INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2020-21

July 12, 2021 **EXAM – Second Call** **Duration: 3h00**

Note: Conveniently justify your answers.

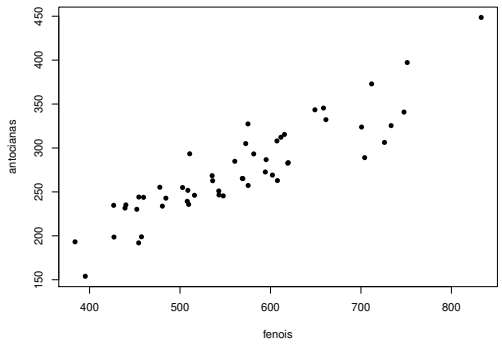
I [9 points]

A study seeks to model the anthocyanin content (in mg/l) of grape berries from the Moreto variety, a red variety grown in the Alentejo region of Portugal. On $n = 52$ genotypes, measurements were made of the mean values of the variable `antocianas` and of five potential predictors: phenol contents (`fenois`, in mg/l); `brix` levels (degrees brix); `pH`; berry weight (`pesobago`, in g); and acidity (`acidez`, in g/l of tartaric acid). Here are some indicators for the observations of each variable, as well as the corresponding matrix of correlations:

	`antocianas`	`pH`	`brix`	`fenois`	`acidez`	`pesobago`
Minimum	153.923	3.827	17.400	383.934	3.200	2.200
Mean	275.0065	3.9703	19.6769	564.3342	3.7683	2.4697
Maximum	448.615	4.103	21.333	832.909	4.650	2.951
Std. Deviation	53.6160	0.0593	0.6468	101.9808	0.2830	0.1357

	antocianas	pH	brix	fenois	acidez	pesobago
antocianas	1.00000	0.27786	0.65778	0.89735	-0.14286	-0.13566
pH	0.27786	1.00000	0.59785	0.32601	-0.49489	0.07244
brix	0.65778	0.59785	1.00000	0.60313	-0.22063	0.07903
fenois	0.89735	0.32601	0.60313	1.00000	-0.13315	-0.11272
acidez	-0.14286	-0.49489	-0.22063	-0.13315	1.00000	0.23733
pesobago	-0.13566	0.07244	0.07903	-0.11272	0.23733	1.00000

1. Below is given the scatterplot relating the variables `antocianas` and `fenois`.



   (a) Calculate the regression line of `antocianas` over `fenois`.

   (b) Discuss in detail the goodness-of-fit of the regression line.

   (c) Is it admissible to state that for each additional mg per litre in phenol contents, there corresponds, in the population, an average increase of 0.5 mg/l in the content of anthocyanins? Answer using a 95% confidence interval, and knowing that the estimate of the variance of the model's random errors is 571.186.

(d) Calculate the usual residual that corresponds to the observation on the upper right-hand corner of the scatterplot.

(e) What is the observation with the biggest leverage? Compute the value of that leverage.

2. An analyst considered that the above scatterplot has a curvature that would justify fitting a third-degree polynomial. Here are the results obtained:

```
> summary(lm(antocianas ~ fenois + I(fenois^2) + I(fenois^3) , data=moretoEx))
[...]
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.114e+02  4.214e+02  -1.925   0.0601
fenois       4.906e+00  2.196e+00   2.234   0.0302
I(fenois^2) -7.789e-03  3.738e-03  -2.084   0.0425
I(fenois^3)  4.445e-06  2.080e-06   2.138   0.0377
---
Residual standard error: 23.2 on 48 degrees of freedom
Multiple R-squared:  0.8238,  Adjusted R-squared:  0.8128
F-statistic: 74.81 on 3 and 48 DF,  p-value: < 2.2e-16
```

(a) Write the equation of the fitted curve.

(b) Formally test the hypotheses that this cubic model's goodness-of-fit is significantly better than that of the initial simple linear regression. Discuss your result.

3. A multiple linear regression was fitted, using all the available predictors, with these results:

```
Call: lm(formula = antocianas ~ . , data = moretoEx)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  245.50139  276.42909   0.888  0.37910
pH          -147.20363   75.49169  -1.950  0.05729
brix          23.78472    7.24378   3.283  0.00196
fenois         0.40251    0.03928  10.247 1.86e-13
acidez        -8.80027   13.46178  -0.654  0.51655
pesobago     -19.45681   24.83780  -0.783  0.43743
---
Residual standard error: 22.31 on 46 degrees of freedom
Multiple R-squared: 0.8439,  Adjusted R-squared: 0.8269
F-statistic: 49.72 on 5 and 46 DF,  p-value: < 2.2e-16
```

(a) Can it be said that, in the population, and with all other predictors remaining constant, an increase in acidity is associated with a decrease in anthocyanins levels? Answer using an appropriate hypothesis test and requiring the burden of the proof for the statement.

(b) A predictor can be excluded from the model without significantly affecting its goodness-of-fit. Identify the predictor whose exclusion would have the least effect on the goodness-of-fit. Justify your answer.

(c) Compute the value of the coefficient of determination for the submodel resulting from the exclusion of the predictor `fenois`. Comment.

II [5 points]

An experiment with the Malvasia grape variety was carried out in Fontanelas, with a view to comparing the yields of 9 different genotypes (called `MV1` to `MV9`). The goal was to select a genotype with a high yield, but which was systematically good in the meteorological conditions of different years. On each of 3 different years (2013, 2014 and 2017) the yields for each genotype were recorded on 5 plots chosen at random. The variance of the yields in the 135 plots was 2.012446 $(\text{kg/plant})^2$. The overall mean and the means for each year, genotype and year/genotype combination, are shown below.

```
Tables of means
Grand mean     ano                  genotipo
2.854985       A2013 A2014 A2017    MV1   MV2   MV3   MV4   MV5   MV6   MV7   MV8   MV9
               1.883 2.175 4.508    2.444 2.508 2.918 2.948 2.886 3.173 3.413 2.795 2.609

 ano:genotipo
       genotipo
ano      MV1   MV2   MV3   MV4   MV5   MV6   MV7   MV8   MV9
  A2013 1.178 1.920 1.736 2.189 2.086 2.108 2.416 1.797 1.513
  A2014 2.534 1.564 2.217 1.725 2.003 2.588 2.638 2.007 2.295
  A2017 3.621 4.041 4.800 4.931 4.570 4.822 5.186 4.580 4.020
```

1. Identify the experimental design and describe in detail the ANOVA model suited to the experiment.

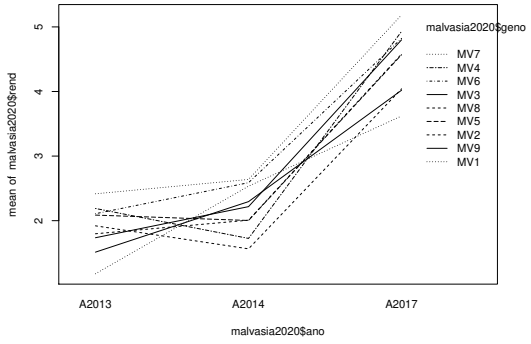
2. Complete the following ANOVA table, indicating how each of the missing values indicated by a question mark is obtained.

```
              Df Sum Sq Mean Sq F value   Pr(>F)
ano           ??? 186.33   93.16    ???   <2e-16
genotipo      ???  11.69    ???    2.566    ---
ano:genotipo ???  10.17    0.64   1.117  0.3490
Residuals     ???   ???     0.57
```

3. Which type of effects should be considered significant? In case you need to carry out more than one test, describe one test in detail and the remaining test(s) more briefly.

4. Compare the yields of genotype `MV7` for the experiment's three years. State which differences should be considered significant. **Note:** The 0.95 quantile in the relevant distribution is 5.369266.

5. Describe and discuss the following plot.



III [2 points]

The genetic mechanisms determining two traits of maize kernels are being studied. The endosperm may be sweet (a recessive trait) or amilaceous (dominant); and the colour of the aleurone (a protein in the endosperm) may be purplish (dominant) or white (recessive). Assuming that each of these traits is governed by a single gene, with independent segregation, it would be expected that in the second generation of the cross between a pure line of maize with purplish aleurone kernels and amilaceous endosperm, with another pure line of maize with white aleurone and sweet endosperm, would result in 9/16 of plants with both dominant traits; 1/16 with both recessive traits; 3/16 of kernels with white aleurone and amilaceous endosperm; and 3/16 of kernels with purplish aleurone and sweet endosperm. An experiment crossing two pure lines as described above, resulted in the second generation counts indicated in the table. Test whether the data are compatible with the genetic hypothesis that was described ($\alpha = 0.05$). Comment your conclusions and, in case of rejection of the Null Hypothesis, discuss the reasons for that rejection.

	Endosperm	
Aleurone	amilaceous (dominant)	sweet (recessive)
purplish (dominant)	248	56
white (recessive)	56	48

IV [4 points]

Consider a multiple linear regression of variable Y over p predictors, fitted using n observations.

1. Describe the right triangle in the space of variables ($\mathbb{R}^n$) that is *directly* related to the fundamental formula of linear regressions. In that triangle, which *geometric* concept corresponds to the ratio between the proportion of Y's variability that can, and that cannot, be accounted for by the regression?

2. Now consider a submodel with k predictors.

   (a) Show that the submodel's *adjusted* R^2 is bigger than that of the full model if and only if the estimate of the random errors' variance is smaller in the submodel than in the full model.

   (b) Show that the inequality $QMRE_c > QMRE_s$ (where the indices c and s denote, respectively, the full model and the submodel) is equivalent to having a value less than 1 in the partial F test statistic comparing these two models.

   (c) Discuss the implication of the conditions in both previous questions for a backward elimination algorithm based on *Student's* t tests.