INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2020-21)

Final Exam

June 29, 2021 Duration: 3h30

I [11.5 points]

The weight of bunches of grapes is an important variable in viticulture, that is closely tied to total production. Its accurate measurement is a time-consuming and destructive operation, which requires that the bunches be picked from the vineyard. A study sought to model bunch weight (variable `Bw`, in g) from variables that can be observed in 2-dimensional images taken by robots that go into vineyards, namely, the number of berries that are visible in an image (count variable `BEv`) and the area of each bunch on its image (variable `Ba`, in cm^2). The dataset used to fit the models had 375 observations, 75 of which from each of 5 varieties: Alvarinho, Cabernet, Syrah, Touriga and Viosinho.

Here are the summary indicators and correlations for the entire dataset:

```
> summary(Todos[,c("BEv","Ba","Bw")])
      BEv              Ba                Bw
Min.   : 8.0   Min.   : 10.60   Min.   : 10.6
1st Qu.:34.0   1st Qu.: 54.52   1st Qu.: 86.0
Median :44.0   Median : 74.12   Median :133.6
Mean   :44.7   Mean   : 74.44   Mean   :137.2
3rd Qu.:55.0   3rd Qu.: 90.67   3rd Qu.:174.8
Max.   :83.0   Max.   :154.62   Max.   :351.0
```

```
> cor(Todos[,c("Bw","BEv","Ba")])
           Bw       BEv        Ba
Bw  1.0000000 0.8627126 0.9167313
BEv 0.8627126 1.0000000 0.8885402
Ba  0.9167313 0.8885402 1.0000000
```

1. Consider the data relative to the bunch weight (`Bw`, in g) *for the 3 varieties: Alvarinho, Syrah and Viosinho*. Consider the output that is in the Appendix I in which some commands of `R` were performed. Whenever possible use the results to answer the following questions:

   (a) Sketch the histogram of `Bw.Vios`, indicated in the output but not plotted.

   (b) Write the necessary calculations for plotting the boxplot of the variable `Bw.Vios`. Draw it please, clearly marking the boxplot limits.

   (c) Given the results presented in the output, can we say that the variable `Bw.Sy` has, on average, smaller values than the `Bw.Vios` variable? Justify properly.

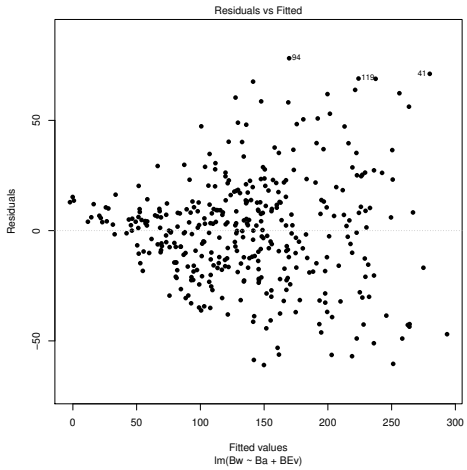
   (d) It is assumed that the weight of variable Alvarinho can be modeled by the gamma distribution. For simplicity consider that only one parameter, $\mu > 0$, is unknown being the density function defined as:

   $$f(x|\mu) = \frac{4}{\mu^2}\, x\, \mathrm{e}^{-\frac{2x}{\mu}}, \quad \text{if } x > 0; \quad 0 \quad \text{if } x \leq 0.$$

   Given a random sample $(X_1, X_2, ..., X_n)$ extracted from that variable, obtain the maximum likelihood estimator of μ.

2. Which potential predictor, `Ba` or `BEv`, would provide the best simple linear regression for the response variable `Bw`, considering all 375 observations? Discuss the resulting goodness-of-fit of the model you chose.

3. Below is a scatterplot of residuals versus fitted values, for the multiple linear regression of `Bw` over both the other variables, fitted using all 375 observations. Describe and discuss the plot and its relevance for the fitted linear model.



4. A multiple linear regression model was fitted *with log-transformations of all three variables*, to predict `log(Bw)` from the other two variables. Below is the resulting output for this model:

```
Call: lm(formula = log(Bw) ~ log(Ba) + log(BEv), data = Todos)
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.40444    0.08839  -4.575 6.49e-06
log(Ba)      0.93918    0.05197  18.072  < 2e-16
log(BEv)     0.32806    0.05762   5.694 2.52e-08
---
Residual standard error: 0.1745 on 372 degrees of freedom
Multiple R-squared: 0.9039,  Adjusted R-squared: 0.9034
F-statistic:  1749 on 2 and 372 DF,  p-value: < 2.2e-16
```

   (a) Write the fitted non-linear equation relating the three original (not log-transformed) variables.

   (b) Build a 95% confidence interval for the coefficient of the log-transformed number of visible berries. Interpret your results, in terms of both the log-transformed and the original variables.

5. A simple linear regression of log(`Bw`) over log(`Ba`), was initially fitted with the entire dataset, resulting in a coefficient of determination $R^2 = 0.8955$. An ANCOVA model was then fitted, allowing for different simple linear regressions of `log(Bw)` over `log(Ba)` in each of the five varieties. Here is the summary output for this model, as well as a submatrix of the (co-)variance matrix of some parameter estimators:

```
> summary(Todos.anc)
Call: lm(formula = log(Bw) ~ log(Ba) * Casta, data = Todos)
[...]
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.79050    0.22287  -3.547 0.000441
log(Ba)                1.35726    0.05363  25.307  < 2e-16
CastaCabernet          0.72740    0.26383   2.757 0.006126
CastaSyrah             0.23466    0.28791   0.815 0.415588
CastaTouriga           0.09209    0.30806   0.299 0.765149
CastaViosinho          0.87185    0.32956   2.646 0.008509
log(Ba):CastaCabernet -0.22688    0.06340  -3.578 0.000392
log(Ba):CastaSyrah    -0.09970    0.06814  -1.463 0.144306
log(Ba):CastaTouriga  -0.06954    0.07441  -0.935 0.350663
log(Ba):CastaViosinho -0.23833    0.07651  -3.115 0.001985
---
Residual standard error: 0.1624 on 365 degrees of freedom
Multiple R-squared:  0.9183,   Adjusted R-squared:  0.9163
F-statistic: 455.9 on 9 and 365 DF,  p-value: < 2.2e-16

> vcov(Todos.anc)[7:8,7:8]
                      log(Ba):CastaCabernet log(Ba):CastaSyrah
```

```
log(Ba):CastaCabernet           0.004019790        0.002876272
log(Ba):CastaSyrah              0.002876272        0.004643707
```

(a) Write the fitted equation for the Cabernet variety, in both the log-transformed and the original units.

(b) Can the Cabernet and Syrah population regression lines relating the log-transformed variables be considered parallel? Provide a formal justification.

(c) Formally test whether this ANCOVA model provides a significantly better fit than the simple linear regression model with a single regression line for the entire dataset. Discuss your result.

6. Taking into consideration the plot in question 3, a researcher suggested fitting a Generalized Linear Model of `Bw` over both other variables, *without log-transformations*, assuming a Gamma distribution for the random component and an identity link function. Here are some results:

```
Call: glm(formula = Bw ~ Ba + BEv, family = Gamma(link = identity),  data = Todos)
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.58838    1.19755 -11.347  < 2e-16
Ba            1.54400    0.09506  16.242  < 2e-16
BEv           0.75675    0.16525   4.579 6.37e-06
---
(Dispersion parameter for Gamma family taken to be 0.03177647)
    Null deviance: 99.125  on 374  degrees of freedom
Residual deviance: 12.079  on 372  degrees of freedom
AIC: 3377.9
```

(a) Why is the Gamma distribution a plausible choice for the random component?

(b) Describe the fitted model and discuss its quality.

II [3 points]

In a grapevine selection study, 32 clones of the variety Vital were evaluated in 3 locations (Bombarral, Cadaval, and Caldas da Rainha). In each location a field trial with the 32 clones was planted according to a completely randomized design with 8 repetitions (that is, in each location there are 8 observations of each one of the clones). Consider the `location` as a fixed effects factor and the `clone` as a random effects factor (i.e., admit that the clones studied constitute a sample of the possible best clones of the Vital variety).

a) Describe in detail the adequate model for the study described above.

In R, with the function `lmer` from the package `lme4`, the following commands were executed:

```
> library(lme4)
> vital<-read.table("vital.txt", header=T)
> vitallmer1<-lmer(rend~local+(1|clone)+(1|local:clone), data=vital)
> summary(vitallmer1)
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: rend ~ local + (1 | clone) + (1 | local:clone)
   Data: vital
Random effects:
 Groups      Name        Variance Std.Dev.
 local:clone (Intercept) 0.07685  0.2772
 clone       (Intercept) 0.33689  0.5804
 Residual                1.74037  1.3192
Number of obs: 768, groups:  local:clone, 96; clone, 32
Fixed effects:
             Estimate Std. Error      df t value Pr(>|t|)
(Intercept)    1.9808     0.1405 59.2485   14.10   <2e-16 ***
localCadaval   1.8447     0.1356 62.0006   13.60   <2e-16 ***
localCaldas    1.5651     0.1356 62.0006   11.54   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> logLik(vitallmer1)
'log Lik.' -1342.896
> vitallmer2<-lmer(rend~local+(1|clone), data=vital)
> logLik(vitallmer2)
'log Lik.' -1344.309
> vitallmer3<-lmer(rend~local+(1|local:clone), data=vital)
> logLik(vitallmer3)
'log Lik.' -1355.286
```

b) Test the variance components associated to the model defined above. Describe in detail only one of the hypothesis tests performed.

c) For the full fitted model, what is the value of the Bayesian Information Criterion?

III [2.5 points]

Consider 5 continuous variables, `pl_orbper`, `pl_orbsmax`, `st_logg`, `pl_bmass` e `sy_dist`, regarding parameter estimates of 1177 planets detected beyond our solar system. The `pl_orbper` corresponds to the orbital period (in days) around the corresponding star, the variable `pl_orbsmax` is the length of the orbital largest semi-axis in astronomic units (an astronomic unit (au) is approximately equal to the average of the distances between the planet Earth and the Sun), the variable `st_logg` is the logarithm in base 10, of the gravity acceleration in $\texttt{cm/s}^2$, the variable `pl_bmass` is the planet mass, measured in Jupiter masses and the variable `sy_dist` represents the distance to planetary system in `parsecs` (a parsec (pc) corresponds approximately to 3.26 light-years). The data was retrieved from Nasa's archive on Exoplanets exploration

`https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS&constraint=default_flag=1`

From the above data was obtained the corresponding correlation matrix (rounded to the 3 decimal places),

	pl_orbper	pl_orbsmax	st_logg	pl_bmasse	sy_dist
pl_orbper	1.000	0.936	-0.004	0.167	-0.073
pl_orbsmax	0.936	1.000	-0.094	0.283	-0.139
st_logg	0.004	-0.094	1.000	-0.412	-0.084
pl_bmasse	0.167	0.283	-0.412	1.000	0.080
sy_dist	-0.073	-0.139	-0.084	0.080	1.000

A clustering analysis was performed on the standardized data set of the 1177 planets using the hierarchical method of the complete linkage and the euclidean distance. From this clustering analysis a partition into 2 clusters was obtained and it turned out that one of the clusters contained only a single planet. Posteriorly the partition into 2 groups was compared with the partition into 3 clusters that is obtained from the same hierarchical clustering analysis using the Rand index.

1. Determine the number of elements of each cluster of the partition into 3 clusters knowing that the Rand index is equal to 0.9881762.
2. Perform a clustering analysis of the set of the 5 variables using the Ward's hierarchical method and a convenient dissimilarity that does not account for the sign of the correlation between the variables. Interpret the result.

IV [3 points]

Let $\mathbf{X}^c$ be an $n \times p$ column-centred data matrix and consider the Singular Value Decomposition $\frac{1}{\sqrt{n-1}}\mathbf{X}^c = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t = \sum_{i=1}^{p} \delta_i \vec{\mathbf{w}}_i \vec{\mathbf{v}}_i^t$, where $\vec{\mathbf{w}}_i$ and $\vec{\mathbf{v}}_i$ are the columns of, respectively, $\mathbf{W}$ and $\mathbf{V}$, and δ_i the corresponding diagonal elements of $\mathbf{\Delta}$.

1. Show that the matrix of orthogonal projections onto the column-space of $\mathbf{X}^{\mathrm{c}}$, $\mathbf{P}_{\mathbf{X}^{\mathrm{c}}} = \mathbf{X}^{\mathrm{c}}(\mathbf{X}^{\mathrm{c}t}\mathbf{X}^{\mathrm{c}})^{-1}\mathbf{X}^{\mathrm{c}t}$, is the same as the matrix of orthogonal projections onto the column-space of matrix $\mathbf{W}$. Interpret this result in terms of the Principal Component Analysis of the data associated with matrix $\mathbf{X}^{\mathrm{c}}$.

2. Let $\mathbf{W}_k$ be the $n \times k$ submatrix defined by the first k columns of matrix $\mathbf{W}$ (associated with the k largest singular values). Knowing that the matrix of orthogonal projections onto its column-space is $\mathbf{P}_k = \mathbf{W}_k\mathbf{W}_k^t = \sum_{j=1}^{k} \vec{\mathbf{w}}_j\vec{\mathbf{w}}_j^t$, show that matrix $\mathbf{P}_k\mathbf{X}^{\mathrm{c}}$ solves the Eckart-Young problem for matrix $\mathbf{X}^{\mathrm{c}}$. Interpret that result in terms of Principal Components.