novo)}} = \sqrt{MQ_{res}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)} =$$

$$\sqrt{\frac{4}{98}\left(1 + \frac{1}{100} + \frac{(0.5 - 0.489)^2}{3.05}\right)} = 0.203$$

$$P(\hat{Y}_h - t^* \times s_{\hat{Y}_{h(novo)}} \leq Y_{h(novo)} \leq \hat{Y}_h + t^* \times s_{\hat{Y}_{h(novo)}}) = 1 - \alpha$$

$$P(0.25 - 1.98 \times 0.203 \leq Y_{h(novo)} \leq 0.25 + 1.98 \times 0.203) = 0.95$$

$$P(-0.152 \leq Y_{h(novo)} \leq 0.652) = 0.95$$

$$P(0.21 \leq E(Y_h) \leq 0.28) = 0.95$$

$$P(-0.152 \leq Y_{h(novo)} \leq 0.652) = 0.95$$

```
predict(m1,newobs,interval='confidence')

##         fit       lwr       upr
## 1 0.2460923 0.2057178 0.2864668

predict(m1,newobs,interval='prediction')

##         fit        lwr       upr
## 1 0.2460923 -0.1588289 0.6510134
```

```r
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54883 -0.11846  0.01314  0.14513  0.51984
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.13436    0.06031  -35.39   <2e-16 ***
## x            4.76091    0.11621   40.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.203 on 98 degrees of freedom
## Multiple R-squared:  0.9448,^^IAdjusted R-squared:  0.9443
## F-statistic:  1678 on 1 and 98 DF,  p-value: < 2.2e-16
```

- Inferências sobre β_1 nos dizem a certeza acerca da relação entre X e Y

- Inferências sobre β_1 nos dizem a certeza acerca da relação entre X e Y
- Inferências sobre β_0 nos dizem se a reta passa pela origem ($\beta_0 = 0$)

OK, PRA QUE SERVE TUDO ISSO?

- Inferências sobre β_1 nos dizem a certeza acerca da relação entre X e Y
- Inferências sobre β_0 nos dizem se a reta passa pela origem ($\beta_0 = 0$)
- Inferências sobre $E(Y_h)$ nos dizem a certeza acerca do valor esperado de Y_i para um dado X_i ($E[Y_i]$)

OK, PRA QUE SERVE TUDO ISSO?

- Inferências sobre β_1 nos dizem a certeza acerca da relação entre X e Y
- Inferências sobre β_0 nos dizem se a reta passa pela origem ($\beta_0 = 0$)
- Inferências sobre $E(Y_h)$ nos dizem a certeza acerca do valor esperado de Y_i para um dado X_i ($E[Y_i]$)
- Inferências sobre $Y_{h(novo)}$ nos dizem a certeza acerca de um valor aleatório de Y_i, para um dado X_i

- Inferências sobre β_1 nos dizem a certeza acerca da relação entre X e Y
- Inferências sobre β_0 nos dizem se a reta passa pela origem ($\beta_0 = 0$)
- Inferências sobre $E(Y_h)$ nos dizem a certeza acerca do valor esperado de Y_i para um dado X_i ($E[Y_i]$)
- Inferências sobre $Y_{h(novo)}$ nos dizem a certeza acerca de um valor aleatório de Y_i, para um dado X_i
- Só isso!

RELEMBRANDO: PARTIÇÃO DE VARIÂNCIA

Um dos objetivos de um modelo de regressão linear é quantificar o quanto da variação total em Y nós podemos explicar através de X

Para isso, usamos o método de partição de variâncias: SQ_{tot}, SQ_{res} e SQ_{reg}

- Soma dos Quadrados Totais?

Um dos objetivos de um modelo de regressão linear é quantificar o quanto da variação total em Y nós podemos explicar através de X

Para isso, usamos o método de partição de variâncias: SQ_{tot}, SQ_{res} e SQ_{reg}

- Soma dos Quadrados Totais? $\sum(Y_i - \bar{Y})^2$

Um dos objetivos de um modelo de regressão linear é quantificar o quanto da variação total em Y nós podemos explicar através de X

Para isso, usamos o método de partição de variâncias: SQ_{tot}, SQ_{res} e SQ_{reg}

- Soma dos Quadrados Totais? $\sum(Y_i - \bar{Y})^2$
- Soma dos Quadrados dos Erros?

RELEMBRANDO: PARTIÇÃO DE VARIÂNCIA

Um dos objetivos de um modelo de regressão linear é quantificar o quanto da variação total em Y nós podemos explicar através de X

Para isso, usamos o método de partição de variâncias: SQ_{tot}, SQ_{res} e SQ_{reg}

- Soma dos Quadrados Totais? $\sum(Y_i - \bar{Y})^2$
- Soma dos Quadrados dos Erros? $\sum(Y_i - \hat{Y}_i)^2$

Um dos objetivos de um modelo de regressão linear é quantificar o quanto da variação total em Y nós podemos explicar através de X

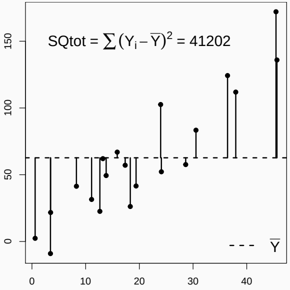
Para isso, usamos o método de partição de variâncias: SQ_{tot}, SQ_{res} e SQ_{reg}

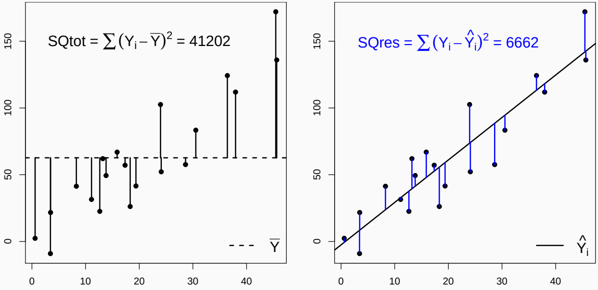
- Soma dos Quadrados Totais? $\sum(Y_i - \bar{Y})^2$
- Soma dos Quadrados dos Erros? $\sum(Y_i - \hat{Y}_i)^2$
- Soma dos Quadrados da Regressão?

Um dos objetivos de um modelo de regressão linear é quantificar o quanto da variação total em Y nós podemos explicar através de X

Para isso, usamos o método de partição de variâncias: SQ_{tot}, SQ_{res} e SQ_{reg}

- Soma dos Quadrados Totais? $\sum(Y_i - \bar{Y})^2$
- Soma dos Quadrados dos Erros? $\sum(Y_i - \hat{Y}_i)^2$
- Soma dos Quadrados da Regressão? $\sum(\hat{Y}_i - \bar{Y})^2$


$SQtot = \sum(Y_i - \overline{Y})^2 = 41202$
$\overline{Y}$
0
50
100
150
0
10
20
30
40



SQtot = $\sum(Y_i - \overline{Y})^2$ = 41202
150
100
50
0
0
10
20
30
40
$\overline{Y}$
SQres = $\sum(Y_i - \hat{Y}_i)^2$ = 6662
150
100
50
0
0
10
20
30
40
$\hat{Y}_i$

RELEMBRANDO: PARTIÇÃO DE VARIÂNCIA

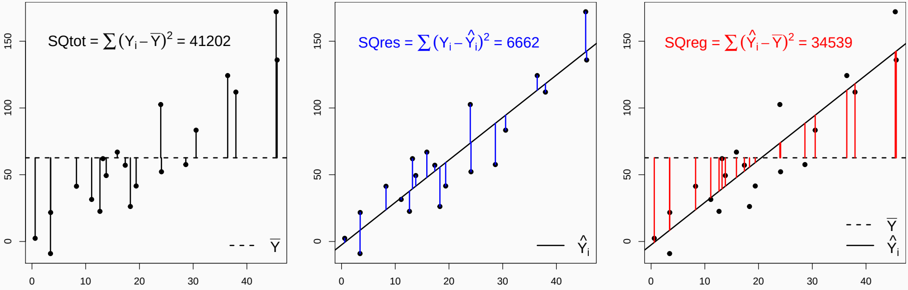


TABELA ANOVA E TESTE F PARA REGRESSÃO

Uma maneira comum para mostrar a partição dos erros em um modelo de regressão é a **tabela ANOVA**

Uma maneira comum para mostrar a partição dos erros em um modelo de regressão é a **tabela ANOVA**

ANOVA significa *ANalysis Of VAriance* (Análise de Variância)

Uma maneira comum para mostrar a partição dos erros em um modelo de regressão é a **tabela ANOVA**

ANOVA significa *ANalysis Of VAriance* (Análise de Variância)

A ANOVA pode ser vista como uma ”versão”da regressão, um método para particionar as variâncias quando X é categórico

Uma maneira comum para mostrar a partição dos erros em um modelo de regressão é a **tabela ANOVA**

ANOVA significa *ANalysis Of VAriance* (Análise de Variância)

A ANOVA pode ser vista como uma ”versão”da regressão, um método para particionar as variâncias quando X é categórico

De fato, ANOVA e Regressão Linear são casos específicos dos chamados Modelos Lineares Gerais (e um teste t é uma ANOVA com apenas dois tratamentos).

Elementos de uma tabela ANOVA completa

Fonte	GL	Soma Quadrados	Média Quadrados	E(Med. Quad.)	F	P
Regressão	1	$SQ_{reg} = \sum(\hat{Y}_i - \bar{Y})^2$	$MQR = \frac{SQ_{reg}}{1}$	$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$	$\frac{MQR}{MQE}$	P($F_{(1,n-2)}$)
Resíduos	$n-2$	$SQ_{res} = \sum(Y_i - \hat{Y}_i)^2$	$MQE = \frac{SQ_{res}}{n-2}$	σ^2		
Total	$n-1$	$SQ_{tot} = \sum(Y_i - \bar{Y})^2$	$MQT = \frac{SQ_{tot}}{n-1}$	σ_Y^2		

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\sum(X_i - \mu)^2}{\sigma^2} =$$

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\sum(X_i - \mu)^2}{\sigma^2} = \chi^2_n$$

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\sum(X_i - \mu)^2}{\sigma^2} = \chi^2_n$$

E qual a distribuição de:

$$\frac{\sum(X_i - \bar{X})^2}{s^2} =$$

Se $X \sim N(\mu, \sigma^2)$, qual a distribuição de:

$$\frac{\sum(X_i - \mu)^2}{\sigma^2} = \chi^2_n$$

E qual a distribuição de:

$$\frac{\sum(X_i - \bar{X})^2}{s^2} = \chi^2_{n-1}$$

E se $U_1 \sim \chi^2_{gl_1}$ e $U_2 \sim \chi^2_{gl_1}$, qual a distribuição de:

$$\frac{U_1/gl_1}{U_2/gl_2} =$$

E se $U_1 \sim \chi^2_{gl_1}$ e $U_2 \sim \chi^2_{gl_1}$, qual a distribuição de:

$$\frac{U_1/gl_1}{U_2/gl_2} = F_{gl_1,gl_2}$$

$$U = \frac{\sum(X_i - \bar{X})^2}{s^2} \sim \chi^2_{n-1} \quad ; \quad \frac{U_1/gl_1}{U_2/gl_2} \sim F_{gl_1,gl_2}$$

$$U = \frac{\sum(X_i - \bar{X})^2}{s^2} \sim \chi^2_{n-1} \quad ; \quad \frac{U_1/gl_1}{U_2/gl_2} \sim F_{gl_1,gl_2}$$

Se a variância é constante, então:

$$U = \frac{\sum(X_i - \bar{X})^2}{s^2} \sim \chi^2_{n-1} \quad ; \quad \frac{U_1/gl_1}{U_2/gl_2} \sim F_{gl_1,gl_2}$$

Se a variância é constante, então:

$$\frac{\frac{\sum(X_{i(1)} - \bar{X}_1)^2}{s^2}}{gl_1} \div \frac{\frac{\sum(X_{i(2)} - \bar{X}_2)^2}{s^2}}{gl_2} =$$

$$U = \frac{\sum(X_i - \bar{X})^2}{s^2} \sim \chi^2_{n-1} \quad ; \quad \frac{U_1/gl_1}{U_2/gl_2} \sim F_{gl_1,gl_2}$$

Se a variância é constante, então:

$$\frac{\frac{\sum(X_{i(1)} - \bar{X}_1)^2}{s^2}}{gl_1} \div \frac{\frac{\sum(X_{i(2)} - \bar{X}_2)^2}{s^2}}{gl_2} =$$

$$\frac{\sum(X_{i(1)} - \bar{X}_1)^2}{gl_1} \div \frac{\sum(X_{i(2)} - \bar{X}_2)^2}{gl_2} \sim F_{gl_1,gl_2}$$

HMMMMMMMMMMMMMMMMMMMMMM....

$$\frac{\sum(X_{i(1)} - \bar{X}_1)^2}{gl_1} \div \frac{\sum(X_{i(2)} - \bar{X}_2)^2}{gl_2} \sim F_{gl_1, gl_2}$$

$$\frac{\sum(X_{i(1)} - \bar{X}_1)^2}{gl_1} \div \frac{\sum(X_{i(2)} - \bar{X}_2)^2}{gl_2} \sim F_{gl_1, gl_2}$$

$$MQ_{reg} = \frac{SQ_{reg}}{gl_{reg}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{1}$$

HMMMMMMMMMMMMMMMMMMMMMM....

$$\frac{\sum(X_{i(1)} - \bar{X}_1)^2}{gl_1} \div \frac{\sum(X_{i(2)} - \bar{X}_2)^2}{gl_2} \sim F_{gl_1, gl_2}$$

$$MQ_{reg} = \frac{SQ_{reg}}{gl_{reg}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{1}$$

$$MQ_{res} = \frac{SQ_{res}}{gl_{erro}} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$$

HMMMMMMMMMMMMMMMMMMMMMM....

$$\frac{\sum(X_{i(1)} - \bar{X}_1)^2}{gl_1} \div \frac{\sum(X_{i(2)} - \bar{X}_2)^2}{gl_2} \sim F_{gl_1, gl_2}$$

$$MQ_{reg} = \frac{SQ_{reg}}{gl_{reg}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{1}$$

$$MQ_{res} = \frac{SQ_{res}}{gl_{erro}} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$$

$$\frac{MQ_{reg}}{MQ_{res}} \sim F_{(1, n-2)}$$

Na tabela ANOVA, havíamos visto que $E(MQ_{reg})$ era:

$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

Na tabela ANOVA, havíamos visto que $E(MQ_{reg})$ era:

$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

E sabemos que $E(MQ_{res}) = \sigma^2$

Na tabela ANOVA, havíamos visto que $E(MQ_{reg})$ era:

$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

E sabemos que $E(MQ_{res}) = \sigma^2$

Se não existe relação entre X e Y, então $\beta_1 = 0$ e temos:

$$F = \frac{MQ_{reg}}{MQ_{res}} = \frac{\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

Na tabela ANOVA, havíamos visto que $E(MQ_{reg})$ era:

$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

E sabemos que $E(MQ_{res}) = \sigma^2$

Se não existe relação entre X e Y, então $\beta_1 = 0$ e temos:

$$F = \frac{MQ_{reg}}{MQ_{res}} = \frac{\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

Mas se essa relação existe, então $\beta_1 > 0$, e $F > 1$

E assim, podemos avaliar o grau de evidência de que $H_0 : \beta_1 = 0$ é verdadeira, dado o nosso modelo:

$$F^* = \frac{MQ_{reg}}{MQ_{res}}$$

O nosso valor p é $P(F(1, n-2)|H_0)$. Ou "qual a probabilidade de observarmos essa proporção entre MQ_{reg} e MQ_{res} se o nosso modelo na verdade não explica nada?".

```r
anova(m)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1  34539   34539  93.316 1.517e-08 ***
## Residuals 18   6662     370
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F^{*} = \frac{MQ_{reg}}{MQ_{res}} = \frac{34539}{370} = 93.3$$

```
anova(m)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1  34539   34539  93.316 1.517e-08 ***
## Residuals 18   6662     370
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F^* = \frac{MQ_{reg}}{MQ_{res}} = \frac{34539}{370} = 93.3$$

$$P(F^*|H_0) = 1.52 \times 10^{-8}$$

REGRESSÃO MÚLTIPLA

O modelo de regressão múltipla é uma extensão do modelo simples.

Para duas variáveis explicativas, temos:

O modelo de regressão múltipla é uma extensão do modelo simples.

Para duas variáveis explicativas, temos:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

O modelo de regressão múltipla é uma extensão do modelo simples.

Para duas variáveis explicativas, temos:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

Os termos fixos nos dão $E[Y]$, e o termo aleatório nos dá $Var[Y]$.

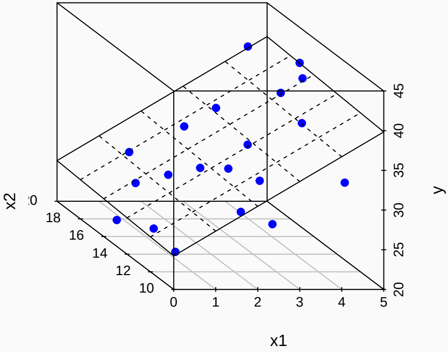
O modelo de regressão múltipla é uma extensão do modelo simples.

Para duas variáveis explicativas, temos:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

Os termos fixos nos dão $E[Y]$, e o termo aleatório nos dá $Var[Y]$.

Ex.: Se $E[Y]$ depende de uma combinação de duas variáveis preditoras (X_1 a X_2), a reta se torna um plano



x2
20
18
16
14
12
10
0
1
2
3
4
5
x1
20
25
30
35
40
45
y

- β_0: intercepto da superfície de resposta. Valor de Y quando $X_1 = X_2 = \ldots X_{k=(p-1)} = 0$. Geralmente não tem um significado explícito.

- β_0: intercepto da superfície de resposta. Valor de Y quando $X_1 = X_2 = \ldots X_{k=(p-1)} = 0$. Geralmente não tem um significado explícito.
- β_1, β_2, ..., β_k: determinam o aumento em $E(Y)$ quando X_k $(k = \{0, p-1\})$ aumenta em 1, e os demais X_k permanecem constantes.

- β_0: intercepto da superfície de resposta. Valor de Y quando $X_1 = X_2 = \ldots X_{k=(p-1)} = 0$. Geralmente não tem um significado explícito.
- β_1, β_2, ..., β_k: determinam o aumento em $E(Y)$ quando X_k $(k = \{0, p-1\})$ aumenta em 1, e os demais X_k permanecem constantes.
- Cada coeficiente representa a contribuição absoluta de X_k para a estimativa de $E[Y]$ (ou $\beta_k = \frac{\delta E[Y]}{\delta X_{(k)}}$)

- β_0: intercepto da superfície de resposta. Valor de Y quando $X_1 = X_2 = \ldots X_{k=(p-1)} = 0$. Geralmente não tem um significado explícito.
- β_1, β_2, ..., β_k: determinam o aumento em $E(Y)$ quando X_k $(k = \{0, p-1\})$ aumenta em 1, e os demais X_k permanecem constantes.
- Cada coeficiente representa a contribuição absoluta de X_k para a estimativa de $E[Y]$ (ou $\beta_k = \frac{\delta E[Y]}{\delta X_{(k)}}$)
- ε_i continua sendo a diferença entre Y_i e $E[Y_i]$

A partição geral da variância segue o mesmo padrão do modelo simples, mas com diferentes graus de liberdade

Fonte	GL	Soma Quadrados	Média Quadrados
Regressão	$p-1$	$SQ_{reg} = \mathbf{b'X'Y} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{Y'JY}$	$MQ_{reg} = \frac{SQ_{reg}}{p-1}$
Resíduos	$n-p$	$SQ_{res} = \mathbf{Y'Y} - \mathbf{b'X'Y}$	$MQ_{res} = \frac{SQ_{res}}{n-p}$
Total	$n-1$	$SQ_{tot} = \mathbf{Y'Y} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{Y'JY}$	$MQ_{tot} = \frac{SQ_{tot}}{n-1}$

- O teste geral para a regressão ainda é feito usando $F^{*} = \frac{MS_{reg}}{MS_{res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{erro}}{SQ_{tot}}$

- O teste geral para a regressão ainda é feito usando $F^* = \frac{MS_{reg}}{MS_{res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{erro}}{SQ_{tot}}$
- Quando novas variáveis são incluídas no modelo, SQ_{res} permanece o mesmo ou diminui, mas nunca aumenta.

- O teste geral para a regressão ainda é feito usando $F^* = \frac{MS_{reg}}{MS_{res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{erro}}{SQ_{tot}}$
- Quando novas variáveis são incluídas no modelo, SQ_{res} permanece o mesmo ou diminui, mas nunca aumenta.

  Por esse motivo, o R^2 aumenta mesmo que a quantidade de variância adicional explicada seja mínima.

- O teste geral para a regressão ainda é feito usando $F^* = \frac{MS_{reg}}{MS_{res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{erro}}{SQ_{tot}}$
- Quando novas variáveis são incluídas no modelo, SQ_{res} permanece o mesmo ou diminui, mas nunca aumenta.

  Por esse motivo, o R^2 aumenta mesmo que a quantidade de variância adicional explicada seja mínima.
- Assim, não se pode confiar em R^2 como uma medida de qualidade do modelo (a interpretação de quantidade de variância explicada continua correta).

- O coeficiente de determinação ajustado (R_a^2) penaliza a razão de somas de quadrados pela razão entre os graus de liberdade:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SQ_{res}}{SQ_{tot}}$$

- O coeficiente de determinação ajustado (R_a^2) penaliza a razão de somas de quadrados pela razão entre os graus de liberdade:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SQ_{res}}{SQ_{tot}}$$

- Dessa maneira, o ganho em explicação é ponderado pelo aumento de $\frac{(n-1)}{(n-p)}$, e o R_a^2 pode até diminuir com a adição de novas variáveis, se a contribuição não for importante.

  (Mas R_a^2 deixa de ter relação com % de variância explicada)

As inferências sobre o modelo (intervalos de confiança e testes de hipótese) seguem o mesmo modelo da regressão simples.

As equações para estimativas dos erros são mais complexas, mas o princípio não se altera.

Os modelos lineares de regressão múltipla apresentam algumas “complicações” a mais quando comparados com os modelos simples:

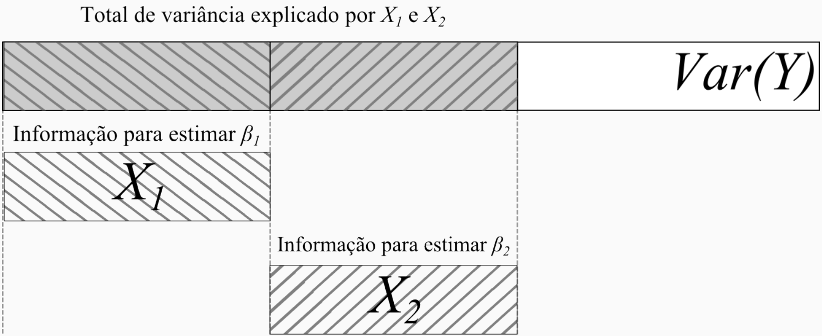
- A existência de correlação entre as variáveis pode atrapalhar a nossa partição de variância (multicolinearidade).
- Os coeficientes β normalmente não são diretamente comparáveis.
- Quando o número de variáveis independentes aumenta, a seleção final daquelas a serem inseridas no modelo torna-se mais difícil.

O modelo de regressão busca explicar parte da variância de Y através da co-variância entre Y e X (partição de variâncias)

Se as variáveis X são independentes, cada porção da variância de Y é explicada separadamente por cada X

Mas se as variáveis preditoras foem correlacionadas, há redundância de informação, reduzindo a quantidade de informação disponívelpara estimação dos coeficientes β

Caso 1: X_k perfeitamente independentes



Caso 1: X_k perfeitamente independentes

Nesse caso, a contribuição de X_1 e X_2 são exatamente as mesmas de dois modelos lineares simples:

```
x1 <- c(4,4,4,4,6,6,6,6)
x2 <- c(2,2,3,3,2,2,3,3)
y <- c(42,39,48,51,49,53,61,60)

cor(x1,x2)

## [1] 0
```

Caso 1: X_k perfeitamente independentes

```
m1 <- lm(y ~ x1);m1

##
## Call:
## lm(formula = y ~ x1)
##
## Coefficients:
## (Intercept)           x1
##      23.500        5.375

anova(m1)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value
## x1         1 231.12 231.125   7.347
## Residuals  6 188.75  31.458
##            Pr(>F)
## x1        0.03508 *
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

Caso 1: X_k perfeitamente independentes

```
m2 <- lm(y ~ x2); m2

##
## Call:
## lm(formula = y ~ x2)
##
## Coefficients:
## (Intercept)           x2
##       27.25         9.25

anova(m2)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value
## x2         1 171.12 171.125  4.1276
## Residuals  6 248.75  41.458
##            Pr(>F)
## x2        0.08846 .
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

Caso 1: X_k perfeitamente independentes

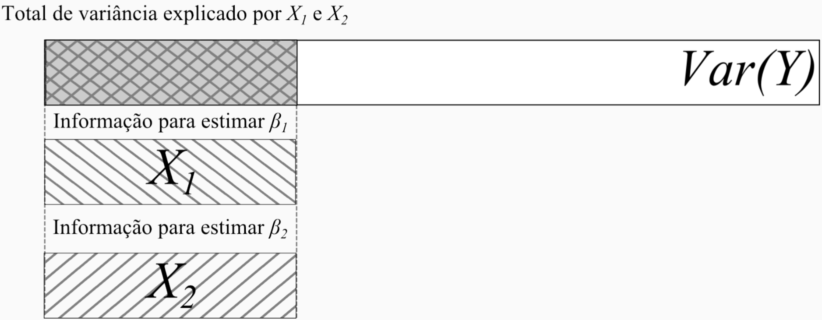
```
m3 <- lm(y ~ x1 + x2); m3

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##       0.375        5.375        9.250

anova(m3)

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value
## x1         1 231.125 231.125  65.567
## x2         1 171.125 171.125  48.546
## Residuals  5  17.625   3.525
##              Pr(>F)
## x1        0.0004657 ***
## x2        0.0009366 ***
## Residuals
## ---
## Signif. codes:
```

Caso 2: X_k perfeitamente correlacionados



Caso 2: X_k perfeitamente correlacionados

Nesse caso, não há variância restante para estimar β_2 após a estimação de β_1:

```
x1 <- c(4,4,4,4,6,6,6,6)
x2 <- x1
y <- c(42,39,48,51,49,53,61,60)

cor(x1,x2)

## [1] 1
```

Caso 2: X_k perfeitamente correlacionados

```
m1 <- lm(y ~ x1 + x2); m1

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##      23.500        5.375           NA

anova(m1)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value
## x1         1 231.12 231.125   7.347
## Residuals  6 188.75  31.458
##            Pr(>F)
## x1        0.03508 *
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

Caso 2: X_k perfeitamente correlacionados

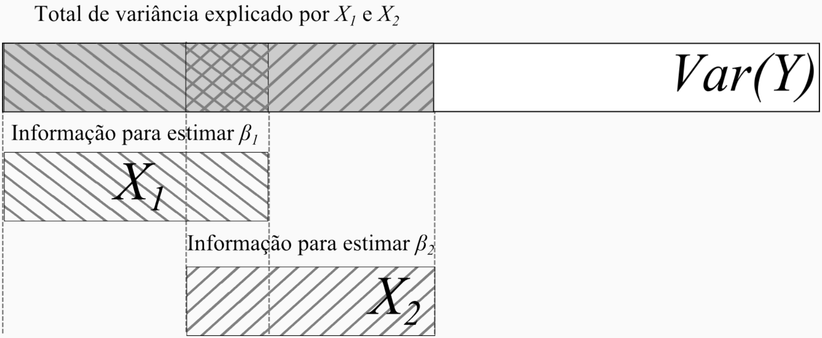
```
m2 <- lm(y ~ x2 + x1); m2

##
## Call:
## lm(formula = y ~ x2 + x1)
##
## Coefficients:
## (Intercept)           x2           x1
##      23.500        5.375           NA

anova(m2)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value
## x2         1 231.12 231.125   7.347
## Residuals  6 188.75  31.458
##            Pr(>F)
## x2        0.03508 *
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

Caso 3: X_k parcialmente correlacionados



Caso 3: X_k parcialmente correlacionados

Nesse caso, há ”menos”variância restante para estimar β_2 após a estimação de β_1:

```
x1 <- c(4,4,4,4,6,6,6,6)
set.seed(154)
x2 <- x1 + runif(8,0,1)
y <- c(42,39,48,51,49,53,61,60)

cor(x1,x2)

## [1] 0.9592065
```

Caso 3: X_k parcialmente correlacionados

```
m1 <- lm(y ~ x1 + x2); m1

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##      23.886        6.878       -1.418

anova(m1)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value
## x1         1 231.12 231.125  6.1739
## x2         1   1.57   1.570  0.0419
## Residuals  5 187.18  37.436
##            Pr(>F)
## x1        0.05552 .
## x2        0.84583
## Residuals
## ---
## Signif. codes:
```

Caso 3: X_k parcialmente correlacionados

```
m2 <- lm(y ~ x2 + x1); m2

##
## Call:
## lm(formula = y ~ x2 + x1)
##
## Coefficients:
## (Intercept)           x2           x1
##      23.886       -1.418        6.878

anova(m2)

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value
## x2         1 202.448 202.448  5.4078
## x1         1  30.246  30.246  0.8080
## Residuals  5 187.180  37.436
##            Pr(>F)
## x2        0.06759 .
## x1        0.40992
## Residuals
## ---
## Signif. codes:
```

Que parte do modelo de regressão esperamos que vá ser afetada pela multicolinearidade?

Que parte do modelo de regressão esperamos que vá ser afetada pela multicolinearidade?

- Os coeficientes β_1, ..., β_k

Que parte do modelo de regressão esperamos que vá ser afetada pela multicolinearidade?

- Os coeficientes β_1, ..., β_k

Qual será o principal efeito da multicolinearidade sobre a especificação do modelo?

Que parte do modelo de regressão esperamos que vá ser afetada pela multicolinearidade?

- Os coeficientes β_1, ..., β_k

Qual será o principal efeito da multicolinearidade sobre a especificação do modelo?

- As propriedades dos estimadores não se alteram (*BLUE*)

Que parte do modelo de regressão esperamos que vá ser afetada pela multicolinearidade?

- Os coeficientes β_1, ..., β_k

Qual será o principal efeito da multicolinearidade sobre a especificação do modelo?

- As propriedades dos estimadores não se alteram (*BLUE*)
- Devido à redução na quantidade de informação disponível, o erro na estimaçãode cada b_k aumenta.

Que parte do modelo de regressão esperamos que vá ser afetada pela multicolinearidade?

- Os coeficientes β_1, ..., β_k

Qual será o principal efeito da multicolinearidade sobre a especificação do modelo?

- As propriedades dos estimadores não se alteram (*BLUE*)
- Devido à redução na quantidade de informação disponível, o erro na estimaçãode cada b_k aumenta.
- Como a informação é redundante, múltiplas combinações de X_k e b_k podem dar o mesmo resultado final.