INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2020-21

November 4, 2020 **First Test** **Duration: 2h00**

I [15 points]

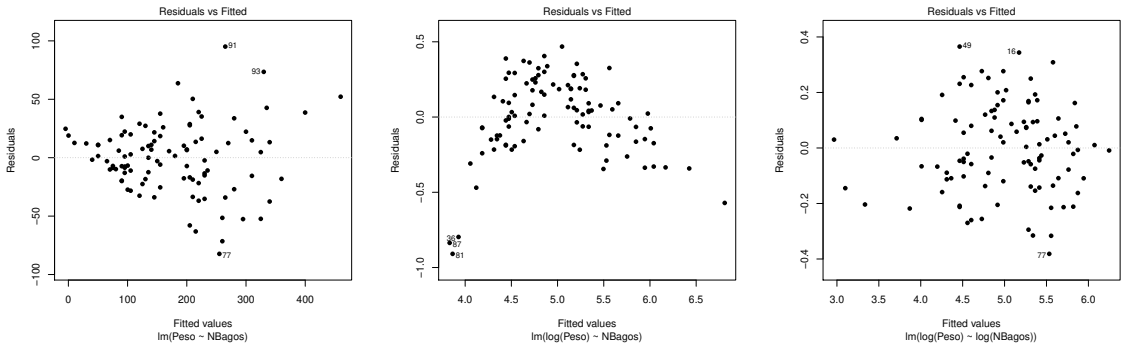
A 2014 study with the grape variety Viosinho attempted to model bunch weights, which is a variable of importance in production. A robot with a camera crossed an ISA experimental vineyard, taking photos of 100 bunches of grapes. These bunches were subsequently collected and weighted (variable `Peso`, in g). In the images obtained, there are two potential predictors of bunch weight: the number of visible berries in a bunch (variable `NBagos`) and the visible area of the bunch (variable `Area`, in cm^2). Here are some indicators, as well as the correlation matrix, of the observed variables.

	Peso	NBagos	Area
Minimum	19.20	10.00	21.92
Maximum	512.00	103.00	210.45
Mean	175.84	46.15	101.02
Std. Deviation	98.24383	18.74355	39.55328

```
          Peso NBagos   Area
Peso    1.0000 0.9530 0.8893
NBagos 0.9530 1.0000 0.9217
Area   0.8893 0.9217 1.0000
```

1. It was decided to fit a simple linear regression model for bunch weight.
   (a) Which predictor variable is best and what is the proportion of observed variability of bunch weight that is accounted for by that model? Test whether that proportion is significantly different from zero. Justify and comment your replies.
   (b) Fit the regression line for the predictor that you chose above. What are the units of measurement and the biological meaning of the slope of the fitted line?

2. It was decided to model bunch weight (`Peso`) from the number of visible berries in the images (`NBagos`), but allowing for the possibility of variable transformations. To help choose a model, (standard) residuals were plotted against fitted values for three models: on the left without any variable transformations; in the middle with a log-transformation of `Peso` only; on the right with a log-transformation of both variables. Discuss the implications of these plots.



3. A linear regression of the (natural) logarithm of bunch weight on the logarithm of the number of visible berries for each bunch was fitted, with the following results:

```
Call: lm(formula = log(Peso) ~ log(NBagos), data = viosinho2014)
Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.27446    0.13686  -2.005   0.0477
log(NBagos)  1.40715    0.03633  38.736   <2e-16
---
Residual standard error: 0.163 on 98 degrees of freedom
Multiple R-squared:  0.9387,  Adjusted R-squared:  0.9381
F-statistic:  1500 on 1 and 98 DF,  p-value: < 2.2e-16
```

(a) What is the estimated variance of the random errors in the model?

(b) The 95% prediction interval for an observation of this model's response variable, given a bunch with 50 visible berries, is $]\,4.9051\,,\,5.5556\,[$.

i. What is the estimated value $\hat{\sigma}_{indiv}$ of the variability associated with this prediction?

ii. Interpret this interval in terms of bunch weights (in g).

(c) Is it possible to describe the underlying relation by stating that the squared bunch weights are proportional to the cube of the number of visible berries? Justify using an appropriate confidence interval. Comment your result.

4. Finally, a multiple linear regression model was fitted using the log-transformations of the response variable and of both predictors. Here are the results:

```
Call: lm(formula = log(Peso) ~ log(NBagos) + log(Area), data = viosinho2014)
Coefficients:
(Intercept)  log(NBagos)    log(Area)
    -0.6237       1.0366       0.3835
```

(a) What can be stated about the coefficient of determination of this fitted model?

(b) Knowing that the image for the lightest bunch (minimum weight) had 11 visible berries and a visible area of 24.22 cm^2, compute the residual for this observation.

(c) What is the fitted relation between the original (non-transformed) variables? Justify your answer. For the lightest bunch, compute the difference between the observed weight and the weight predicted by this non-linear relation. Comment your result.

II [5 points]

1. In the study of the relation between yield per plant y and crop density x, the equation $y = \frac{1}{c+d\,x}$ is sometimes used. Show that this model correspond to assuming that the relative growth rate of $y(x)$ is proportional to $y(x)$. Which linearizing transformation allows this type of relation to be studied using a simple linear regression?

2. Consider a simple linear regression model, to be fitted with n pairs of observations.

(a) Describe the model.

(b) Deduce the probability distribution of the observations Y_i of the response variable.

(c) Knowing that $\frac{SQRE}{\sigma^2} \frown \chi^2_{n-2}$, deduce an unbiased estimator of the random errors' variance.

3. Consider a multiple linear regression in a descriptive context, with p predictors and fitted using n observations.

(a) What is the matrix model $\mathbf{X}$ and what is its column-space $\mathcal{C}(\mathbf{X})$?

(b) Show that the matrix of orthogonal projections onto $\mathcal{C}(\mathbf{X})$ is symmetric and idempotent.