Lista 4 - Planejamento e Pesquisa 1

Victor Ribeiro Baião Decanini
Bruno de Castro Paul Schultze
Guilherme Tamborra
Gustavo de Oliveira Kanno
Marcos Soares Rodrigues
Rodrigo Marcel Araujo Oliveira
Rubens Santos Andrade Filho

18 de junho de 2020

Conteúdo

1 Exercício 1 **2**

1.1 a) Escreva e ajuste um modelo hierárquico para determinar se existem efeitos fixos de classe e de escola em relação à extroversão (extro). 2

1.2 b) Especifique e verifique as suposições do mesmo. 3

1.3 c) Repita (a) e (b) para capacidade social 3

1.4 d) Considere agora que as escolas foram sorteadas de maneira aleatória, qual seria o modelo a ser ajustado para a capacidade social? Obtenha a estimativa da componente de variância pelo método da análise de variância. 5

2 Exercício 2 **5**

2.1 Formule e ajuste um modelo apropriado para os dados de marketing apresentados na sala de aula 5

3 Exercício 3 **7**

3.1 Reproduza no R e apresente as saídas correspondentes do exemplo da resistência das fibras. 7

1 Exercício 1

1.1 a) Escreva e ajuste um modelo hierárquico para determinar se existem efeitos fixos de classe e de escola em relação à extroversão (extro).

O modelo hierárquico foi definido como:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

Na qual:

- y_{ijk} é a ijk-ésima observação da variável resposta (extroversão);
- μ é a média;
- α_i é o efeito do i-ésimo nível da escola;
- $\beta_{j(i)}$ é o efeito do j-ésimo nível da classe na escola de i-ésimo nível;
- $\varepsilon_{(ij)k}$ é o erro aleatório

Ao realizar uma ANOVA para os modelos fixos com hierarquia, obtivemos:

```
> aov.fit <- aov(extro ~ school/class, data = data)
> summary(aov.fit)
              Df Sum Sq Mean Sq F value Pr(>F)
school         5  95908   19182 19801.3 <2e-16 ***
school:class  18   7386     410   423.6 <2e-16 ***
Residuals   1176   1139       1
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

Concluímos então à partir do p-valor, que tanto o fator escola quanto sua interação com o fator classe são significativos, ao nível de significância de 0.05.

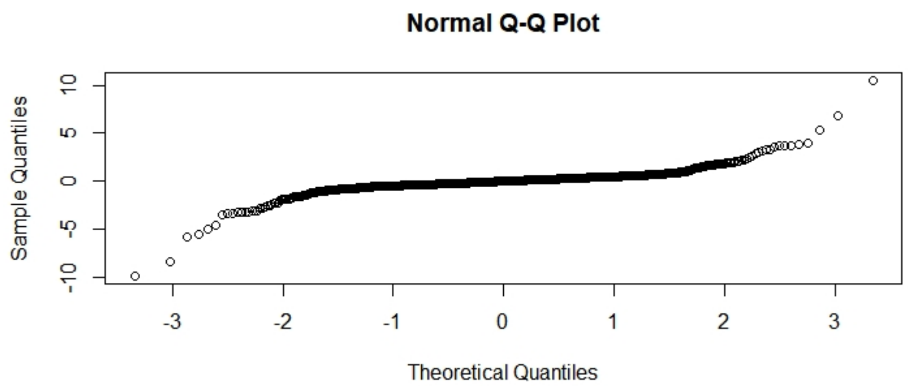
Agora, vamos propor um modelo linear misto para o mesmo problema, afim de identificar possíveis interações entre os fatores aninhados.

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.7949 -0.3299  0.0048  0.3379 10.4515

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       39.9971     0.1392 287.355  < 2e-16 ***
schoolII          12.2428     0.1968  62.195  < 2e-16 ***
schoolIII         16.8128     0.1968  85.411  < 2e-16 ***
schoolIV          20.6421     0.1968 104.864  < 2e-16 ***
schoolV           24.8935     0.1968 126.462  < 2e-16 ***
schoolVI          29.9259     0.1968 152.027  < 2e-16 ***
schoolI:classb     5.7780     0.1968  29.353  < 2e-16 ***
schoolII:classb    1.3942     0.1968   7.083 2.43e-12 ***
schoolIII:classb   0.9738     0.1968   4.947 8.62e-07 ***
schoolIV:classb    1.0348     0.1968   5.257 1.74e-07 ***
schoolV:classb     1.1847     0.1968   6.018 2.35e-09 ***
schoolVI:classb    1.9259     0.1968   9.784  < 2e-16 ***
schoolI:classc     8.5169     0.1968  43.267  < 2e-16 ***
schoolII:classc    2.5020     0.1968  12.710  < 2e-16 ***
schoolIII:classc   2.0060     0.1968  10.191  < 2e-16 ***
schoolIV:classc    2.1766     0.1968  11.057  < 2e-16 ***
schoolV:classc     2.1075     0.1968  10.706  < 2e-16 ***
schoolVI:classc    4.8824     0.1968  24.803  < 2e-16 ***
schoolI:classd    10.3912     0.1968  52.789  < 2e-16 ***
schoolII:classd    3.5719     0.1968  18.146  < 2e-16 ***
schoolIII:classd   2.9154     0.1968  14.811  < 2e-16 ***
schoolIV:classd    3.1436     0.1968  15.970  < 2e-16 ***
schoolV:classd     3.4605     0.1968  17.580  < 2e-16 ***
schoolVI:classd   10.4551     0.1968  53.113  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9842 on 1176 degrees of freedom
Multiple R-squared:  0.9891,    Adjusted R-squared:  0.9889
F-statistic:  4636 on 23 and 1176 DF,  p-value: < 2.2e-16
```

Como podemos ver, é notável que há interação entre todos os fatores dos níveis hierárquicos.

1.2 b) Especifique e verifique as suposições do mesmo.

Tivemos que supor que:
Os $\varepsilon_{(ij)k} \sim N(0, \sigma^2)$ e são independentes.
Para verificar normalidade, usamos o gráfico QQ dos resíduos exposto abaixo:



Como podemos ver, o gráfico nos da indício, através das caudas deslocadas, de que os dados não seguem uma distribuição normal, e para verificar, usamos os testes de Shapiro-Wilk e teste de Levene:

```
> qqnorm(y = model.fit$residuals)
> shapiro.test(model.fit$residuals)

        Shapiro-Wilk normality test

data:  model.fit$residuals
W = 0.72019, p-value < 2.2e-16

> library(car)
> leveneTest(model.fit)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group   23  36.261 < 2.2e-16 ***
      1176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

O teste de Shapiro-Wilk confirma a rejeição da normalidade dos dados, e o teste de Levene nos mostra que as variâncias não são homogêneas.

1.3 c) Repita (a) e (b) para capacidade social

O modelo hierárquico foi definido como:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

Na qual:

- y_{ijk} é a ijk-ésima observação da variável resposta (extroversão);
- μ é a média;
- α_i é o efeito do i-ésimo nível da escola;
- $\beta_{j(i)}$ é o efeito do j-ésimo nível da classe na escola de i-ésimo nível;
- $\varepsilon_{(ij)k}$ é o erro aleatório

Ao realizar uma ANOVA para os modelos fixos com hierarquia, obtivemos:

```
> aov2.fit <- aov(social ~ school/class, data = data)
> summary(aov2.fit)
               Df Sum Sq Mean Sq F value Pr(>F)
school          5   1459   291.9   1.216 0.2992
school:class   18   7824   434.7   1.811 0.0198 *
Residuals    1176 282229   240.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

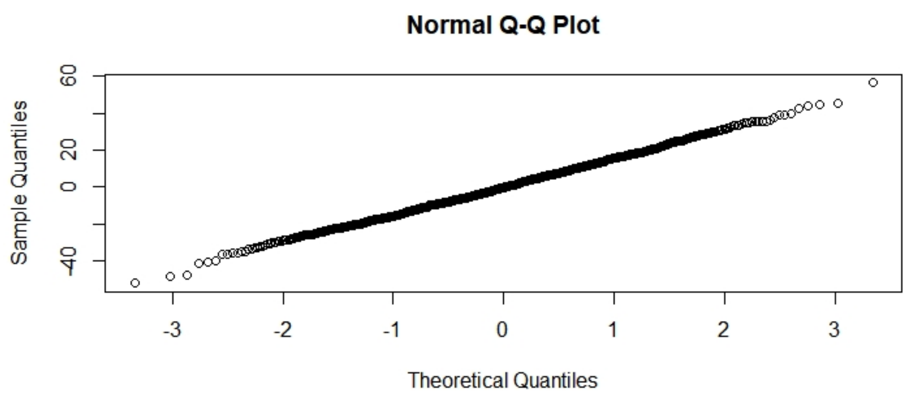
Concluímos então à partir do p-valor, que o fator escola é significativa, mas a interação com o fator classe não ao nível de significância de 0.05. A interação acima tem um p-valor de 0.29.

Logo, não rejeitamos a hipótese nula de igualdade dos efeitos de α para o fator escola, mas rejeitamos a hipótese nula de igualdade dos efeitos de β para a interação escola$\sim$ *classe*.

Agora, vamos propor um modelo linear misto para o mesmo problema, afim de identificar possíveis interações entre os fatores aninhados.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-52.334 -10.237  -0.233  10.298  56.477 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      103.43457    2.19085  47.212  < 2e-16 ***
schoolII          -8.63957    3.09833  -2.788  0.00538 ** 
schoolIII         -5.23306    3.09833  -1.689  0.09149 .  
schoolIV           2.98951    3.09833   0.965  0.33480    
schoolV           -4.79213    3.09833  -1.547  0.12221    
schoolVI          -2.06729    3.09833  -0.667  0.50476    
schoolI:classb    -7.08227    3.09833  -2.286  0.02244 *  
schoolII:classb    2.76188    3.09833   0.891  0.37289    
schoolIII:classb  -0.56538    3.09833  -0.182  0.85524    
schoolIV:classb   -8.52408    3.09833  -2.751  0.00603 ** 
schoolV:classb     1.00093    3.09833   0.323  0.74671    
schoolVI:classb   -3.90333    3.09833  -1.260  0.20798    
schoolI:classc    -0.81285    3.09833  -0.262  0.79310    
schoolII:classc    4.31330    3.09833   1.392  0.16414    
schoolIII:classc   6.05403    3.09833   1.954  0.05094 .  
schoolIV:classc   -7.37993    3.09833  -2.382  0.01738 *  
schoolV:classc    -2.11129    3.09833  -0.681  0.49573    
schoolVI:classc   -1.49666    3.09833  -0.483  0.62915    
schoolI:classd    -7.95631    3.09833  -2.568  0.01035 *  
schoolII:classd    5.42191    3.09833   1.750  0.08039 .  
schoolIII:classd   3.88728    3.09833   1.255  0.20986    
schoolIV:classd   -6.09958    3.09833  -1.969  0.04923 *  
schoolV:classd    -0.22833    3.09833  -0.074  0.94127    
schoolVI:classd   -0.01993    3.09833  -0.006  0.99487    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.49 on 1176 degrees of freedom
Multiple R-squared:  0.03185,	Adjusted R-squared:  0.01291 
F-statistic: 1.682 on 23 and 1176 DF,  p-value: 0.02308
```

Para o modelo acima, tivemos que supor que:
Os $\varepsilon_{(ij)k} \sim N(0, \sigma^2)$ e são independentes.
Para verificar normalidade, usamos o gráfico QQ dos resíduos exposto abaixo:



Como podemos ver, o gráfico já nos mostra um modelo bem ajustado à distribuição Normal dos dados e a homogeneidade das variâncias, mas para confirmar nossa hipótese, e para verificar, usamos os testes de Shapiro-Wilk e teste de Bartlett:

```
> qqnorm(y = model2.fit$residuals)
> shapiro.test(model2.fit$residuals)

	Shapiro-Wilk normality test

data:  model2.fit$residuals
W = 0.9989, p-value = 0.693

> bartlett.test(social ~ interaction(school, class), data)

	Bartlett test of homogeneity of variances

data:  social by interaction(school, class)
Bartlett's K-squared = 16.605, df = 23, p-value = 0.8282

> |
```

Como esperado, o teste de Shapiro-Wilk não rejeita a hipótese nula e confirma a normalidade dos dados. E o teste de Bartlett nos confirma a hipótese de que as variâncias são homogêneas.

1.4 d) Considere agora que as escolas foram sorteadas de maneira aleatória, qual seria o modelo a ser ajustado para a capacidade social? Obtenha a estimativa da componente de variância pelo método da análise de variância.

Continuamos com o modelo hierárquico e, já que vamos assumir o fator escolas como aleatório, por estar aninhado, o fator classe também é. Nesse caso, nosso modelo tem dois fatores aleatórios definidos por:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

Na qual:

- y é a ijk-ésima observação da variável resposta (capacidade social);
- μ é a média;
- α_i é o efeito aleatório do i-ésimo nível da escola;
- $\beta_{j(i)}$ é o efeito aleatório do j-ésimo nível da classe na escola de i-ésimo nível;
- $\varepsilon_{(ij)k}$ é o erro aleatório

Temos que os fatores são aleatórios $\Rightarrow$ α, β e ε são variáveis aleatórias independentes, e por isso: $\alpha \sim N(0, \sigma^2_\alpha)$, $\beta_{j(i)} \sim N(0, \sigma^2_\beta)$ e $\varepsilon_{(ij)k} \sim N(0, \sigma^2)$

```
> aov3.fit <- aov(social ~ school/class, data = data)
> summary(aov3.fit)
               Df Sum Sq Mean Sq F value Pr(>F)
school          5   1459   291.9   1.216 0.2992
school:class   18   7824   434.7   1.811 0.0198 *
Residuals    1176 282229   240.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

então as estimativas têm valor:

$\hat{\sigma}^2 = \frac{MS_\alpha - MS_E}{b \cdot n} = \frac{291.9 - 240}{4 \cdot 50} \approx 0.26$

$\hat{\sigma}^2_{\beta\alpha} = \frac{MS_{\beta\alpha} - MS_E}{n} = \frac{437.7 - 240}{50} \approx 3.89$

2 Exercício 2

2.1 Formule e ajuste um modelo apropriado para os dados de marketing apresentados na sala de aula

O modelo apropriado para este caso é:

$$y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}$$

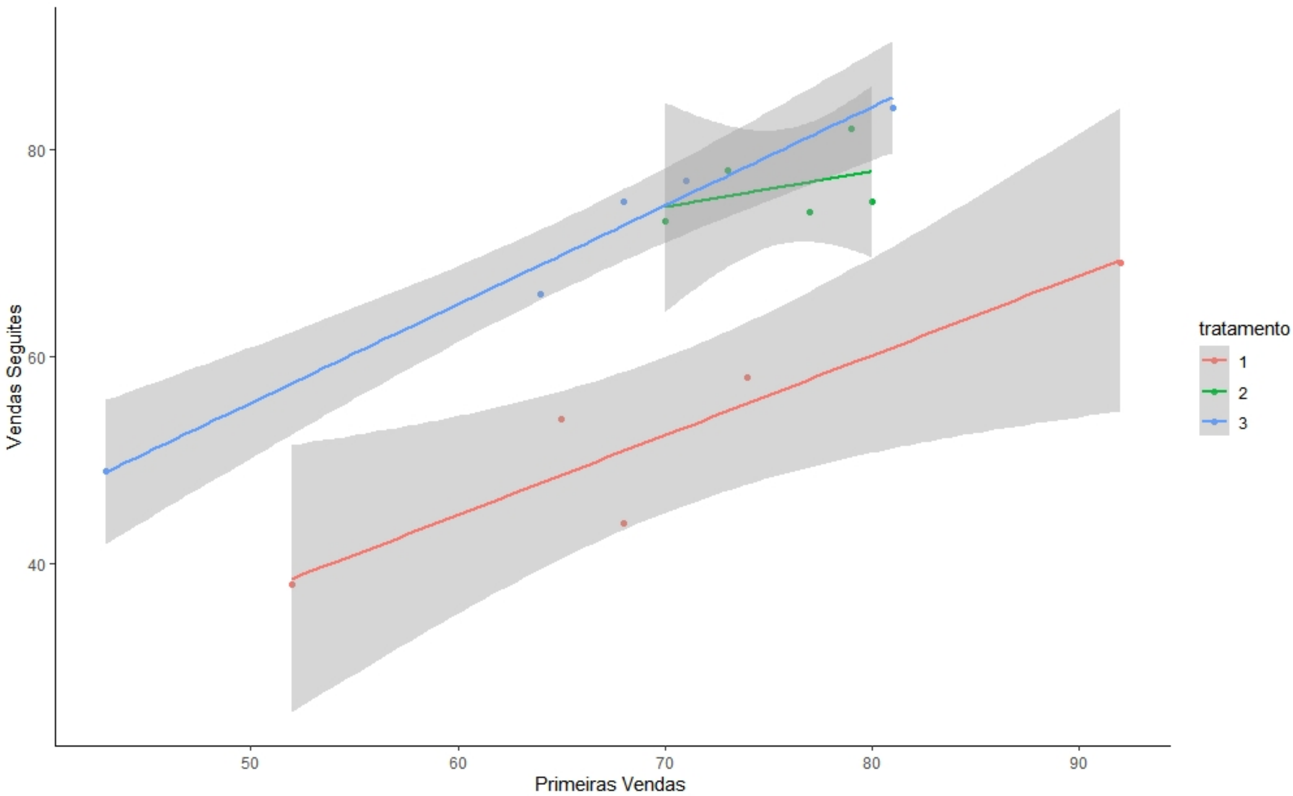
Na qual:

- y é a ij-ésima observação da variável resposta (valor das vendas);
- β_{0i} é o efeito do tratamento da imagem atual na área do estacionamento, imagem nova na área do estacionamento ou imagem nova em frente aos caixas;
- β_1 é o coeficiente que mede a relação linear;
- x_{ij} é o valor da venda do produto nas 3 semanas;
- ε_{ij} é o erro aleatório

Usamos a regressão linear para obter os resultados a seguir:

```
> tratamento = c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3)
> vendas.primeiras = c(92, 68, 74, 52, 65, 77, 80, 70, 73, 79, 64, 43, 81, 68, 71)
> vendas.seguintes = c(69, 44, 58, 38, 54, 74, 75, 73, 78, 82, 66, 49, 84, 75, 77)
> tab <- cbind(tratamento, vendas.primeiras, vendas.seguintes)
> tab <- as.data.frame(tab)
> aov2.fit <- aov(vendas.seguintes ~  factor(tratamento) + vendas.primeiras + 0, data = tab)
> summary(aov2.fit)
                   Df Sum Sq Mean Sq F value   Pr(>F)    
factor(tratamento)  3  67659   22553 1405.33 1.72e-14 ***
vendas.primeiras    1   1191    1191   74.19 3.21e-06 ***
Residuals          11    177      16                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> model.fit <- lm(vendas.seguintes ~ factor(tratamento) + vendas.primeiras + 0, data = tab)
> summary(model.fit)

Call:
lm(formula = vendas.seguintes ~ factor(tratamento) + vendas.primeiras + 
    0, data = tab)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7636 -2.7666  0.7781  2.4288  5.7406 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
factor(tratamento)1 -5.99860    7.03500  -0.853   0.4120    
factor(tratamento)2 13.12687    7.56107   1.736   0.1104    
factor(tratamento)3 15.60815    6.58624   2.370   0.0372 *  
vendas.primeiras     0.83474    0.09691   8.614 3.21e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.006 on 11 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9965 
F-statistic:  1073 on 4 and 11 DF,  p-value: 3.588e-14

> |
```

Abaixo temos um gráfico com os três tratamentos. É notável que o tratamento I apresentou uma queda nas semanas seguintes, já o tratamento II se manteve estável, e o tratamento III apresentou um aumento nas semanas seguintes.



3 Exercício 3

3.1 Reproduza no R e apresente as saídas correspondentes do exemplo da resistência das fibras.

Conforme o exemplo,consideramos o seguinte modelo:

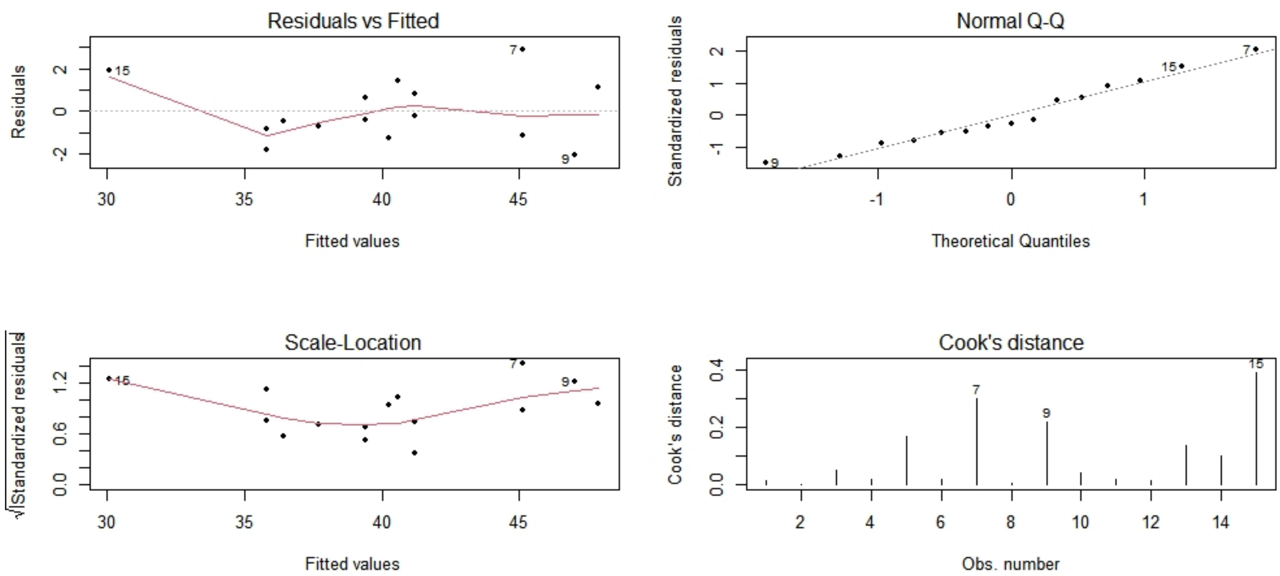
$$y_{ij} = \mu + \beta_i + \beta(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij}$$

```
> maquina = c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3)
> y = c(36, 41, 39, 42, 49, 40, 48, 39, 45, 44, 35, 37, 42, 34, 32)
> x = c(20, 25, 24, 25, 32, 22, 28, 22, 30, 28, 21, 23, 26, 21, 15)
> tab <- cbind(maquina, y, x)
> tab <- as.data.frame(tab)
> tab$x.media <- tab$x - mean(tab$x)
> tab$maquina <- as.factor(tab$maquina)
> model.fit <- lm(data = tab, y ~ maquina + x.media, contrasts = list(maquina = contr.sum))
> summary(model.fit)

Call:
lm(formula = y ~ maquina + x.media, data = tab, contrasts = list(maquina = contr.sum))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0160 -0.9586 -0.3841  0.9518  2.8920 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  40.2000     0.4118  97.611  < 2e-16 ***
maquina1      0.1824     0.5950   0.307    0.765    
maquina2      1.2192     0.6201   1.966    0.075 .  
x.media       0.9540     0.1140   8.365 4.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.595 on 11 degrees of freedom
Multiple R-squared:  0.9192,	Adjusted R-squared:  0.8972 
F-statistic: 41.72 on 3 and 11 DF,  p-value: 2.665e-06

> 
```

```
> aov.fit <- anova(model.fit)
> aov.fit
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
maquina    2 140.400  70.200  27.593 5.170e-05 ***
x.media    1 178.014 178.014  69.969 4.264e-06 ***
Residuals 11  27.986   2.544                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

```
> car::Anova(model.fit, type = "III")
Anova Table (Type III tests)

Response: y
             Sum Sq Df   F value    Pr(>F)    
(Intercept) 24240.6  1 9527.8944 < 2.2e-16 ***
maquina        13.3  2    2.6106    0.1181    
x.media       178.0  1   69.9694 4.264e-06 ***
Residuals      28.0 11                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Como podemos ver, o fator máquina tem estatística F=2.61 com p-valor $\geq$ 0.05, informando assim a não rejeição da hipótese nula, ou seja, a resistência das fibras independe da máquina utilizada.



Podemos ver através do gráfico de resíduos que o modelo proposto está bem ajustado, sugerindo normalidade dos dados.