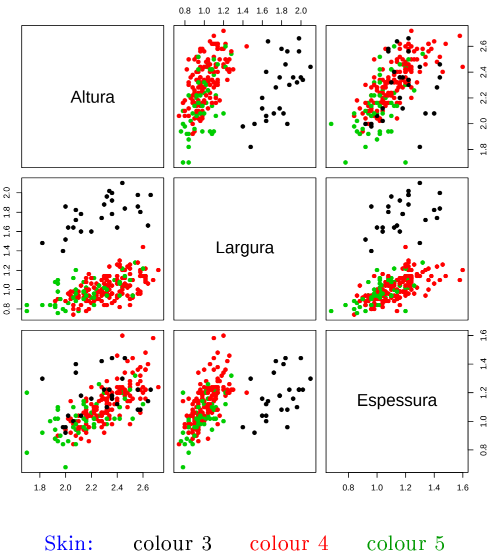
INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2020-21)

April 19, 2021 **Test – Linear Model** Duration: 2h30

I [16 points]

A study of garlic involved the measurements (in cm) of the height (variable `Altura`), width (`Largura`) and depth (`Espessura`) of garlic cloves. The skin colour of the cloves, with three categories (that were coded as colours 3, 4 and 5) was also registered. Below are the means and variances of each variable, for all the cloves and for the cloves with each skin colour. Also shown are the linear correlation coefficients (for all cloves), as well as the scatterplots for each pair of numerical variables with a colour code for the skin colours.



Means

dataset	size	Altura	Largura	Espessura
All	223	2.282870	1.095605	1.112242
Colour 3	26	2.273077	1.777692	1.178462
Colour 4	143	2.330070	1.016084	1.131958
Colour 5	54	2.162593	0.977778	1.028148

Variances

dataset	Altura	Largura	Espessura
All	0.04117731	0.07903465	0.02178864
Colour 3	0.05315015	0.03255446	0.02281354
Colour 4	0.03005563	0.01532540	0.02028487
Colour 5	0.04608372	0.01519497	0.01575122

Correlations (entire dataset)

```
             Altura   Largura Espessura
Altura    1.0000000 0.2546324 0.6402415
Largura   0.2546324 1.0000000 0.4229474
Espessura 0.6402415 0.4229474 1.0000000
```

1. A first model only considered the observations of cloves with skin colour 4. The clove depth was modelled as a linear regression on height and width, with the following results:

```
Call:
lm(formula = Espessura ~ Altura + Largura, data = alho.cordentes[alho.cordentes$Cor==4, ])
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31984    0.10028  -3.190  0.00176
Altura       0.46123    0.05141   8.973 1.70e-15
Largura      0.37113    0.07199   5.155 8.46e-07
---
Residual standard error: 0.08846 on 140 degrees of freedom
Multiple R-squared:  0.6197,  Adjusted R-squared:  0.6143
F-statistic: ??? on ??? and ??? DF,  p-value: < 2.2e-16
```

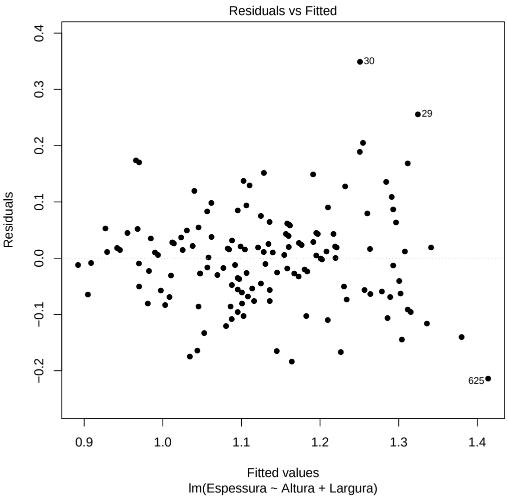
(a) Write the equation of the fitted regression plane in 3-dimensional space.

(b) What is the fitted value for the response variable `Espessura` when both predictors take their observed mean value for the data used to fit the model (skin colour 4)?

(c) Discuss and test ($\alpha = 0.05$) the goodness-of-fit of this model.

(d) For a fixed clove height, calculate a 95% confidence interval for the mean variation in its depth associated with an increase in the width of 1 mm (0.1 cm).

(e) Below is the scatterplot of (usual) residuals versus fitted values of the response variable. Discuss it.



(f) The observed clove depth (`Espessura`) for the observation labeled `30`, at the top of the plot above, is 1.60. Its *standarized* residual is 3.9924. *Without doing any calculations*, what can be said about the fitted width value for this observation?

2. Now consider ANOVA models of one of the numerical variables on the factor skin colour.

   (a) *Without doing any calculations*, which response variable would you expect to have the most significant skin colour effects? Justify your reply.

   (b) For the response variable you chose in the previous question, build the summary table and discuss your results. What conclusions can you draw regarding the relation between your chosen response variable and clove skin colour?

3. It was now decided to model clove width (`Largura`) as a function of clove height (variable `Altura`), but allowing for different linear regressions for cloves of each skin colour.

   (a) A first model admits that for each skin colour, there may be a totally different regression line. Here are the results for this model:

```
Call: lm(formula = Largura ˜ Altura * Cor, data = alho.cordentes)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.69794    0.21405   3.261  0.00129
Altura       0.47502    0.09370   5.069 8.54e-07
Cor4        -0.60257    0.24645  -2.445  0.01528
Cor5        -0.47481    0.26148  -1.816  0.07077
Altura:Cor4 -0.07987    0.10730  -0.744  0.45745
Altura:Cor5 -0.12606    0.11644  -1.083  0.28016
---
Residual standard error: 0.108 on 217 degrees of freedom
Multiple R-squared:  0.8557,  Adjusted R-squared:  0.8524
F-statistic: 257.4 on 5 and 217 DF,  p-value: < 2.2e-16
```

   i. Write down the equation of the fitted regression line for the cloves with skin colour 4.

   ii. Test whether the slope of the fitted lines for skin colours 3 and 5 are significantly different.

   (b) A second model assumed that the regression lines for all three skin colours were parallel, but could have different intercepts. Here are the results:

```
Call: lm(formula = Largura ˜ Altura + Cor, data = alho.cordentes)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.88138    0.08898   9.905   <2e-16
Altura       0.39432    0.03802  10.370   <2e-16
Cor4        -0.78408    0.02309 -33.962   <2e-16
Cor5        -0.75635    0.02608 -29.006   <2e-16
---
Residual standard error: 0.1078 on 219 degrees of freedom
Multiple R-squared:  0.8549,  Adjusted R-squared:  0.8529
F-statistic: 430.2 on 3 and 219 DF,  p-value: < 2.2e-16
```

Discuss in detail a test to determine whether this model has a significantly worse fit than the previous one. What are the practical implications of this result?

(c) What is the coefficient of determination for a single regression line of `Largura` over `Altura`, regardless of skin colours? Is it significantly different from zero? What practical conclusions and general lessons can be drawn from this result?

II [4 points]

1. In the Linear Model, the probability distribution of the vector of response values is known to be $\vec{\mathbf{Y}} \frown \mathcal{N}_n(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n)$. Show that the vector $\frac{1}{\sigma}(\vec{\mathbf{Y}}-\mathbf{X}\vec{\boldsymbol{\beta}})$ has the Standard Multinormal distribution $\mathcal{N}_n(\vec{\mathbf{0}}, \mathbf{I}_n)$.

2. Let $\mathbf{X}_c$ be the model matrix of a Linear Model with p predictors and $\mathbf{X}_s$ be the model matrix of one of its submodels, with only $k<p$ predictors. Let $\mathbf{H}_c$ and $\mathbf{H}_s$ be the corresponding hat matrices.

   (a) Show that the column-space of $\mathbf{X}_s$ is contained in the column-space of $\mathbf{X}_c$, that is, that any vector $\vec{\mathbf{y}} \in \mathcal{C}(\mathbf{X}_s)$ must also belong to $\mathcal{C}(\mathbf{X}_c)$.

   (b) Show that it must be the case that $\mathbf{H}_c\mathbf{H}_s = \mathbf{H}_s$, and also $\mathbf{H}_s = \mathbf{H}_s\mathbf{H}_c$. Interpret these results.

   (c) The application of Cochran's Theorem justifying the partial F test that compares the full model and the submodel requires that, in the decomposition $\mathbf{I}_n = (\mathbf{I}_n - \mathbf{H}_c) + (\mathbf{H}_c - \mathbf{H}_s) + (\mathbf{H}_s - \mathbf{P}_{\vec{\mathbf{1}}_n}) + \mathbf{P}_{\vec{\mathbf{1}}_n}$, the first two terms on the right-hand side of the equation be matrices of orthogonal projections which, when multiplied give a matrix of zeros, $\mathbf{0}_n$.

      i. Knowing that the matrices of orthogonal projections onto subspaces of $\mathbb{R}^n$ are the $n \times n$ symmetric and idempotent matrices, show that both $(\mathbf{I}_n - \mathbf{H}_c)$ and $(\mathbf{H}_c - \mathbf{H}_s)$ are orthogonal projections matrices. **Note:** You may assume that both hat matrices are matrices of orthogonal projections.

      ii. Show that $(\mathbf{I}_n - \mathbf{H}_c)(\mathbf{H}_c - \mathbf{H}_s) = \mathbf{0}_n$.