INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2020-21

January 11, 2021 **FINAL EXAM - First call** **Duration: 3 hours**

Note: Justify your answers.

I [9 points]

A study of pastures sought to model the amount of dry matter (variable `Produtividade`, in kg/ha). Potential predictor variables of three kinds were available: chemical composition of soils, composition of the pastures and a large number of vegetation indices obtained from satellite images. Values of all these variables were obtained in 47 different locations in the Portuguese region of Alentejo.

1. A multiple linear regression model was fitted, with a logarithmic transformation of the response variable `Produtividade` (no transformation of the predictors). Here are some results:

```
Residual standard error: 0.3396 on 25 degrees of freedom
Multiple R-squared:  0.7199,  Adjusted R-squared:  0.4846
F-statistic:  3.06 on 21 and 25 DF,  p-value: ????
```

(a) How many predictors were used in this model? Comment on the validity of the following statement: *"this model explains nearly* 72% *of the variability of the observed productivities"*.

(b) Test the model's goodness-of-fit and comment your results.

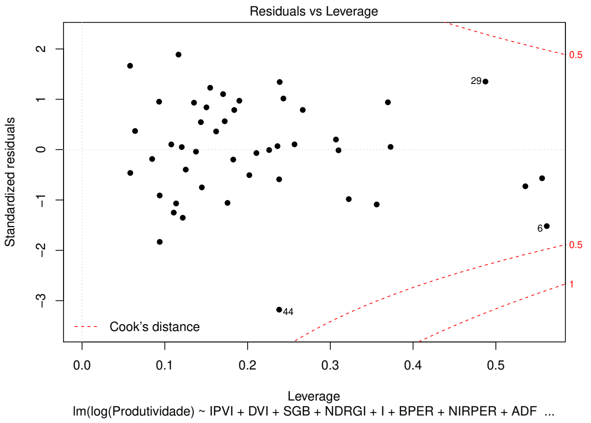
(c) A backward elimination algorithm produced a submodel with the following results.

```
Residual standard error: 0.2978 on 37 degrees of freedom
Multiple R-squared:  0.6814,  Adjusted R-squared:  0.6039
F-statistic: 8.791 on 9 and 37 DF,  p-value: 6.413e-07
```

i. How can we explain that, when compared with the full model, there is a considerable increase in the value of the adjusted R^2, while the standard R^2 decreases?

ii. Formally test whether this submodel is significantly different from the full model that was fitted initially.

iii. Describe and comment the following plot obtained from the fitted submodel. It is known that observation 44 has the largest Cook's distance. Based on the information provided by the graph, calculate an *approximate* value for this Cook's distance and comment it.



2. Here are the minimum, mean and maximum values for the variables used in the above submodel, as well as their standard deviations and their *coefficients of linear correlation with log-productivity*:

	IPVI	DVI	SGB	NDRGI	I	BPER	NIRPER	ADF	NDF	log(Prod)
min	0.55	20.65	-14.89	-0.03	164.89	0.25	0.45	366.37	568.52	6.087
mean	0.639	60.661	3.155	0.053	221.592	0.315	0.652	431.847	682.493	7.719
max	0.75	115.64	31.89	0.15	317.02	0.38	0.96	546.03	804.55	8.639
s	0.049	20.529	9.785	0.048	35.538	0.029	0.120	30.886	53.746	0.4731
r	0.6291	0.6552	-0.3048	-0.0072	-0.5461	0.301	0.6255	0.13	0.4409	1

(a) Identify the best predictor of log-productivity in a simple linear regression. Calculate the equation of the fitted regression line and the value of its coefficient of determination.

(b) Calculate the Residual Sum of Squares for the fitted regression.
Nota: If you were unable to compute this value, use the value 5.90 in what follows.

(c) Is it possible to state that the regression line's slope is significantly larger than zero? Answer using a Hypothesis Test.

(d) Compute a 95% confidence interval of the expected value of log-productivity, when the predictor's value is equal to its sample mean value.

(e) Deduce the non-linear relation between *productivity* (in kg/ha) and the predictor that you chose, that corresponds to the fitted regression line. What is the estimated value of the relative rate of change of productivity, as a function of your chosen predictor?

II [5 points]

Experiments were carried out with the Vital grape variety to determine whether its yield differs between sites and between fields where it is intended to set up vineyards. Three different sites were tested: Bombarral, Cadaval and Caldas da Rainha. In each site, the experiments used eight different fields which had different soil characteristics, both within and between sites. Each field was divided into seven plots, and in each plot the yield, in kg/ha, was observed. The mean yield in all 168 plots was 3.60162 kg/plant and the corresponding variance was 3.08414 (kg/plant)^2. The mean yield for all fields in Bombarral was 2.280, in Cadaval 4.361, and in Caldas da Rainha 4.164. The mean yields in each of the 24 fields were the following:

```
                            fields
Bombarral 1.823 2.080 2.270 2.694 2.030 2.696 2.331 2.313
Cadaval   5.772 3.994 3.718 4.617 5.357 4.048 3.639 3.742
Caldas    1.791 3.064 3.596 5.652 4.891 5.054 5.538 3.729
```

1. Which experimental design was used? Describe, in detail, the appropriate ANOVA model.

2. Knowing that the Residual Mean Square is 1.673 and that the computed value of the F test statistic for field effects is 3.592, complete the ANOVA model's summary table, indicating how you obtained each of the missing values.

3. Test whether there are significant site effects and discuss your conclusion.

4. Was it important to use different fields in each site? Answer, briefly justifying your answer using an appropriate hypothesis test.

5. The largest observed mean yield was for the first field in Cadaval. From what other field yields is this largest yield sigificantly different (at the $\alpha=0.05$ significance level)?

III [2 points]

A study was carried out to determine the viability of grafting in vines of the Azal variety, a variety of Portugal's Vinhos Verdes region. Three different rootstocks, called 1103P, R110 and 196/17, were used. The analyst initially grafted 1754 plants with the rootstock 1103P; 1787 with R110; and 1747 with 196/17. These plants were left to stratify during 3 weeks, after which they were planted in a greenhouse. After some time, the number of viable (commercializable) plants was counted. The results are given in the table below.

Rootstock	Viable plants	Non-viable plants	Total
1103P	1226	528	1754
R110	1295	492	1787
196/17	1357	390	1747
Total	3878	1410	5288

We wish to determine whether the performance of the rootstocks, as regards the viability of grafted plants, is similar.

1. What kind of Hypothesis Test should be used? In particular, describe the hypotheses.
2. Calculate the contribution to the final value of the test statistic of the non-viable plants with rootstock `196/17`.
3. Knowing that the computed value of the test statistic is $X^2_{calc} = 28.123$, what is your conclusion at the $\alpha = 0.05$ significance level?

IV [4 points]

1. Consider the one-way (one factor) ANOVA model, with k levels and n_i observations of the response variable Y in the factor's i-th level.
   (a) Characterize the model matrix, $\mathbf{X}$.
   (b) Show that for any vector in the column-space $\mathcal{C}(\mathbf{X})$, the vector elements corresponding to observations of a common factor level, all share the same value.
2. Consider a multiple linear regression with p predictors and fitted using n observations.
   (a) Write the multiple linear regression Model *using vector/matrix notation.*
   (b) Assuming the model, deduce the probability distribution of the random vector $\vec{\mathbf{Y}}$ of response variable observations.
3. Consider a generalization of the multiple linear regression model, in which the covariance matrix of the random error vector is a *generic* symmetric matrix Σ, which we assume known.
   (a) What is being assumed if we allow Σ to *not* be a diagonal matrix (that is, it may have non-zero off-diagonal elements)?
   (b) Knowing that the probability distribution of vector $\vec{\mathbf{Y}}$ is now $\mathcal{N}_n(\,\mathbf{X}\vec{\boldsymbol{\beta}}\,,\,\Sigma\,)$, deduce the probability distribution of the usual vector of estimators, $\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$. Are the estimators unbiased?