EXERCISE 2

$G = 4$ treatments, $n_g = 5$ for $g = 1, \dots, 4$ $\Rightarrow$ $N = \sum_{g=1}^{4} n_g = 20$

a) we have a linear model with dummy variables to encode the treatments.

Define:

$$x_{i2} = \begin{cases} 1 & \text{if tree}_i \text{ has treatment 2} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i4} = \begin{cases} 1 & \text{if tree}_i \text{ has treatment 4} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if tree}_i \text{ has treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

then,

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \qquad i = 1, \dots, 20$$

in matrix form:

observations are sorted according to the treatment:

$$\underline{Y} = [\underbrace{Y_1, \dots, Y_5}_{\text{group 1}}, \underbrace{Y_6, \dots, Y_{10}}_{\text{group 2}}, \underbrace{Y_{11}, \dots, Y_{15}}_{\text{group 3}}, \underbrace{Y_{16}, \dots, Y_{20}}_{\text{group 4}}]^T \qquad \underline{Y} \sim N_{20}(X\underline{\beta}, \sigma^2 I_{20})$$

the model matrix

$$X = [\underline{1} \; \underline{x}_2 \; \underline{x}_3 \; \underline{x}_4] = \begin{array}{c} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} i=1 \\ \vdots \\ i=5 \\ i=6 \\ \vdots \\ i=10 \\ i=11 \\ \vdots \\ i=15 \\ i=16 \\ \vdots \\ i=20 \end{array}$$

dimension $N \times G = 20 \times 4$

$\underline{\beta}$ is a 4-dim. vector

$$\underline{\beta} = [\beta_1 \; \beta_2 \; \beta_3 \; \beta_4]^T$$

$\underline{\varepsilon}$ is a N-dim vector

$$\underline{\varepsilon} = [\varepsilon_1 \dots \varepsilon_{20}]^T \qquad \underline{\varepsilon} \sim N_{20}(\underline{0}, \sigma^2 I_{20})$$

b) sample space $\mathcal{Y} = \mathbb{R}^{20}$

parameter space $\Theta = \mathbb{R}^4 \times (0, +\infty)$

c) in general, the sum of squares decomposition is

$$\underbrace{\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2}_{SST} = \underbrace{\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\hat{y}_{ig} - \bar{y})^2}_{SSR} + \underbrace{\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \hat{y}_{ig})^2}_{SSE}$$

however we know that, in this model:

$\hat{y}_{ig} = \bar{y}_g$ for $i = 1, \dots, n_g$

hence

$$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\hat{y}_{ig} - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{y}_g - \bar{y})^2 = \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2 = n \sum_{g=1}^{G} (\bar{y}_g - \bar{y})^2$$

and that

$$\sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 \cdot \frac{1}{(n_g - 1)} = s_g^2 \Rightarrow \sum_{i=1}^{n_g} (y_{ig} - \hat{y}_{ig})^2 = \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 = (n_g - 1) s_g^2 = (n-1) s_g^2$$

Hence:

$$\underset{\substack{\| \\ SST}}{\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2} = \underbrace{n \sum_{g=1}^{G} (\bar{y}_g - \bar{y})^2}_{\substack{\text{BETWEEN-GROUP} \\ \text{SUM OF SQUARES} \\ \| \\ SSR}} + \underbrace{(n-1) \sum_{g=1}^{G} s_g^2}_{\substack{\text{WITHIN-GROUP} \\ \text{SUM OF SQUARES} \\ \| \\ SSE}}$$

The overall mean $\bar{y} = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} y_{ig} = \frac{1}{N} \sum_{g=1}^{G} n_g \bar{y}_g$

in this case $n_g = 5 = n$ for all $g = 1, \dots, G$ $\Rightarrow$ $\bar{y} = \frac{1}{N} \sum_{g=1}^{G} n \cdot \bar{y}_g = \frac{1}{Gn} \cdot n \sum_{g=1}^{G} \bar{y}_g = \frac{1}{G} \sum_{g=1}^{G} \bar{y}_g$

With the data, $\bar{y} = 0.5035$

We can compute the regression sum of squares (between-groups SS) as

$$5 \cdot [(0.184 - 0.5035)^2 + (0.332 - 0.5035)^2 + (0.164 - 0.5035)^2 + (1.334 - 0.5035)^2] = 4.682 = SSR$$

The error sum of squares (within-groups SS) is

$$4 \cdot [0.016 + 0.319 + 0.015 + 0.737] = 4.3564 = SSE$$

Finally, $SST = SSR + SSE = 9.038815$

d) the type of treatment does not have an effect $\Leftrightarrow$ the groups have the same mean weight

$\Leftrightarrow$ $\beta_1 = \beta_1 + \beta_2 = \beta_1 + \beta_3 = \beta_1 + \beta_4$

$\Leftrightarrow$ $\beta_2 = \beta_3 = \beta_4 = 0$

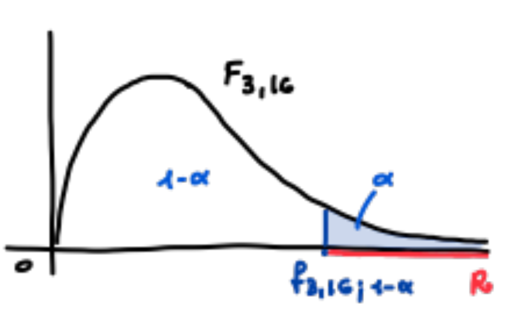
$$\begin{cases} H_0: \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1: \exists r \in \{2, 3, 4\} : \beta_r \neq 0 \end{cases}$$ test of overall significance

the test statistic is

$$F = \frac{\tilde{\Sigma}^2 - \hat{\Sigma}^2}{\hat{\Sigma}^2} \cdot \frac{N - G}{G - 1} \overset{H_0}{\sim} F_{G-1, N-G} = F_{3,16}$$

and the reject region at a 5% significance level is

$$R = (f_{3,16; 0.95}; +\infty)$$
$$= (3.2389; +\infty)$$



the observed value of the test

$$f^{obs} = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \cdot \frac{16}{3}$$

where $\hat{\sigma}^2$ is the estimate of the variance of the full model (H_1)

$\tilde{\sigma}^2$ is the estimate under the null model (H_0)

$$\hat{\sigma}^2 = \frac{SSE}{N}, \qquad \tilde{\sigma}^2 = \frac{SST}{N}$$

$$\Rightarrow f^{obs} = \frac{SST - SSE}{SSE} \cdot \frac{16}{3} = \frac{SSR}{SSE} \cdot \frac{16}{3} = \frac{\text{between-groups SS}}{\text{within-groups SS}} \cdot \frac{16}{3} = \frac{4.682}{4.3564} \cdot \frac{16}{3} = 5.7325$$

$f^{obs} \in R$ $\Rightarrow$ reject H_0 at a 5% significance level.

e) We know that, according to the specified model

$$\begin{cases} \mu_1 = \beta_1 \\ \mu_2 = \beta_1 + \beta_2 \\ \mu_3 = \beta_1 + \beta_3 \\ \mu_4 = \beta_1 + \beta_4 \end{cases} \Leftrightarrow \begin{cases} \beta_1 = \mu_1 \\ \beta_2 = \mu_2 - \mu_1 \\ \beta_3 = \mu_3 - \mu_1 \\ \beta_4 = \mu_4 - \mu_1 \end{cases}$$

Moreover, the MLE of μ_g is $\bar{y}_g = \frac{1}{n} \sum_{i=1}^{n} y_{ig}$ $\quad g = 1, \dots, G$

Hence, the MLE of $\underline{\beta}$ is

$$\hat{\underline{\beta}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \bar{y}_3 - \bar{y}_1 \\ \bar{y}_4 - \bar{y}_1 \end{bmatrix} = \begin{bmatrix} 0.184 \\ 0.148 \\ -0.02 \\ 1.15 \end{bmatrix}$$

The estimated model is $\hat{y}_i = 0.184 + 0.148\, x_{i1} - 0.02\, x_{i2} + 1.15\, x_{i3}$

f) $\hat{\underline{B}} \sim N_4(\underline{\beta}, (X^T X)^{-1} \sigma^2)$

and

$\hat{B}_2 \sim N(\beta_2, 0.40\, \sigma^2)$

g) $\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$

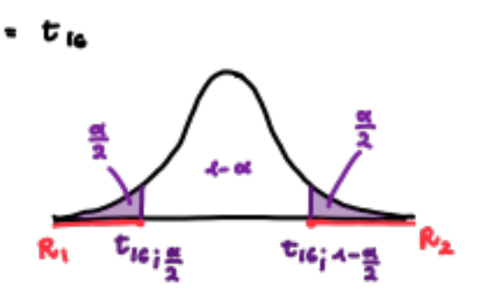
the distribution of $\hat{B}_2$ is given above, but σ^2 is unknown

$$\frac{\hat{B}_2}{\sqrt{0.40\, \sigma^2}} \overset{H_0}{\sim} N(0,1)$$

We estimate σ^2:

$$\frac{\hat{B}_2}{\sqrt{0.40\, \hat{\Sigma}^2}} \overset{H_0}{\sim} t_{N-G} = t_{16}$$

Two-tail reject region



$$R = R_1 \cup R_2$$
$$= (-\infty; -t_{16; 1-\frac{\alpha}{2}}) \cup (t_{16; 1-\frac{\alpha}{2}}; +\infty)$$
$$= (-\infty; -2.119) \cup (2.119; +\infty)$$

$\frac{\alpha}{2} = 0.025$

$1 - \frac{\alpha}{2} = 0.975$

The observed value of the test is:

$$t^{obs} = \frac{\hat{\beta}_2}{\sqrt{0.40\, \hat{\sigma}^2}} = \frac{0.148}{\sqrt{0.40 \cdot \frac{4.3564}{20}}} = 0.5014$$

$t^{obs} \notin R$ | do not reject H_0

Hence, treatments (1) and (2) are not statistically different

Applying no treatment or only fertilizer have the same effect on the trees' weight.

h) $R^2 = \frac{SSR}{SST} = 0.5180$

It represents the proportion of variability of Y that is explained by the model.

Here, we explain about half of the total variability.

i) $\sum_{i=1}^{20} e_i = 0$ yes, because the model includes the intercept.

$\Rightarrow$ $\underline{1} \in \mathcal{C}(X)$ and we know that $\underline{e}^T \underline{v} = 0$ for all $\underline{v} \in \mathcal{C}(X)$

$\Rightarrow$ $\underline{e}^T \underline{1} = \sum_{i=1}^{20} e_i = 0$

$\sum_{i=1}^{5} e_i = 0$ yes: $\sum_{i=1}^{5} e_i = \sum_{i=1}^{20} e_i - \sum_{i=6}^{20} e_i = \underline{e}^T \underline{1} - (\underline{e}^T (\underline{x}_2 + \underline{x}_3 + \underline{x}_4))$

$$= \underline{e}^T (\underbrace{\underline{1} - \underline{x}_2 - \underline{x}_3 - \underline{x}_4}_{\substack{\text{linear combination} \\ \text{of vectors } \in \mathcal{C}(X) \Rightarrow \in \mathcal{C}(X)}}) = 0$$

$\sum_{i=1}^{10} e_i = 0$ yes $\sum_{i=1}^{10} e_i = \underline{e}^T (\underline{1} - \underline{x}_3 - \underline{x}_4) = 0$

$\sum_{i=4}^{12} e_i = 0$ no