INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2020-21)

Test – Generalised Linear Models and Mixed Linear Models

May 31, 2021 Duration: 2h30

I [9 points]

A study seeks to estimate the number of berries in bunches of grapes (a count variable `BE`) based on three other variables: the bunch weight (variable `Bw`, in g) and two 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 model had observations on 75 bunches of each of 5 varieties, for a total of 375 observations, but since the goal was a model that could be applied to any variety, the observations were considered in their entirety.

Here are some summary indicators:

```
> summary(Todos[,c("BE","BEv","Ba","Bw")])
       BE              BEv              Ba               Bw
 Min.   :  8.0   Min.   : 8.0   Min.   : 10.60   Min.   : 10.6
 1st Qu.: 61.0   1st Qu.:34.0   1st Qu.: 54.52   1st Qu.: 86.0
 Median : 85.0   Median :44.0   Median : 74.12   Median :133.6
 Mean   : 87.7   Mean   :44.7   Mean   : 74.44   Mean   :137.2
 3rd Qu.:113.5   3rd Qu.:55.0   3rd Qu.: 90.67   3rd Qu.:174.8
 Max.   :218.0   Max.   :83.0   Max.   :154.62   Max.   :351.0
```

1. Given the nature of the random component `BE`, what probability distribution (among those considered in class) do you consider most appropriate? Justify your answer.

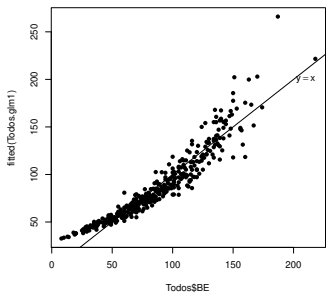
2. Regardless of your reply to the previous question, two Generalised Linear Models with a Poisson response variable were fitted, that differed in their link function. Here are the results:

```
> summary(Todos.glm1)
Call: glm(formula = BE ~ BEv + Bw + Ba,
    family = poisson(link = log),  data = Todos)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.3338620  0.0201937 165.094   <2e-16
BEv          0.0166706  0.0007675  21.721   <2e-16
Bw           0.0029365  0.0001973  14.881   <2e-16
Ba          -0.0011815  0.0005132  -2.302   0.0213
---
    Null deviance: 5970.14  on 374  degrees of freedom
Residual deviance:  676.76  on 371  degrees of freedom
AIC: 3012.5
```

```
> summary(Todos.glm2)
Call: glm(formula = BE ~ BEv + Bw + Ba,
    family = poisson(link = identity),  data = Todos)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.07425    1.14234  -3.567 0.000362
BEv          1.38669    0.07167  19.348  < 2e-16
Bw           0.34403    0.02071  16.613  < 2e-16
Ba          -0.23393    0.05079  -4.606  4.1e-06
---
    Null deviance: 5970.14  on 374  degrees of freedom
Residual deviance:  267.95  on 371  degrees of freedom
AIC: 2603.7
```

   (a) Describe in detail the model that was fitted on the left (model `Todos.glm1`).

   (b) Below is the scatterplot of berries per bunch (horizontal axis) and corresponding values fitted by the model `Todos.glm1` (model on the left), together with the $y = x$ line. Comment.



(c) Indicate the mean number of berries that the model on the right (model `Todos.glm2`) would associate to a bunch that weighted 20 g and whose image had an area of 15 cm^2 and 10 visible berries. Comment this value, also taking into account that the corresponding value fitted by the other model is 34.521.

(d) Which of these two models would you choose, based on the available information? Justify your answer.

(e) Consider a modification to the model `Todos.glm2` (on the right): assume that the distribution of the random component is Normal. Comment that model. How would it be possible to compare its results with those of model `Todos.glm2`?

3. The above models include a predictor whose measurement requires a manual weighting of the bunches (`Bw`). Seeking a model whose systematic component only involves measurements that can be made on images that are automatically collected, a Poisson model was fitted, with an identity link function, but only two predictors: `BEv` and `Ba`. The resulting residual deviance was 547.3. Perform a Likelihood Ratio Test to determine whether this new model's goodness-of-fit is significantly worse than that of the corresponding three-predictor model. Comment.