16.1 Exemplo de aplicação

Dois fungicidas (trt) foram avaliados experimentalmente a campo em delineamento inteiramente casualizado com a cultura da soja. A variável resposta (y) é o rendimento de grãos, em sacas/ha. Ocorre que, por ocasião da aplicação dos fungicidas, as parcelas apresentavam diferentes níveis de incidência e severidade da doença. Assim, a covariável (x) representando a área foliar (%) **livre** da doença em cada parcela foi também mensurada para fins de controle experimental. O conjunto de dados é ancova.csv.

```
> soja <- read.csv("https://raw.githubusercontent.com/
    arsilva87/statsbook/main/datasets/ancova.csv",
    colClasses = c("factor", "numeric", "numeric"))
> str(soja)
'data.frame':   20 obs. of  3 variables:
 $ trt: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1  ...
 $ x  : num  58.4 50.9 60.5 51.3 68.4 51.7 49.8 57.1 ...
 $ y  : num  51.4 47.9 48.5 41.3 58.4 39.7 44.8 45.1 ...
```

Observe a seguir as médias de rendimento. Há 1.84 sc/ha (110.4 kg/ha) de diferença. Mas, pela ANOVA, essa diferença não é detectável, mesmo não havendo problemas com os pressupostos do modelo.

```
> aggregate(y ~ trt, data = soja, FUN = mean)
  trt     y
1   1 48.28
2   2 50.12

> anova(lm(y ~ trt, data = soja))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
trt        1  16.93  16.928  0.5326 0.4749
Residuals 18 572.07  31.782
```

Embora não haja efeito dos tratamentos fungicidas sobre a covariável x, é razoável supor que o efeito (τ_i) dos fungicidas sobre os dados de rendimento de grãos (y_{ij}) seja influenciado pelo valor de x nas respectivas parcelas (x_{ij}). Assim, deve-se incluir no modelo estatístico para o rendimento de grãos o efeito (b) dessa covariável.

$$y_{ij} = \mu + b(x_{ij} - \bar{x}) + \tau_i + \epsilon_{ij}$$

Note que tal modelo é uma mistura dos modelos de regressão linear e de análise de variância. O termo b representa o coeficiente de regressão, efeito linear de x sobre y em ambos os tratamentos. Observe que x é centrada na média, para que sua inclusão não afete a estimativa da média geral μ. Observe o ajuste do modelo de ANCOVA no R:

```
> mod <- lm(y ~ I(x - mean(x)) + trt, data = soja)
> mod

Call:
lm(formula = y ~ I(x - mean(x)) + trt, data = soja)

Coefficients:
   (Intercept)  I(x - mean(x))             trt2
       47.0969          0.8273           4.2061
```

Em que se obtém as estimativas: $\bar{y}_1^* = 47.09$, $\bar{y}_2^* = 47.09 + 4.2061 = 51.2961$, das médias de tratamentos, e $\hat{b} = 0.8273$ o efeito da covariável x.

Uma vez que os valores *ajustados* de rendimento (y) para a covariável x são dados por:

$$y_{ij} - b(x_{ij} - \bar{x}) = \mu + \tau_i + \epsilon_{ij}$$

logo percebe-se que as *médias ajustadas* de rendimento podem ser obtidas por:

$$\begin{aligned}
\bar{y}_i^* &= \bar{y}_i - \hat{b}(\bar{x}_i - \bar{x}) \\
\bar{y}_1^* &= 48.28 - 0.8273(57.48 - 56.05) = 47.09 \\
\bar{y}_2^* &= 50.12 - 0.8273(54.62 - 56.05) = 51.30
\end{aligned}$$

As diferenças entre tratamentos em relação aos valores de x contribuem com a diferença em rendimento. Assim, note que o modelo de ANCOVA permitiu *amplificar* para 4.21 sc/ha as diferenças entre tratamentos, de modo que as médias ajustadas são estatisticamente diferentes (p = 0.0026). Perceba também que um grau de liberdade foi retirado do resíduo para estimar o efeito da covariável x.

```
> anova(mod)
Analysis of Variance Table

Response: y
               Df Sum Sq Mean Sq F value    Pr(>F)
I(x - mean(x))  1 391.49  391.49  58.305 6.843e-07 ***
trt             1  83.36   83.36  12.415  0.002608 **
Residuals      17 114.15    6.71
```

```
---
Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1
```

A figura 16.1 nos ajuda a perceber o *ajuste* feito pela covariável x no efeito dos tratamentos sobre y. Note que os dados de cada tratamento estão dispersos ao longo de retas (efeito linear de x sobre y). Essas retas possuem a mesma inclinação $\hat{b} = 0.8273$, mas estão deslocadas no eixo x pela diferença existente no campo em termos de ataque da doença nas parcelas, tornando de certa forma 'injusta' a comparação dos tratamentos. As médias ajustadas, por outro lado, representam as médias de rendimento dos tratamentos agora centradas num valor comum de ataque da doença ($\bar{x}$), a média geral de x.

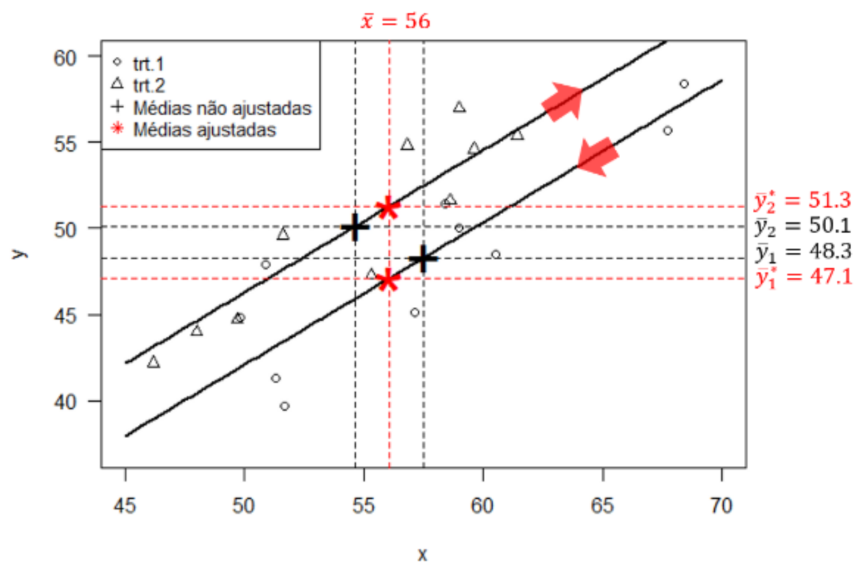

Figura 16.1: Ilustração da análise de covariância

O fato de haver apenas dois tratamentos dispensa testes de médias. Não obstante, a diferença de rendimento entre tratamentos poderia ser testada através de testes post-hoc, utilizando as médias ajustadas[1]. A seguir o resultado do teste t (LSD de Fisher) com o pacote *emmeans* (Lenth, 2019).

```
> library(emmeans)
> pairs(lsmeans(mod, "trt"))
 contrast estimate   SE df t.ratio p.value
 1 - 2       -4.21 1.19 17  -3.523  0.0026
```

[1]Ou médias de mínimos quadrados (*least square means*, do inglês).