AULA 6: REGRESSÃO MÚLTIPLA, ANOVA E ANCOVA

Análise Estatística e Modelagem de Dados Ecológicos

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

31 de Março de 2015

Programa de Pós Graduação em Ecologia e Biodiversidade - UNESP

Regressão Múltipla

Diferenças entre Regressão Simples e Múltipla

ANOVA: Análise de Variância

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

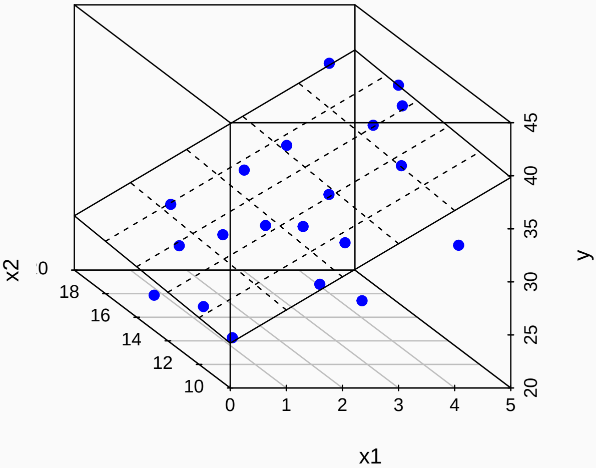
Para duas variáveis explicativas, temos:

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

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

Se $E(Y)$ depende de uma combinação de duas variáveis preditoras (X_1 e X_2), a reta se torna um plano

UM É POUCO, DOIS É BOM, TRÊS É MELHOR AINDA



- 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.

INTERPRETAÇÃO DOS TERMOS DO MODELO

- 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 $\beta_k = \frac{\delta E(Y)}{\delta X_{(k)}}$)

INTERPRETAÇÃO DOS TERMOS DO MODELO

- 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	$p-1$	$SQ_{Reg} = \mathbf{b'X'Y} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{Y'JY}$	$MR_{Reg} = \frac{SQ_{Reg}}{p-1}$
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 $F^{*} = \frac{MQ_{Reg}}{MQ_{Res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{Reg}}{SQ_{Tot}} = 1 - \frac{SQ_{Res}}{SQ_{Tot}}$

COEFICIENTE DE DETERMINAÇÃO MÚLTIPLO

- O teste geral para a regressão ainda é feito usando $F^{*} = \frac{MQ_{Reg}}{MQ_{Res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{Reg}}{SQ_{Tot}} = 1 - \frac{SQ_{Res}}{SQ_{Tot}}$
- Quando novas variáveis são incluídas no modelo, pode-se mostrar matematicamente que $SQ_{Res} nuncaaumenta$.

- O teste geral para a regressão ainda é feito usando $F^* = \frac{MQ_{Reg}}{MQ_{Res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{Reg}}{SQ_{Tot}} = 1 - \frac{SQ_{Res}}{SQ_{Tot}}$
- Quando novas variáveis são incluídas no modelo, pode-se mostrar matematicamente que $SQ_{Res} nuncaaumenta.$

  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 $F^* = \frac{MQ_{Reg}}{MQ_{Res}}$, e a quantidade de variância explicada é representada por $R^2 = \frac{SQ_{Reg}}{SQ_{Tot}} = 1 - \frac{SQ_{Res}}{SQ_{Tot}}$
- Quando novas variáveis são incluídas no modelo, pode-se mostrar matematicamente que $SQ_{Res} nunca aumenta.$

  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 procedimentos diagnósticos também são os mesmos, com a adição de scatterplots dos resíduos verus cada variável X_k.

DIFERENÇAS ENTRE REGRESSÃO SIMPLES E MÚLTIPLA

Os modelos lineares de regressão múltipla apresentam algumas “complicações” extras quando comparados aos modelos simples:

- A existência de correlação entre as variáveis pode atrapalhar a nossa partição de variância (multicolinearidade).

Os modelos lineares de regressão múltipla apresentam algumas “complicações” extras quando comparados aos 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.

Os modelos lineares de regressão múltipla apresentam algumas "complicações" extras quando comparados aos modelos simples:

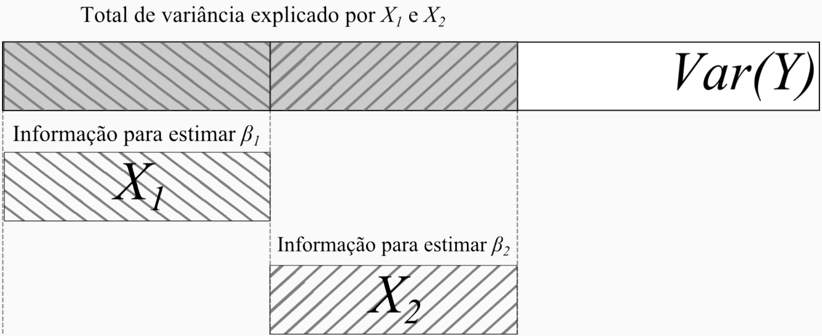
- A existência de correlação entre as variáveis pode atrapalhar a nossa partição de variância (multicolinearidade).
- Os coeficientes β normalmente não são diretamente comparáveis.
- Quando o número de variáveis independentes aumenta, a decisão sobre quais são mais ou menos importantes é mais difícil.

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 disponível para 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  Pr(>F)
## x1         1 231.12 231.125   7.347 0.03508 *
## Residuals  6 188.75  31.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

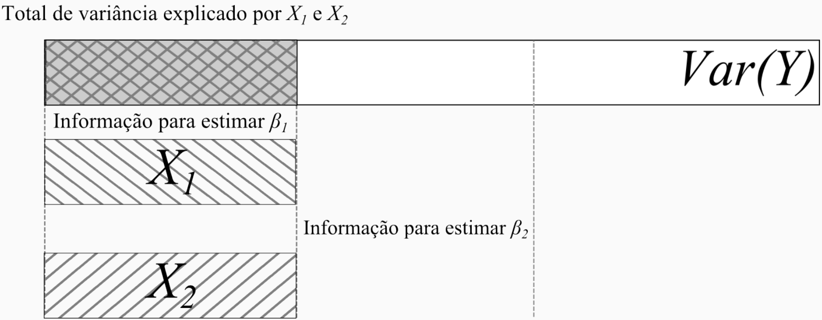
Caso 1: X_k perfeitamente independentes

```
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  Pr(>F)
## x2         1 171.12 171.125  4.1276 0.08846 .
## Residuals  6 248.75  41.458
## ---
## 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    Pr(>F)
## x1         1 231.125 231.125  65.567 0.0004657 ***
## x2         1 171.125 171.125  48.546 0.0009366 ***
## Residuals  5  17.625   3.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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
```

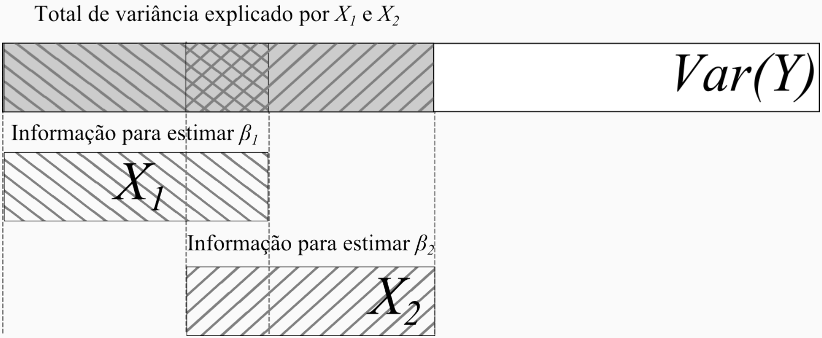
Caso 2: X_k perfeitamente correlacionados

```
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  Pr(>F)
## x1         1 231.12 231.125   7.347 0.03508 *
## Residuals  6 188.75  31.458
## ---
## 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  Pr(>F)
## x2         1 231.12 231.125   7.347 0.03508 *
## Residuals  6 188.75  31.458
## ---
## 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  Pr(>F)
## x1         1 231.12 231.125  6.1739 0.05552 .
## x2         1   1.57   1.570  0.0419 0.84583
## Residuals  5 187.18  37.436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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  Pr(>F)
## x2         1 202.448 202.448  5.4078 0.06759 .
## x1         1  30.246  30.246  0.8080 0.40992
## Residuals  5 187.180  37.436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 de 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 de cada b_k aumenta
- Como a informação é redudante, múltiplas combinações de X_k e b_k podem dar o mesmo resultado final

EXEMPLO

```
set.seed(1500)
x1 <- runif(50,0,20)
x2 <- x1 + runif(50,0,5)
y <- 24 + 1.2 *x1 + 2.1*x2 + rnorm(50,0,20)
m1 <- lm(y ~ x1)
summary(m1)

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.567 -12.840   1.822  12.161  46.037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7099     5.6739   5.941 3.08e-07 ***
## x1            3.0194     0.4763   6.339 7.61e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.93 on 48 degrees of freedom
## Multiple R-squared:  0.4556,^^IAdjusted R-squared:  0.4443
## F-statistic: 40.18 on 1 and 48 DF,  p-value: 7.607e-08
```

EXEMPLO

```r
m2 <- lm(y ~ x2)
summary(m2)

##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.280 -11.190  -3.074  10.861  42.214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.9196     6.3456   4.242    1e-04 ***
## x2            2.9624     0.4456   6.649 2.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.49 on 48 degrees of freedom
## Multiple R-squared:  0.4794,^^IAdjusted R-squared:  0.4686
## F-statistic: 44.21 on 1 and 48 DF,  p-value: 2.541e-08
```

EXEMPLO

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

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.823 -11.745  -3.351  10.786  41.832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.4629     7.0189   3.913 0.000293 ***
## x1            0.3541     1.8629   0.190 0.850079
## x2            2.6348     1.7818   1.479 0.145895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.69 on 47 degrees of freedom
## Multiple R-squared:  0.4798,^^IAdjusted R-squared:  0.4577
## F-statistic: 21.68 on 2 and 47 DF,  p-value: 2.134e-07
```

Podemos quantificar a existência de multicolinearidade através da medida de **tolerância**:

$T = 1 - R_k^2$

Podemos quantificar a existência de multicolinearidade através da medida de **tolerância**:

$T = 1 - R_k^2$

R_k^2 vem da regressão $\mathbf{X_k} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + \varepsilon$

Podemos quantificar a existência de multicolinearidade através da medida de **tolerância**:

$T = 1 - R_k^2$

R_k^2 vem da regressão $\mathbf{X_k} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + \varepsilon$

Normalmente, expressamos a tolerância na forma inversa, o que denominamos **Fator de Inflação da Variância** (*Variance Inflation Factor, VIF*)

$VIF = \frac{1}{T} = \frac{1}{1 - R_k^2}$

Podemos quantificar a existência de multicolinearidade através da medida de **tolerância**:

$T = 1 - R_k^2$

R_k^2 vem da regressão $\mathbf{X_k} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + \varepsilon$

Normalmente, expressamos a tolerância na forma inversa, o que denominamos **Fator de Inflação da Variância** (*Variance Inflation Factor, VIF*)

$$VIF = \frac{1}{T} = \frac{1}{1 - R_k^2}$$

O VIF nos dá a proporção do quanto o erro é "inflado"pela variável. Se X_k tem um VIF de 1.8, isso significa que o erro do coeficiente b_k é 80% maior do que o esperado se não houvesse colinearidade.

```r
library(car)
vif(m3)

##       x1       x2
## 15.67194 15.67194
```

A partir de que valor devemos nos preocupar com o VIF?

A partir de que valor devemos nos preocupar com o VIF?

Não existe uma regra fixa, mas em geral:

A partir de que valor devemos nos preocupar com o VIF?

Não existe uma regra fixa, mas em geral:

VIF > 4 pede que a correlação entre os preditores seja melhor investigada

A partir de que valor devemos nos preocupar com o VIF?

Não existe uma regra fixa, mas em geral:

VIF > 4 pede que a correlação entre os preditores seja melhor investigada

VIF > 10 representa multicolinearidade severa, precisa ser corrigida de qualquer maneira

Podemos resolver o problema da multicolinearidade de diversas maneiras:

Podemos resolver o problema da multicolinearidade de diversas maneiras:

1) Através de uma combinação entre as variáveis (ex.: $X_1 + X_2$)

Podemos resolver o problema da multicolinearidade de diversas maneiras:

1) Através de uma combinação entre as variáveis (ex.: $X_1 + X_2$)

2) Usando os resíduos da regressão entre X_1 e X_2

Podemos resolver o problema da multicolinearidade de diversas maneiras:

1) Através de uma combinação entre as variáveis (ex.: $X_1 + X_2$)

2) Usando os resíduos da regressão entre X_1 e X_2

3) Ortogonalização (ex.: análise de componentes principais)

1) Através de uma combinação entre as variáveis (ex.: $X_1 + X_2$)

```
x.novo <- x1 + x2
m4 <- lm(y ~ x.novo)
summary(m4)

##
## Call:
## lm(formula = y ~ x.novo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.328 -13.047  -1.519  11.687  43.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.7061     6.0136   4.940 9.89e-06 ***
## x.novo        1.5202     0.2305   6.596 3.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.56 on 48 degrees of freedom
## Multiple R-squared:  0.4754,^^IAdjusted R-squared:  0.4645
## F-statistic:  43.5 on 1 and 48 DF,  p-value: 3.066e-08
```

2) Usando os resíduos da regressão entre X_1 e X_2

```
mx <- lm(x2 ~ x1)
rx <- residuals(mx)
m5 <- lm(y ~ x1 + rx)
summary(m5)

##
## Call:
## lm(formula = y ~ x1 + rx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.823 -11.745  -3.351  10.786  41.832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7099     5.6050   6.014 2.56e-07 ***
## x1            3.0194     0.4706   6.416 6.29e-08 ***
## rx            2.6348     1.7818   1.479    0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.69 on 47 degrees of freedom
## Multiple R-squared:  0.4798,^^IAdjusted R-squared:  0.4577
## F-statistic: 21.68 on 2 and 47 DF,  p-value: 2.134e-07
```

3) Ortogonalização (ex.: análise de componentes principais)

```
pca <- princomp(~ x2 + x1)
m6 <- lm(y ~ pca$scores[,1] + pca$scores[,2])
summary(m6)

##
## Call:
## lm(formula = y ~ pca$scores[, 1] + pca$scores[, 2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.823 -11.745  -3.351  10.786  41.832
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      64.9237     2.7841  23.319  < 2e-16 ***
## pca$scores[, 1]  -2.1499     0.3279  -6.556 3.86e-08 ***
## pca$scores[, 2]   1.5637     2.5569   0.612    0.544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.69 on 47 degrees of freedom
## Multiple R-squared:  0.4798,^^IAdjusted R-squared:  0.4577
## F-statistic: 21.68 on 2 and 47 DF,  p-value: 2.134e-07
```

ANOVA: ANÁLISE DE VARIÂNCIA

A ANOVA pode ser vista como uma regressão usando uma variável X categórica

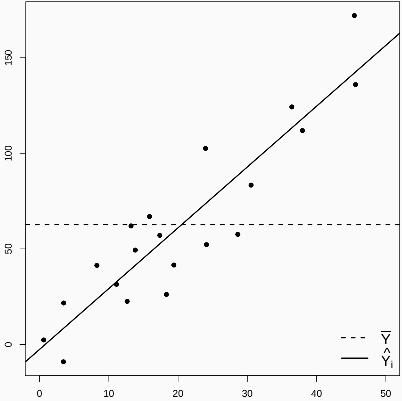
A ANOVA pode ser vista como uma regressão usando uma variável X categórica

Como X não é contínua, o foco é determinar se existem diferenças em $E(Y)$ para cada nível de X, sem a pressuposição de relação linear entre X e Y

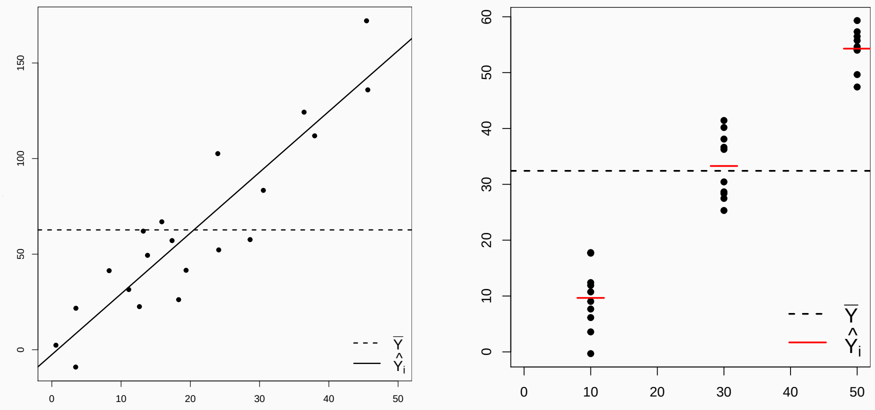
A ANOVA pode ser vista como uma regressão usando uma variável X categórica

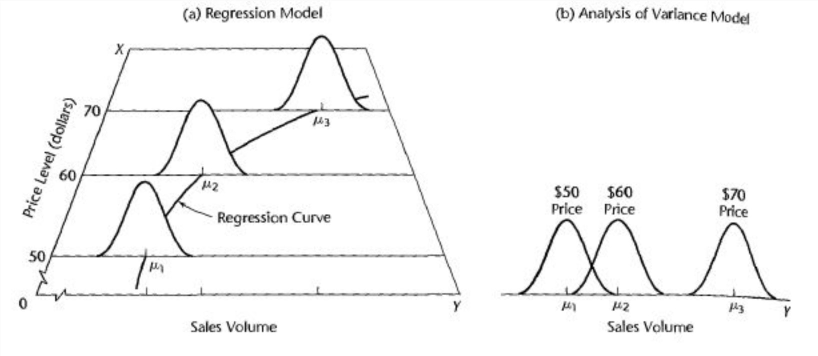
Como X não é contínua, o foco é determinar se existem diferenças em $E(Y)$ para cada nível de X, sem a pressuposição de relação linear entre X e Y

ANOVA e Regressão Linear são casos específicos dos chamados Modelos Lineares Gerais



150
100
50
0
0
10
20
30
40
50
$\overline{Y}$
$\hat{Y}_i$


0
10
20
30
40
50
150
100
50
0
60
$\bar{Y}$
$\hat{Y}_i$


(a) Regression Model
X
70
60
50
0
Price Level (dollars)
μ_3
μ_2
μ_1
Regression Curve
Y
Sales Volume
(b) Analysis of Variance Model
$50 Price
$60 Price
$70 Price
μ_1
μ_2
μ_3
Y
Sales Volume

A notação matemática utilizada para a ANOVA é um pouco diferente da regressão:

A notação matemática utilizada para a ANOVA é um pouco diferente da regressão:

- Y ainda é a variável resposta

A notação matemática utilizada para a ANOVA é um pouco diferente da regressão:

- Y ainda é a variável resposta
- Em vez de ter um X contínuo, temos um X categórico (fator), com diferentes níveis ou *tratamentos* que são identificados por $i = (1, 2, \ldots, r)$

A notação matemática utilizada para a ANOVA é um pouco diferente da regressão:

- Y ainda é a variável resposta
- Em vez de ter um X contínuo, temos um X categórico (fator), com diferentes níveis ou *tratamentos* que são identificados por $i = (1, 2, \ldots, r)$
- Cada X_i tem um certo número de observações, n_i, e o número total de obseravções é n
- Como agora usamos i para identificar os níveis de X, usamos j para descrever uma observação específica em cada nível: $j = (1, 2, \ldots, n_i)$

O nosso modelo de regressão era:

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

O nosso modelo de regressão era:

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

O modelo ANOVA é parecido, mas tem menos parâmetros:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

O nosso modelo de regressão era:

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

O modelo ANOVA é parecido, mas tem menos parâmetros:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Ou seja, cada valor de Y (Y_{ij}) é determinado pela média do grupo i, mais um erro aleatório (distância entre Y_{ij} e μ_i). Este é o chamado “modelo celular”, ou modelo de médias celulares.

Também podemos especificar o modelo da seguinte maneira:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Também podemos especificar o modelo da seguinte maneira:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

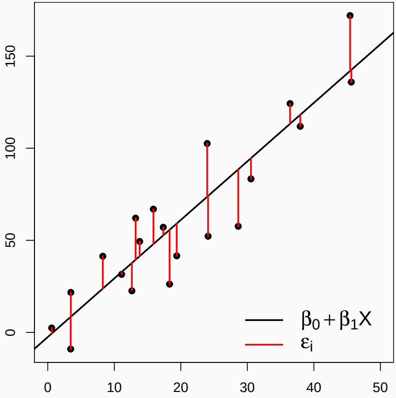
Aqui, cada valor de Y (Y_{ij}) é determinado pela média global de Y (μ), somada ou subtraída de um coeficiente α que varia de acordo com o nível i, mais um erro aleatório. ($\mu + \alpha_i = \mu_i$).

Também podemos especificar o modelo da seguinte maneira:

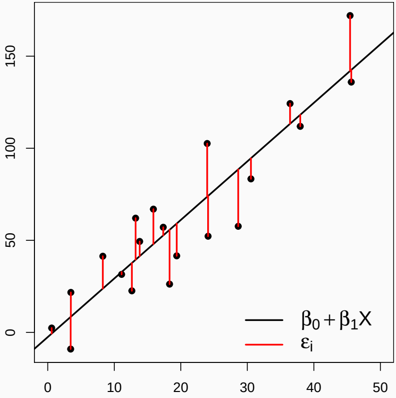
$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

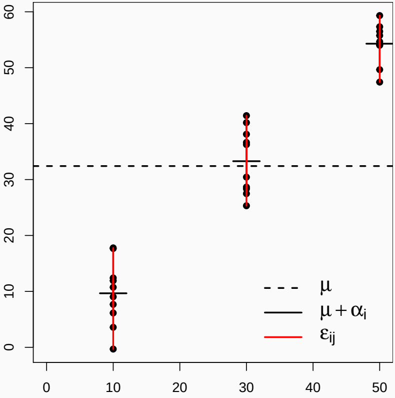
Aqui, cada valor de Y (Y_{ij}) é determinado pela média global de Y (μ), somada ou subtraída de um coeficiente α que varia de acordo com o nível i, mais um erro aleatório. ($\mu + \alpha_i = \mu_i$).

Este modelo é chamado de “modelo de efeitos”, já que individualiza os efeitos de cada tratamento.



150
100
50
0
0
10
20
30
40
50
$\beta_0 + \beta_1 X$
ε_i


0
50
100
150
0
10
20
30
40
50
$\beta_0 + \beta_1 X$
ε_i


0
10
20
30
40
50
60
0
10
20
30
40
50
μ
$\mu + \alpha_i$
ε_{ij}

O modelo ANOVA segue pressuposições similares ao modelo de regressão:

O modelo ANOVA segue pressuposições similares ao modelo de regressão:

- A distribuição dos erros segue $\varepsilon \sim N(0, \sigma^2)$

O modelo ANOVA segue pressuposições similares ao modelo de regressão:

- A distribuição dos erros segue $\varepsilon \sim N(0, \sigma^2)$
- σ^2 é constante para todos os i níveis de X $(\sigma_1^2 = \sigma_2^2 \ldots = \sigma_i^2 = \sigma^2)$

O modelo ANOVA segue pressuposições similares ao modelo de regressão:

- A distribuição dos erros segue $\varepsilon \sim N(0, \sigma^2)$
- σ^2 é constante para todos os i níveis de X $(\sigma_1^2 = \sigma_2^2 \ldots = \sigma_i^2 = \sigma^2)$
- Os erros ε_i são independentes entre si

O modelo ANOVA segue pressuposições similares ao modelo de regressão:

- A distribuição dos erros segue $\varepsilon \sim N(0, \sigma^2)$
- σ^2 é constante para todos os i níveis de X $(\sigma_1^2 = \sigma_2^2 \ldots = \sigma_i^2 = \sigma^2)$
- Os erros ε_i são independentes entre si

Se essas pressuposições são verdadeiras, então:

$E(Y_{ij}) = \mu + \alpha_i$, $Var(Y_{ij}) = \sigma^2$

Podemos desdobrar nosso modelo ANOVA para que se assemelhe a uma regressão. No caso de uma ANOVA com três níveis de X:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Podemos desdobrar nosso modelo ANOVA para que se assemelhe a uma regressão. No caso de uma ANOVA com três níveis de X:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$Y_{ij} = \mu + \alpha_1 X_{i=1} + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij}$ ou

Podemos desdobrar nosso modelo ANOVA para que se assemelhe a uma regressão. No caso de uma ANOVA com três níveis de X:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$Y_{ij} = \mu + \alpha_1 X_{i=1} + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij}$ ou

$$Y_{ij} = \mu_1 + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij}$$

Podemos desdobrar nosso modelo ANOVA para que se assemelhe a uma regressão. No caso de uma ANOVA com três níveis de X:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$Y_{ij} = \mu + \alpha_1 X_{i=1} + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij} \text{ ou}$$

$$Y_{ij} = \mu_1 + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij}$$

Nesse caso, X_i é chamado de variável **indicadora** (ou *dummy*) com valor $X_i = 1$ para $Y = Y_{ij}$, e $X_i = 0$ quando $Y \neq Y_{ij}$

EXEMPLO: VARIÁVEL INDICADORA

Y	X
62.00	A
60.00	A
63.00	A
63.00	B
67.00	B
71.00	B
72.00	C
70.00	C
75.00	C

$$A = X_{i=1}$$
$$B = X_{i=2}$$
$$C = X_{i=3}$$

Y	X	X_1	X_2	X_3
62.00	A	1	0	0
60.00	A	1	0	0
63.00	A	1	0	0
63.00	B	0	1	0
67.00	B	0	1	0
71.00	B	0	1	0
72.00	C	0	0	1
70.00	C	0	0	1
75.00	C	0	0	1

$$Y_{ij} = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Y	X	X_1	X_2	X_3
62.00	A	1	0	0
60.00	A	1	0	0
63.00	A	1	0	0
63.00	B	0	1	0
67.00	B	0	1	0
71.00	B	0	1	0
72.00	C	0	0	1
70.00	C	0	0	1
75.00	C	0	0	1

Y	X	X_1	X_2	X_3
62.00	A	1	0	0
60.00	A	1	0	0
63.00	A	1	0	0
63.00	B	0	1	0
67.00	B	0	1	0
71.00	B	0	1	0
72.00	C	0	0	1
70.00	C	0	0	1
75.00	C	0	0	1

$$Y_{ij} = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Para $Y = Y_{Aj}$:

$$Y_{Aj} = \mu + \alpha_1 \times 1 + \alpha_2 \times 0 + \alpha_3 \times 0 + \varepsilon_{Aj}$$
$$Y_{Aj} = \mu + \alpha_1 + \varepsilon_{Aj}$$

Y	X	X_1	X_2	X_3
62.00	A	1	0	0
60.00	A	1	0	0
63.00	A	1	0	0
63.00	B	0	1	0
67.00	B	0	1	0
71.00	B	0	1	0
72.00	C	0	0	1
70.00	C	0	0	1
75.00	C	0	0	1

$$Y_{ij} = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Para $Y = Y_{Aj}$:

$$Y_{Aj} = \mu + \alpha_1 \times 1 + \alpha_2 \times 0 + \alpha_3 \times 0 + \varepsilon_{Aj}$$
$$Y_{Aj} = \mu + \alpha_1 + \varepsilon_{Aj}$$

Para $Y = Y_{Bj}$:

$$Y_{Bj} = \mu + \alpha_1 \times 0 + \alpha_2 \times 1 + \alpha_3 \times 0 + \varepsilon_{Bj}$$
$$Y_{Bj} = \mu + \alpha_2 + \varepsilon_{Bj}$$

Y	X	X_1	X_2	X_3
62.00	A	1	0	0
60.00	A	1	0	0
63.00	A	1	0	0
63.00	B	0	1	0
67.00	B	0	1	0
71.00	B	0	1	0
72.00	C	0	0	1
70.00	C	0	0	1
75.00	C	0	0	1

$$Y_{ij} = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Para $Y = Y_{Aj}$:

$$Y_{Aj} = \mu + \alpha_1 \times 1 + \alpha_2 \times 0 + \alpha_3 \times 0 + \varepsilon_{Aj}$$
$$Y_{Aj} = \mu + \alpha_1 + \varepsilon_{Aj}$$

Para $Y = Y_{Bj}$:

$$Y_{Bj} = \mu + \alpha_1 \times 0 + \alpha_2 \times 1 + \alpha_3 \times 0 + \varepsilon_{Bj}$$
$$Y_{Bj} = \mu + \alpha_2 + \varepsilon_{Bj}$$

Para $Y = Y_{Cj}$:

$$Y_{Cj} = \mu + \alpha_1 \times 0 + \alpha_2 \times 0 + \alpha_3 \times 1 + \varepsilon_{Cj}$$
$$Y_{Cj} = \mu + \alpha_3 + \varepsilon_{Cj}$$

EXEMPLO: VARIÁVEL INDICADORA

Y	X	X_2	X_3
62.00	A	0	0
60.00	A	0	0
63.00	A	0	0
63.00	B	1	0
67.00	B	1	0
71.00	B	1	0
72.00	C	0	1
70.00	C	0	1
75.00	C	0	1

$$Y_{ij} = \mu_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Y	X	X_2	X_3
62.00	A	0	0
60.00	A	0	0
63.00	A	0	0
63.00	B	1	0
67.00	B	1	0
71.00	B	1	0
72.00	C	0	1
70.00	C	0	1
75.00	C	0	1

Y	X	X_2	X_3
62.00	A	0	0
60.00	A	0	0
63.00	A	0	0
63.00	B	1	0
67.00	B	1	0
71.00	B	1	0
72.00	C	0	1
70.00	C	0	1
75.00	C	0	1

$$Y_{ij} = \mu_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Para $Y = Y_{Aj}$:

$$Y_{Aj} = \mu_1 + \alpha_2 \times 0 + \alpha_3 \times 0 + \varepsilon_{Aj}$$
$$Y_{Aj} = \mu_1 + \varepsilon_{Aj}$$

Y	X	X_2	X_3
62.00	A	0	0
60.00	A	0	0
63.00	A	0	0
63.00	B	1	0
67.00	B	1	0
71.00	B	1	0
72.00	C	0	1
70.00	C	0	1
75.00	C	0	1

$$Y_{ij} = \mu_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Para $Y = Y_{Aj}$:

$$Y_{Aj} = \mu_1 + \alpha_2 \times 0 + \alpha_3 \times 0 + \varepsilon_{Aj}$$
$$Y_{Aj} = \mu_1 + \varepsilon_{Aj}$$

Para $Y = Y_{Bj}$:

$$Y_{Bj} = \mu_1 + \alpha_2 \times 1 + \alpha_3 \times 0 + \varepsilon_{Bj}$$
$$Y_{Bj} = \mu_1 + \alpha_2 + \varepsilon_{Bj}$$

Y	X	X_2	X_3
62.00	A	0	0
60.00	A	0	0
63.00	A	0	0
63.00	B	1	0
67.00	B	1	0
71.00	B	1	0
72.00	C	0	1
70.00	C	0	1
75.00	C	0	1

$$Y_{ij} = \mu_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon_{ij}$$

Para $Y = Y_{Aj}$:

$$Y_{Aj} = \mu_1 + \alpha_2 \times 0 + \alpha_3 \times 0 + \varepsilon_{Aj}$$
$$Y_{Aj} = \mu_1 + \varepsilon_{Aj}$$

Para $Y = Y_{Bj}$:

$$Y_{Bj} = \mu_1 + \alpha_2 \times 1 + \alpha_3 \times 0 + \varepsilon_{Bj}$$
$$Y_{Bj} = \mu_1 + \alpha_2 + \varepsilon_{Bj}$$

Para $Y = Y_{Cj}$:

$$Y_{Cj} = \mu_1 + \alpha_2 \times 0 + \alpha_3 \times 1 + \varepsilon_{Cj}$$
$$Y_{Cj} = \mu_1 + \alpha_3 + \varepsilon_{Cj}$$

No nosso modelo de regressão, podíamos particionar a variância como $SQ_{Tot} = SQ_{Reg} + SQ_{Res}$:

$$SQ_{Tot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

No nosso modelo de regressão, podíamos particionar a variância como $SQ_{Tot} = SQ_{Reg} + SQ_{Res}$:

$$SQ_{Tot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

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

No nosso modelo de regressão, podíamos particionar a variância como $SQ_{Tot} = SQ_{Reg} + SQ_{Res}$:

$$SQ_{Tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

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

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})$$

Podemos fazer algo semelhante com o nosso modelo ANOVA:

Podemos fazer algo semelhante com o nosso modelo ANOVA:

$$SQ_{Tot} =$$

Podemos fazer algo semelhante com o nosso modelo ANOVA:

$$SQ_{Tot} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$$

Podemos fazer algo semelhante com o nosso modelo ANOVA:

$$SQ_{Tot} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$$

$$SQ_{Res} =$$

Podemos fazer algo semelhante com o nosso modelo ANOVA:

$$SQ_{Tot} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$$

$$SQ_{Res} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

Podemos fazer algo semelhante com o nosso modelo ANOVA:

$$SQ_{Tot} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$$

$$SQ_{Res} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

$$SQ_{Reg} =$$

Podemos fazer algo semelhante com o nosso modelo ANOVA:

$$SQ_{Tot} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$$

$$SQ_{Res} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

$$SQ_{Reg} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_i - \bar{Y})^2$$

Podemos fazer algo semelhante com o nosso modelo ANOVA:

$$SQ_{Tot} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$$

$$SQ_{Res} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

$$SQ_{Reg} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_i - \bar{Y})^2$$

Essa partição pode ser interpretada como:

Essa partição pode ser interpretada como:

SQ_{Tot}: Variância total de Y

Essa partição pode ser interpretada como:

SQ_{Tot}: Variância total de Y

SQ_{Res}: Variância **intra**-grupos

Essa partição pode ser interpretada como:

SQ_{Tot}: Variância total de Y

SQ_{Res}: Variância **intra**-grupos

SQ_{Reg}: Variância **entre grupos**

Através da comparação entre SQ_{Res} e SQ_{Reg}, podemos avaliar o quanto as diferenças **entre grupos** são importantes, em relação à variação **intra-**grupos:

Essa partição pode ser interpretada como:

SQ_{Tot}: Variância total de Y

SQ_{Res}: Variância **intra**-grupos

SQ_{Reg}: Variância **entre grupos**

Através da comparação entre SQ_{Res} e SQ_{Reg}, podemos avaliar o quanto as diferenças **entre grupos** são importantes, em relação à variação **intra-**grupos:

$H_0 : \mu + \varepsilon_{ij}$

$H_a : Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

A construção da tabela ANOVA é similar ao que vimos para regressão, mas a ênfase agora é na diferença entre variação intra- e inter- grupos:

Fonte	GL	Soma Quadrados	Média Quadrados	E(Med. Quad.)	F	P
Tratamento	$r-1$	$SQ_{Reg} = \sum_{i=1}^{r} n_i(\bar{Y}_i - \bar{Y})^2$	$MQ_{Reg} = \frac{SQ_{Reg}}{r-1}$	$r\sum \frac{\alpha_i^2}{r-1} + \sigma^2$	$\frac{MQ_{Reg}}{MQ_{Res}}$	P($F_{(r-1,n-1)}$)
Resíduos	$(n-r)$	$SQ_{Reg} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$MQ_{Res} = \frac{SQ_{Res}}{n-r}$	σ^2		
Total	$n-1$	$SQ_{Tot} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	$MQ_{Tot} = \frac{SQ_{Tot}}{n-1}$	σ_Y^2		

Para o modelo ANOVA, $E(MQ_{Reg})$ é:

$$r\sum \frac{\alpha_i^2}{r-1} + \sigma^2$$

Para o modelo ANOVA, $E(MQ_{Reg})$ é:

$$r\sum \frac{\alpha_i^2}{r-1} + \sigma^2$$

E $E(MQ_{Res}) = \sigma^2$

Para o modelo ANOVA, $E(MQ_{Reg})$ é:

$r\sum \dfrac{\alpha_i^2}{r-1} + \sigma^2$

E $E(MQ_{Res}) = \sigma^2$

Se não existe diferença entre os tratamentos, então $\alpha_1 = \alpha_2 = \ldots = \alpha_i = 0$ e temos:

$$F = \frac{MQ_{Reg}}{MQ_{Res}} = \frac{r\sum \frac{\alpha_i^2}{r-1} + \sigma^2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

Para o modelo ANOVA, $E(MQ_{Reg})$ é:

$r\sum\frac{\alpha_i^2}{r-1}+\sigma^2$

E $E(MQ_{Res}) = \sigma^2$

Se não existe diferença entre os tratamentos, então $\alpha_1 = \alpha_2 = \ldots = \alpha_i = 0$ e temos:

$$F = \frac{MQ_{Reg}}{MQ_{Res}} = \frac{r\sum\frac{\alpha_i^2}{r-1}+\sigma^2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

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

O resultado de uma ANOVA nos diz apenas o quanto há de evidencia sobre a hipótese H_0 ser falsa, ou seja, os diferentes níveis (r) do tratamento tem de fato um efeito sobre $E(Y)$

O resultado de uma ANOVA nos diz apenas o quanto há de evidencia sobre a hipótese H_0 ser falsa, ou seja, os diferentes níveis (r) do tratamento tem de fato um efeito sobre $E(Y)$

Se $r = 2$ ...a ANOVA é equivalente a um teste t

Mas se $r \geq 3$, logo surge uma pergunta: será que **todos** os níveis são diferentes, ou apenas alguns deles?

Podemos respoder à essa pergunta de duas maneiras: usando testes post-hoc, ou usando contrastes.

Um dos testes post-hoc mais comuns é o “Teste das Diferenças Honestamente Significantes” de Tukey (*Tukey’s HSD test*)

Um dos testes post-hoc mais comuns é o "Teste das Diferenças Honestamente Significantes" de Tukey (*Tukey's HSD test*)

O teste calcula uma diferença mínima entre os tratamentos que pode ser considerada significativa, usando:

$$HSD = q\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) MSRes}$$

Atenção: neste caso, i e j se referem a dois valores diferentes de r

Um dos testes post-hoc mais comuns é o “Teste das Diferenças Honestamente Significantes” de Tukey (*Tukey’s HSD test*)

O teste calcula uma diferença mínima entre os tratamentos que pode ser considerada significativa, usando:

$$HSD = q\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) MSRes}$$

Atenção: neste caso, i e j se referem a dois valores diferentes de r

q vem de uma tabela específica para o teste

O teste HSD de tukey pode ser visto como uma série de testes t, com nível de significância ajustado para múltiplos testes.

O teste HSD de tukey pode ser visto como uma série de testes t, com nível de significância ajustado para múltiplos testes.

Após calcular o valor de HSD, comparam-se as médias de cada tratamento.

O teste HSD de tukey pode ser visto como uma série de testes t, com nível de significância ajustado para múltiplos testes.

Após calcular o valor de HSD, comparam-se as médias de cada tratamento.

Os valores p e intervalos de confiança para cada diferença pareada podem ser interpretados como grau de incerteza.

Através do uso de contrastes, podemos não só comparar diferenças entre níveis, mas diferentes combinações destes valores

Através do uso de contrastes, podemos não só comparar diferenças entre níveis, mas diferentes combinações destes valores

Os contrastes são valores específicos, escolhidos de maneira a serem ortogonais (não-correlacionados), e que nos permitem controlar a partição das somas dos quadrados.

Através do uso de contrastes, podemos não só comparar diferenças entre níveis, mas diferentes combinações destes valores

Os contrastes são valores específicos, escolhidos de maneira a serem ortogonais (não-correlacionados), e que nos permitem controlar a partição das somas dos quadrados.

Criando contrastes:

- Associe um valor inteiro (positivo, negativo ou zero) para cada tratamento

Criando contrastes:

- Associe um valor inteiro (positivo, negativo ou zero) para cada tratamento
- Tratamentos que devem ser agrupados recebem o mesmo número

Criando contrastes:

- Associe um valor inteiro (positivo, negativo ou zero) para cada tratamento
- Tratamentos que devem ser agrupados recebem o mesmo número
- Tratamentos excluídos da comparação recebem contraste zero

Criando contrastes:

- Associe um valor inteiro (positivo, negativo ou zero) para cada tratamento
- Tratamentos que devem ser agrupados recebem o mesmo número
- Tratamentos excluídos da comparação recebem contraste zero
- A soma final de todos os contrastes deve ser zero

Podemos criar múltipos contrastes, e testar várias hipóteses simultâneas, se obedecermos mais duas regras:

- Se existem r tratamentos, então podemos ter no máximo $r-1$ contrastes

Podemos criar múltipos contrastes, e testar várias hipóteses simultâneas, se obedecermos mais duas regras:

- Se existem r tratamentos, então podemos ter no máximo $r-1$ contrastes
- A soma dos produtos cruzados de todos os contrastes precisa ser zero (ortogonalidade)

Podemos criar múltipos contrastes, e testar várias hipóteses simultâneas, se obedecermos mais duas regras:

- Se existem r tratamentos, então podemos ter no máximo $r - 1$ contrastes
- A soma dos produtos cruzados de todos os contrastes precisa ser zero (ortogonalidade)

Criar contrastes é mais ou menos como resolver um sudoku...

E onde entram esses contrastes no modelo?

Lembram do modelo expandido da regressão, com variáveis indicadoras?

$Y_{ij} = \mu + \alpha_1 X_{i=1} + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij}$ ou

E onde entram esses contrastes no modelo?

Lembram do modelo expandido da regressão, com variáveis indicadoras?

$Y_{ij} = \mu + \alpha_1 X_{i=1} + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij}$ ou

Os contrastes também são variáveis indicadoras!

E onde entram esses contrastes no modelo?

Lembram do modelo expandido da regressão, com variáveis indicadoras?

$Y_{ij} = \mu + \alpha_1 X_{i=1} + \alpha_2 X_{i=2} + \alpha_3 X_{i=3} + \varepsilon_{ij}$ ou

Os contrastes também são variáveis indicadoras!

Comparação entre cada tratamento (original):

Y	X	X_1	X_2	X_3
62.00	A	1	0	0
60.00	A	1	0	0
63.00	A	1	0	0
63.00	B	0	1	0
67.00	B	0	1	0
71.00	B	0	1	0
72.00	C	0	0	1
70.00	C	0	0	1
75.00	C	0	0	1

Comparação entre B e C:

Y	X	C_1
62.00	A	0
60.00	A	0
63.00	A	0
63.00	B	1
67.00	B	1
71.00	B	1
72.00	C	-1
70.00	C	-1
75.00	C	-1

Comparação de A versus B e C combinados:

Y	X	C_1
62.00	A	2
60.00	A	2
63.00	A	2
63.00	B	-1
67.00	B	-1
71.00	B	-1
72.00	C	-1
70.00	C	-1
75.00	C	-1

Vamos utilizar o exemplo em Gotelli and Ellison (2004) (publicado em Ellison et al. Ecology 77: 2431-2444, 1996), sobre o efeito da presença de esponjas sobre crescimento radicular de *Rhizophora mangle*.

Vamos utilizar o exemplo em Gotelli and Ellison (2004) (publicado em Ellison et al. Ecology 77: 2431-2444, 1996), sobre o efeito da presença de esponjas sobre crescimento radicular de *Rhizophora mangle*.

O experimento consistiu em quatro tratamentos:

Vamos utilizar o exemplo em Gotelli and Ellison (2004) (publicado em Ellison et al. Ecology 77: 2431-2444, 1996), sobre o efeito da presença de esponjas sobre crescimento radicular de *Rhizophora mangle*.

O experimento consistiu em quatro tratamentos:

- Control: sem manipulação

Vamos utilizar o exemplo em Gotelli and Ellison (2004) (publicado em Ellison et al. Ecology 77: 2431-2444, 1996), sobre o efeito da presença de esponjas sobre crescimento radicular de *Rhizophora mangle*.

O experimento consistiu em quatro tratamentos:

- Control: sem manipulação
- Foam: espuma sintética

Vamos utilizar o exemplo em Gotelli and Ellison (2004) (publicado em Ellison et al. Ecology 77: 2431-2444, 1996), sobre o efeito da presença de esponjas sobre crescimento radicular de *Rhizophora mangle*.

O experimento consistiu em quatro tratamentos:

- Control: sem manipulação
- Foam: espuma sintética
- Haliclona: enxerto da esponja *Haliclona implexiformis*

Vamos utilizar o exemplo em Gotelli and Ellison (2004) (publicado em Ellison et al. Ecology 77: 2431-2444, 1996), sobre o efeito da presença de esponjas sobre crescimento radicular de *Rhizophora mangle*.

O experimento consistiu em quatro tratamentos:

- Control: sem manipulação
- Foam: espuma sintética
- Haliclona: enxerto da esponja *Haliclona implexiformis*
- Tedania: enxerto da esponja *Tedania ignis*

Vamos utilizar o exemplo em Gotelli and Ellison (2004) (publicado em Ellison et al. Ecology 77: 2431-2444, 1996), sobre o efeito da presença de esponjas sobre crescimento radicular de *Rhizophora mangle*.

O experimento consistiu em quatro tratamentos:

- Control: sem manipulação
- Foam: espuma sintética
- Haliclona: enxerto da esponja *Haliclona implexiformis*
- Tedania: enxerto da esponja *Tedania ignis*

E a variável dependente foi: Y = taxa mensal de crescimento radicular, em mm.dia

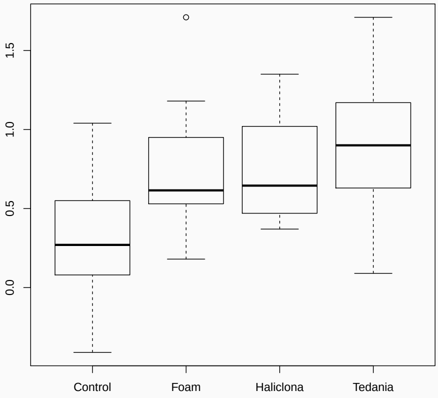
ANOVA: EXEMPLO

```
str(mangue)

## 'data.frame':^^I56 obs. of  2 variables:
##  $ trat : Factor w/ 4 levels "Control","Foam",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cresc: num  -0.05 0.19 0.84 0.11 0.08 -0.18 0.38 1.04 0.55 0.24 ...

mangue[1:5,]

##      trat cresc
## 1 Control -0.05
## 2 Control  0.19
## 3 Control  0.84
## 4 Control  0.11
## 5 Control  0.08

mangue[15:20,]

##    trat cresc
## 15 Foam  0.73
## 16 Foam  0.56
## 17 Foam  0.98
## 18 Foam  0.53
## 19 Foam  0.32
## 20 Foam  0.95
```

```
summary(mangue)

##        trat         cresc
##  Control  :14   Min.   :-0.410
##  Foam     :14   1st Qu.: 0.395
##  Haliclona:14   Median : 0.620
##  Tedania  :14   Mean   : 0.680
##                 3rd Qu.: 1.010
##                 Max.   : 1.710

tapply(mangue$cresc,mangue$trat,mean)

##   Control      Foam Haliclona   Tedania
## 0.3292857 0.7121429 0.7650000 0.9135714

tapply(mangue$cresc,mangue$trat,sd)

##   Control      Foam Haliclona   Tedania
## 0.4316598 0.3930782 0.3478671 0.4415638
```

ANOVA: EXEMPLO

```
boxplot(cresc ~ trat,mangue)
```



ANOVA: EXEMPLO

```
m1 <- lm(cresc ~ trat,mangue)
summary(m1)

##
## Call:
## lm(formula = cresc ~ trat, data = mangue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82357 -0.28643 -0.05786  0.24750  0.99786
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.3293     0.1083   3.040 0.003694 **
## tratFoam        0.3829     0.1532   2.500 0.015623 *
## tratHaliclona   0.4357     0.1532   2.845 0.006341 **
## tratTedania     0.5843     0.1532   3.815 0.000363 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4052 on 52 degrees of freedom
## Multiple R-squared:  0.2335,^^IAdjusted R-squared:  0.1893
## F-statistic: 5.281 on 3 and 52 DF,  p-value: 0.002963
```

ANOVA: EXEMPLO

```r
anova(m1)

## Analysis of Variance Table
##
## Response: cresc
##           Df Sum Sq Mean Sq F value   Pr(>F)
## trat       3 2.6014 0.86713  5.2807 0.002963 **
## Residuals 52 8.5388 0.16421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA: EXEMPLO

```
m1 <- aov(cresc ~ trat,mangue)
summary(m1)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## trat         3  2.601  0.8671   5.281 0.00296 **
## Residuals   52  8.539  0.1642
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m1)

## Analysis of Variance Table
##
## Response: cresc
##           Df Sum Sq Mean Sq F value   Pr(>F)
## trat       3 2.6014 0.86713  5.2807 0.002963 **
## Residuals 52 8.5388 0.16421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA: EXEMPLO

```
# O teste F sugere uma diferença "significativa"...mas entre quais tratamentos?
TukeyHSD(m1)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = cresc ~ trat, data = mangue)
##
## $trat
##                         diff         lwr       upr     p adj
## Foam-Control      0.38285714 -0.02364685 0.7893611 0.0717335
## Haliclona-Control 0.43571429  0.02921029 0.8422183 0.0312321
## Tedania-Control   0.58428571  0.17778172 0.9907897 0.0020035
## Haliclona-Foam    0.05285714 -0.35364685 0.4593611 0.9857121
## Tedania-Foam      0.20142857 -0.20507542 0.6079326 0.5576368
## Tedania-Haliclona 0.14857143 -0.25793257 0.5550754 0.7669696
```

ANOVA: EXEMPLO

```r
# Testes pareados geram um pouco de confusão...
# Melhor usar contrastes!
#
# Será que esponjas tem um efeito diferente da espuma artificial? (0,2,-1,-1)
c1 <- c(rep(0,14),rep(2,14),rep(-1,28))
mangue$c1 <- c1
m2 <- lm(cresc ~ c1, mangue)
anova(m2)

## Analysis of Variance Table
##
## Response: cresc
##           Df  Sum Sq Mean Sq F value Pr(>F)
## c1         1  0.1509 0.15088  0.7414  0.393
## Residuals 54 10.9893 0.20351
```

```r
# Testando múltiplas hipóteses usando contrastes

# 1) Será que esponjas tem um efeito diferente da espuma artificial? (0,2,-1,-1)

# 2) Será que a adição de eponja/espuma faz diferença em relação ao controle? (3,-1,-1,-1)

# 3) Será que existe diferença entre Tedania e Haliclonia? (0,1,-1,-1)

# 4) Será que...opa! Só podemos usar até r-1 contrastes, e r = 4!

c2 <- c(rep(3,14),rep(-1,42))
c3 <- c(rep(0,28),rep(1,14),rep(-1,14))

mangue$c2 <- c2
mangue$c3 <- c3

sum(mangue$c1*mangue$c2*mangue$c3)

## [1] 0
```

```r
# Testando múltiplas hipóteses usando contrastes

# 1) Será que esponjas tem um efeito diferente da espuma artificial? (0,2,-1,-1)

# 2) Será que a adição de eponja/espuma faz diferença em relação ao controle? (3,-1,-1,-1)

# 3) Será que existe diferença entre Tedania e Haliclonia? (0,1,-1,-1)

m3 <- lm(cresc ~ c1 + c2 + c3, mangue)

anova(m3)

## Analysis of Variance Table
##
## Response: cresc
##           Df Sum Sq Mean Sq F value   Pr(>F)
## c1         1 0.1509 0.15088  0.9188 0.342223
## c2         1 2.2960 2.29601 13.9824 0.000461 ***
## c3         1 0.1545 0.15451  0.9410 0.336519
## Residuals 52 8.5388 0.16421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA: EXEMPLO

```r
# Qual o efeito prático de se usarem diferentes contrastes?
# Obter diferentes partições de variância!
anova(m3)

## Analysis of Variance Table
##
## Response: cresc
##           Df Sum Sq Mean Sq F value   Pr(>F)
## c1         1 0.1509 0.15088  0.9188 0.342223
## c2         1 2.2960 2.29601 13.9824 0.000461 ***
## c3         1 0.1545 0.15451  0.9410 0.336519
## Residuals 52 8.5388 0.16421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m3)[1:3,2]

## [1] 0.1508762 2.2960095 0.1545143

sum(anova(m3)[1:3,2])

## [1] 2.6014
```

```
# Qual o efeito prático de se usarem diferentes contrastes?
# Obter diferentes partições de variância!
sum(anova(m3)[1:3,2])

## [1] 2.6014

anova(m1)

## Analysis of Variance Table
##
## Response: cresc
##           Df Sum Sq Mean Sq F value   Pr(>F)
## trat       3 2.6014 0.86713  5.2807 0.002963 **
## Residuals 52 8.5388 0.16421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Uma importante distinção ao se realizar uma ANOVA é quanto ao tipo de efeito do tratamento:

Efeito Fixo:

r representa o conjunto completo de todos os níveis de interesse possíveis, ou seja, uma população. Exemplo: Temos quatro alunos de estatística, será que o desempenho ao longo do curso foi diferente para cada um destes quatro?

Uma importante distinção ao se realizar uma ANOVA é quanto ao tipo de efeito do tratamento:

Efeito Fixo:

r representa o conjunto completo de todos os níveis de interesse possíveis, ou seja, uma população. Exemplo: Temos quatro alunos de estatística, será que o desempenho ao longo do curso foi diferente para cada um destes quatro?

Efeito Aleatório:

r representa um subconjunto de todos os níveis possíves de X, ou seja, uma amostra. Será que existe variação na performance dos alunos de Estatística, com base nesses quatro que eu observei?

A classificação dos efeitos se dá em função do tipo de inferência que se espera fazer para os níveis:

A classificação dos efeitos se dá em função do tipo de inferência que se espera fazer para os níveis:

"Aluno" como fator de efeito fixo: estamos preocupados em inferir se existem diferenças entre os quatro alunos especificados, em termos de desempenho.

A classificação dos efeitos se dá em função do tipo de inferência que se espera fazer para os níveis:

“Aluno” como fator de efeito fixo: estamos preocupados em inferir se existem diferenças entre os quatro alunos especificados, em termos de desempenho.

“Aluno” como fator de efeito aleatório: estamos preocupados em inferir se existe varição no desempenho dos alunos em geral, no curso de de estatística. Os quatro alunos avaliados são uma amostra aleatória do universo de todos os possíveis alunos da disciplina.

Quando os níveis de X são aleatórios, o modelo é espcificado da mesma maneira:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Quando os níveis de X são aleatórios, o modelo é espcificado da mesma maneira:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

A grande diferença é que α agora vem de um valor aleatório, e por isso tem uma distribuição com média e variância:

$$\alpha_i \sim N(0, \sigma^2_\alpha)$$

Quando os níveis de X são aleatórios, o modelo é espcificado da mesma maneira:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

A grande diferença é que α agora vem de um valor aleatório, e por isso tem uma distribuição com média e variância:

$$\alpha_i \sim N(0, \sigma^2_\alpha)$$

E o nosso modelo agora depende de duas variâncias: σ^2 e σ^2_α

O mais importante, contudo, é que nossa hipótese nula é bastante diferente.

O mais importante, contudo, é que nossa hipótese nula é bastante diferente.

Para o modelo de efeitos fixos, queremos avaliar se: $\alpha_1 = \alpha_2 = \ldots = \alpha_i$ (não há diferença entre tratamentos)

O mais importante, contudo, é que nossa hipótese nula é bastante diferente.

Para o modelo de efeitos fixos, queremos avaliar se: $\alpha_1 = \alpha_2 = \ldots = \alpha_i$ (não há diferença entre tratamentos)

Mas para o modelo de efeitos aleatórios, queremos avaliar se:

$\sigma^2_\alpha = 0$ (não há variação entre os diferentes tratamentos)

O mais importante, contudo, é que nossa hipótese nula é bastante diferente.

Para o modelo de efeitos fixos, queremos avaliar se:
$\alpha_1 = \alpha_2 = \ldots = \alpha_i$ (não há diferença entre tratamentos)

Mas para o modelo de efeitos aleatórios, queremos avaliar se:

$\sigma^2_\alpha = 0$ (não há variação entre os diferentes tratamentos)

De fato, no modelo aleatório, não faz sentido testar diferenças entre tratamentos: a cada nova realização do experimento, esses tratamentos seriam diferentes!

Para o modelo ANOVA com efeitos aleatórios, $E(MQ_{Reg})$ é:

$r\sigma_\alpha^2 + \sigma^2$

Para o modelo ANOVA com efeitos aleatórios, $E(MQ_{Reg})$ é:

$r\sigma_\alpha^2 + \sigma^2$

E ainda temos $E(MQ_{Res}) = \sigma^2$

Para o modelo ANOVA com efeitos aleatórios, $E(MQ_{Reg})$ é:

$r\sigma_{\alpha}^2 + \sigma^2$

E ainda temos $E(MQ_{Res}) = \sigma^2$

Se todos os níveis do fator são iguais, então $\sigma_{\alpha}^2 = 0$ e temos:

$$F = \frac{MQ_{Reg}}{MQ_{Res}} = \frac{r\sigma_{\alpha}^2 + \sigma^2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

Para o modelo ANOVA com efeitos aleatórios, $E(MQ_{Reg})$ é:

$r\sigma_\alpha^2 + \sigma^2$

E ainda temos $E(MQ_{Res}) = \sigma^2$

Se todos os níveis do fator são iguais, então $\sigma_\alpha^2 = 0$ e temos:

$$F = \frac{MQ_{Reg}}{MQ_{Res}} = \frac{r\sigma_\alpha^2 + \sigma^2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

Mas se os níveis são diferentes, então $\sigma_\alpha^2 > 0$, e $F > 1$

Importante: O teste F para ANOVA com efeitos fixos e aleatórios é igual apenas para análises com um um unico fator.

ANCOVA: Análise de Covariância

- Uma "mistura" de regressão e ANOVA
- Mistura variáveis explicativas categóricas (*dummy*) e contínuas
- Pode incluir termos de interação
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boldsymbol{\beta_3 X_1 X_2} + \varepsilon$

Extensão da ANOVA para mais de uma variável explicativa (todas categóricas)

- Geralmente também inclui termos de interação.
- Com três ou mais fatores, a interpretação fica bastante difícil.
- Fortemente baseada em desenhos experimentais controlados, que permitam a partição correta da variância, sem multicolinearidade ou pseudoreplicação.

A variável categórica altera o intercepto $(X_2 = (0, 1))$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Para o nível 1 de X_2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \times 0 + \varepsilon = \beta_0 + \beta_1 X_1 + \varepsilon$$

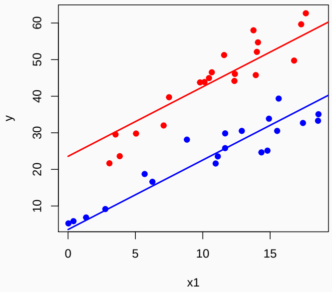
Para o nível 2 de X_2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \times 1 + \varepsilon$$

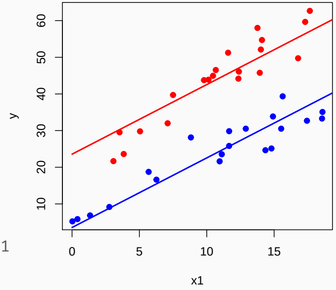
$$Y = (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon$$

ANCOVA

```r
set.seed(234)
x1 <- runif(40,0,20)
x2 <- c(rep(0,20),rep(1,20))
col <- x2; col[col==0] <- "blue"; col[col==1] <- "red"
y <- 3 + 2*x1 + 20*x2 + rnorm(20,0,5)

m <- lm(y ~ x1 + x2)

plot(x1,y, pch=19,col=col)

x21novo <- data.frame(x1 = seq(0,20,by=1), x2=rep(0,21))
x22novo <- data.frame(x1 = seq(0,20,by=1), x2=rep(1,21))

p1 <- predict(m,x21novo)
p2 <- predict(m,x22novo)

lines(seq(0,20,by=1),p1, lwd=2,col='blue')
lines(seq(0,20,by=1),p2, lwd=2,col='red')
```



```
summary(m)
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7356 -3.4624  0.9155  2.5250  8.3066
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.5609     1.7138   2.078   0.0447 *
## x1            1.8990     0.1323  14.358   <2e-16 ***
## x2           20.0080     1.3754  14.547   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.349 on 37 degrees of freedom
## Multiple R-squared:  0.9192,^^IAdjusted R-squared:  0.9148
## F-statistic: 210.5 on 2 and 37 DF,  p-value: < 2.2e-16
```



O termo de interação altera a inclinação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Para o nível 1 de X_2:

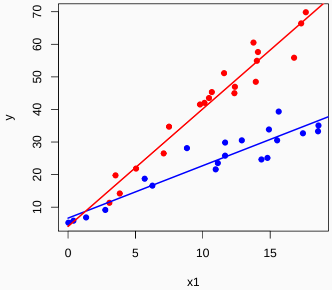
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \times 0 + \beta_3 \times X_1 \times 0 + \varepsilon = Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Para o nível 2 de X_2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \times 1 + \beta_3 \times X_1 \times 1 + \varepsilon$$

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times X_1 + \varepsilon$$

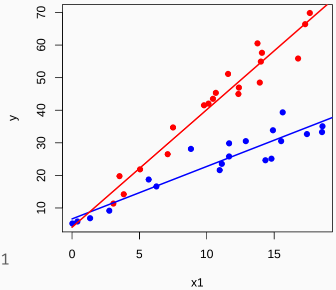
```
set.seed(234)
x1 <- runif(40,0,20)
x2 <- c(rep(0,20),rep(1,20))
y <- 3 + 2*x1 + 6*x2 + 1.2*x1*x2 + rnorm(20,0,5)

m <- lm(y ~ x1 * x2)

m <- lm(y ~ x1 * x2)

x21novo <- data.frame(x1 = seq(0,20,
              by=1), x2=rep(0,21))
x22novo <- data.frame(x1 = seq(0,20
                      ,by=1), x2=rep(1,21))

p1 <- predict(m,x21novo)
p2 <- predict(m,x22novo)

plot(x1,y, pch=19, col=col)

lines(seq(0,20,by=1),p1, lwd=2,col='blue')
lines(seq(0,20,by=1),p2, lwd=2,col='red')
```



ANCOVA

```r
summary(m)
```

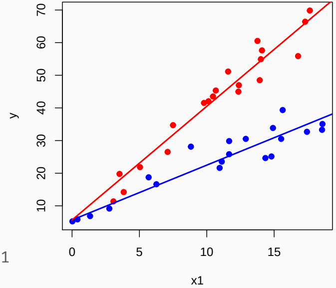
```
##
## Call:
## lm(formula = y ~ x1 * x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7702 -2.1950 -0.2836  2.7892  7.5328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6411     1.7961   3.697 0.000721 ***
## x1            1.6103     0.1475  10.920 5.62e-13 ***
## x2           -2.5561     2.9011  -0.881 0.384121
## x1:x2         1.9989     0.2453   8.149 1.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.875 on 36 degrees of freedom
## Multiple R-squared:  0.9511,^^IAdjusted R-squared:  0.947
## F-statistic: 233.4 on 3 and 36 DF,  p-value: < 2.2e-16
```



ANCOVA

```
m <- lm(y ~ x1 + x1:x2)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x1 + x1:x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2419 -2.3296 -0.0221  2.7361  7.6408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.6613     1.4062   4.026  0.00027 ***
## x1            1.6808     0.1235  13.607 5.63e-16 ***
## x1:x2         1.8030     0.1033  17.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.864 on 37 degrees of freedom
## Multiple R-squared:  0.95,^^IAdjusted R-squared:  0.9473
## F-statistic: 351.8 on 2 and 37 DF,  p-value: < 2.2e-16
```



Quando usar ANCOVA?

- Explicar diferenças em uma relação contínua entre grupos (ex. fertilizante em duas espécies de planta)
- O termo aditivo (novo intercepto) é interpretado como: mesma taxa de crescimento (inclinação), mas com níveis basais diferentes
- O termo de interação é intepretado como taxas (inclinações) diferentes. Ou seja, a resposta do fertilizante *interage* com o tipo de espécie para determinar uma resposta.
- Se esses coeficientes são pequenos e/ou tem p-valor alto, interpreta-se que há pouca evidência de diferenças na relação entre X_1 e Y para os diferentes grupos.