Facultat de Matemàtiques i Estadística, UPC

Linear and Generalized Linear Models

Practical Exercises

Jordi Valero and Marta Pérez-Casany

School year 2018-2019

Contents

1 Linear Models

1.1 Linear Regression

1) It is well known that the excess of weight is one of the factors that has a negative influence in the cholesterol level of human beings. In an experiment with children from 9 to 20 years old, measures of their cholesterol level (C), the weight (W) the height (H) and the age (A) have been recorded. The data appear in file COL.csv.

a) Compute the regression line for modeling the cholesterol as a function of weight.

b) Plot the regression line jointly with the confidence intervals and the prediction intervals.

c) Perform the appropriate plots to check:

- Tendency and homogeneity of variances. Plot the residuals as a function of the predicted values.
- Outliers. Plot the studentized residuals as a function of any of the predictors, observation number ... jointly with the horizontal lines at -2 and 2.
- Influence values. Plot the dffits as a function of any of the following variables: predicted values, observation number ...; jointly with the horizontal lines at $\pm 2\sqrt{\frac{p}{n}}$.

d) Interpret the results. Is there any contradiction?
Hint: Perform the dispersion plot of C as a function of W, jointly with the regression lines for ages. The R instruction is:

```
COL$GA<-factor(COL$A)
scatterplot(C~W|GA,smooth=F,col=1:20,data=COL)
```

2) The file REG8.csv contains simulated data that allow to compute 8 regression lines. The first column, REG (numbers from 1 to 8), indicates the data associated to each line, the rest of columns correspond to X and Y.

Compute the 8 regression lines and compare them. Answer the following questions:

a) Do they have the same regression coefficients?

b) Do they have the same coefficient of determination?

c) Do they give the same conclusion in the ANOVA test?

d) Are all the lines reasonable? Which problems do you observe?

3) With the same data set that in exercise 1, perform a linear regression to model the cholesterol level as a function of weight, height and age. That is that one assumes that

$$C_i = \beta_0 + \beta_1 W_i + \beta_2 A_i + \beta_4 H_i + \varepsilon_i,$$

with $\varepsilon_i \sim N\left(0, \sigma^2\right)$. In matrix notation this is equal to: $Y = X \cdot \beta + \varepsilon$.

Perform the ANOVA analysis assuming that the corresponding assumptions are verified, and do what is required in the following points:

a) For each regression parameter compute its punctual and confidence interval estimations (95%) and perform the test $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$.

b) Punctual estimation of σ^2.

c) Define the hypothesis corresponding to the "omnibus" test and perform the contrast.

d) Define the hypothesis of the ANOVA test, using the sums of squares of type I and type III. Compare the results with the ones obtained in section a).

e) For $W = 65,\ A = 15\ and\ H = 150$

- Compute the confidence interval $IC_{95\%}\left(E\left[C|_{W=65,\ A=15\ and\ H=150}\right]\right)$
- Compute the predicted interval (95%) for cholesterol
- Perform the test to know if $E\left[C|_{W=65,\ A=15\ and\ H=150}\right] = 190$ or not.

f) Plot and interpret the following figures

- Residuals vs predicted.
- Standardized residuals vs predicted.
- Hat values.
- Cook's distances.
- Studentized residuals vs predicted.
- dffits values.

g) For each explanatory variable compute the variance inflation factor, VIF, using the tolerance.

h) Assume that $W_0 = -10 + 0.5H$ is a pattern of weight as a function of height, and that the excess of weight is computed as: $EW = W - W_0$. Compute the regression

$$C_i = \beta_0' + \beta_1' EW_i + \beta_2' A_i + \beta_4' H_i + \varepsilon_i.$$

Which of the results are different from the ones obtained before?

i) Analyze how the regression parameters change in both models and also the rest of the calculus if the independent variables are previously centered at zero. That is if one considers

as explanatory variables $X - \overline{X}$, or $X - X_c$ with $X_c \simeq \overline{X}$. In this case $W_c = 65$, $A_c = 15$ *and* $H_c = 150$.

4) An enterprise in charge of the woodland, of a given number of cities from the *Metropolitan area of Barcelona*, regularly controls the diameter of the crown of the lawnmowers (special kind of tree). The measures are denoted as DCrown. Given that to measure the DCrown is usually quite expensive, they are interested in obtaining a way to estimate this measure from other data that are more simple to obtain. To that end, they consider that interesting measures could be the perimeter of the trunk at 1,5m of height (PT), the perimeter in the base (PB), the trunk's height (HT) and the age of the tree (A) in years. The perimeters and the height are measured in meters.

With this objective, they have recorded those variables in a random sample of 311 trees. The results are contained in the file "`dcrown.csv`". It is important to observe that, given that the base perimeter is larger than the crown perimeter and they are highly correlated, to avoid collinearity problems instead of the basal perimeter, the ratio of the base and the truck perimeters, *RP=PB/PT*, will be considered as an indicator of the shape of the trunk.

The following two models have been considered:

modA: The linear regression model of the random variable $DCopa$ with explanatory variables PT, RP, HT and A.

modB: The linear regression model of the random variable $\log(DCrown)$ with explanatory variables $\log(PT)$, $\log(RP)$, $\log(HT)$ and $\log(A)$.

a) Assuming that the hypothesis of the linear model are verified, fit *modA* and answer the following questions:

(1) Are significantly different from zero the four coefficients for the explanatory variables with a significance level of 5%?

(2) Which is the estimation of the residual variance?

(3) Looking at the plot of the studentized residuals, how many observations do not belong to the interval $(-2, 2)$?
Which percentage do they represent? Is it reasonable this percentage?

b) Assuming that the hypothesis of the linear model are verified, fit *modB* and answer the following questions:

(1) Are significantly different from zero the four coefficients for the explanatory variables with a significance level of 5%?

(2) Which is the estimation of the residual variance?

(3) Looking at the plot of the studentized residuals, how many observations do not belong to the interval $(-2, 2)$?
Which percentage do they represent? Is it reasonable this percentage?

c) Choose between *modA* and *modB*, which of the two is the best? Explain why?

d) With the elected model and for the trees that have 10 years old with $PT = 0.4, PB = 0.6, and HT = 2.3$, compute the prediction of *DCrown* and the confidence interval for the prediction with a confidence level of 95%.

5) Using a program able to deal with matrices (R, Excel, etc..) compute in a matrix calculus the following statistics and check that the results are the same as the ones obtained with `R` in exercise 3):

- $\widehat{\beta}$, $\widehat{y}$ and $\widehat{\varepsilon}$.
- *Model Df, Error Df, Model Sum Sq, Error Sum Sp, Model Mean Sq, Error Mean Sq,* F_{value} and p_{value}.
- $Var\left(\widehat{\beta}\right)$, $IC_{95\%}(\beta)$, $Var(\widehat{y})$ and $Var(\widehat{\varepsilon})$.
- Assuming that the ANOVA hypothesis are verified,
  - For each β_i compute $IC_{95\%}(\beta_i)$

  perform the hypothesis test at (5%):
  - Anova for the regression.
  - For each β_i, test if $\beta_i = 0$ or not.

  for $W = 65,\ A = 15\ y\ H = 150$
  - Compute the confidence interval $IC_{95\%}$ for $E[C]$ when $W = 65,\ A = 15\ and\ H = 150$
  - Compute the predicted interval $PI_{95\%}$ for cholesterol at the same conditions as before.
  - Perform the test to know if $E[C] = 190$ or not, when $W = 65,\ A = 15\ and\ H = 150$.

1.2 One way Anova

6) To see if the dose of a sweetener improves the fattering of piglets, one experiment was performed. A set of piglets with similar conditions where selected and 5 different sweetener doses where considered and randomly assigned to the piglets. The response variable is the average daily gain, ADG, and the explanatory variable is the sweetener dose:

a) Define and fit the linear model appropriate to this situation.

b) Perform the ANOVA test and obtain conclusions from it.

c) Using the Tukey method at 5%,

- Which levels differ in their expected ADG? (*emmeans*)
- Explain the results obtained before in a compact form. (*CLD*)

d) Check if the linear model assumptions can be assumed to be true by performing the model adequacy checking.

7) An experiment in the markets of the city of Barcelona has been realized, in order to compare the origin of the sole fish. Four different origins have been considered and denoted by A, B, C and D. For each sole, the following variables have been recorded: Quality (and index that summarizes several variables), cephalic-somatic index (CSI), pH, Humidity and Proteins. The experimental data are in the file "Sole.csv".

a) With respect to the variable quality (Quality):

(1) Define an appropriate linear model to describe it as a function of other variables and fit it.

(2) Perform the analysis of variance test.

(3) Using the Tukey method (5%), investigate in which levels do really exists differences in the quality. Express the results in a compact way.

(4) Are the conditions of the linear model verified? Perform the model adequacy checking.

b) Do the same analysis than before, taking as response variable each one of the following variables:

- CSI
- pH
- Humidity
- Proteins

8) With the data in exercise 6),

a) Write down the matrix form of the following linear models:

- $y_{ij} = \mu_i + \varepsilon_{ij}$
- $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ con $\sum_{i=1}^{k} \alpha_i = 0$
- $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ con $\alpha_1 = 0$

b) For each one of the models of the last section, compare the parameter estimations using the Tukey method.

1.3 Two way and more than Two way ANOVA

9) In a pate tasting, each person involved has to taste and give punctuation to 5 different pates, which were randomly assigned. The file PATE.csv contains the punctuations. The columns correspond to:

- the person that give the punctuation
- pate code
- punctuations of: color, smell, texture, taste and order. The marks were between 0 and 10, and the order is the preferred order of the taster where 1 means the preferred by the taster.

For each one of the variables that have been scored,

a) Find the appropriate linear model assuming that conditions for the ANOVA are satisfied. Perform the corresponding tests and interpret them, including multiple comparisons. Is it necessary to apply any kind of data transformation?.

b) Compare the results with the ones obtained considering that it may also exists a taster effect.

c) Are the ANOVA conditions verified?. And the conditions for the Friedman test, are they verified?. Justify it theoretically and by means of diagnosis.

10) For the cheese producers, the rendibility is the relation between the weight of the cheese and the amount of milk used to produce it. One is interested in analyzing if the type of milk used changes the rendibility. Three milk types are considered: cow, sheep and goat. For each milk type, one is interested in comparing the rendibility as a function of the thermic milk treatment (plane or pasteurized) and the presence of a given additive denoted by CaCl2. The data are in the cheese.csv file.

a) Assuming that the ANOVA conditions are satisfied, for each type of milk answer the following questions and justify your answer:

- Which treatment is the best?
- The addition of CaCl2 does it increase the rendibility?
- Does the thermic treatment increase the rendibility?

b) Do you think that the ANOVA hypothesis are verified? Answer this question theoretically and by means of diagnosis techniques.

11) For each table separately, compute the marginals means table. Also compute the *"emmeans"* and answer:

Table 1		Factor 1	
Data		A	B
Factor 2	1	18 19 21 22	9 11
	2	19 21	8 9 11 12

Table 2		Factor 1	
Data		A	B
Factor 2	1	18 19 21 22	9 11
	2	9 11	19 21

Table 3		Factor 1	
Data		A	B
Factor 2	1	18 19 21 22	10 12
	2	22 24	12 13 15 16

a) Looking at the marginal means

- Does it exists any difference between the levels of factor 1?
- Does it exists any difference between the levels of factor 2?

b) Looking at the ***emmeans*** computed by hand

- Does it exists any difference between the levels of factor 1?
- Does it exists any difference between the levels of factor 2?

c) Performing the ANOVA test with two factors and looking at the SS1 ($\alpha = 5\%$)

- Does it exists any difference between the levels of factor 1?
- Does it exists any difference between the levels of factor 2?

d) Performing the ANOVA test with two factors and looking at the SS3 ($\alpha = 5\%$)

- Does it exists any difference between the levels of factor 1?
- Does it exists any difference between the levels of factor 2?

e) Using the multiple comparisons (emmeans) ($\alpha = 5\%$)

- Does it exists any difference between the levels of factor 1?
- Does it exists any difference between the levels of factor 2?

12) The file PRAC2F.csv has 8 columns. The first two are two factors with 3 and 4 levels respectively. The last 6 contain the experimental results (simulated) of 6 random variables. For each one of the variables perform the following tasks:

a) Explain a real situation (specifying the name of the variables that appear in the experiment and the procedure to obtain the data) that matches with the given file. Using the two-way ANOVA analysis with ($\alpha = 0.05$) perform:

- the descriptive statistics obtaining the means table

$F1 \diagdown F2$	1	2	3	4	Total
1					
2					
3					
Total					

- the means and standard errors diagram, $\overline{X}$ (x-axis) and $S_{\overline{X}}$ (y-axis) .
- the ANOVA test at 5%.
- the separated means of each one of the factors and treatments. It is convenient to write the separated means in the means table.
- the dispersion diagram of residuals vs predicted to see if there is any tendency and to check if there is homocedasticity.

b) With the results obtained, answer the following questions:

- Is there any significant effect? What can be deduced from the means table?
- Which is the error variance estimation?
- Does the first factor have any effect? How is it detected in the means table and in the $\overline{X} - S_{\overline{X}}$ diagram?
- Does the second factor have any effect? How is it detected in the means table and in the $\overline{X} - S_{\overline{X}}$ diagram?
- Does it exists interaction? How is it detected in the means table and in the $\overline{X} - S_{\overline{X}}$ diagram?
- Which is (are) the best treatment?
- Which is (are) the best level of the first factor?
- Which is (are) the best level of the second factor?
- The best treatment, is it the combination of the best levels of each factor? Why?
- If the first level of the first factor needs to be used, which treatment is in your opinion the best?
- If the second level of the first factor needs to be used, which treatment is in your opinion the best?
- If the third level of the first factor needs to be used, which treatment is in your opinion the best?
- If the first level of the second factor needs to be used, which treatment is in your opinion the best?
- If the second level of the second factor needs to be used, which treatment is in your opinion the best?
- If the third level of the second factor needs to be used, which treatment is in your opinion the best?
- If the fourth level of the second factor needs to be used, which treatment is in your opinion the best?
- Do you see any anomalous tendency in the residuals?
- Do you see any tendency in the residual variance?

c) Using the two-way ANOVA without interaction perform the same analysis as in section a), and answer the questions of section b) that keep having sense. Do you think that the two-way ANOVA is an appropriate model for these data?

d) Do the same as before assuming that the levels of $F2$ change by changing the levels of $F1$, which is known as a *nested* design. The R notation for doing it is $(F1 + F1 : F2)$.

13) One wants to know if the flour degradation (W2) depends on: 1) the wheat type and 2) the presence/absence of insects in the flour. The experimental data appear in the file wheat.csv (the file wheat4.csv, contains the same data but only for four varieties), their columns are: the variety (VAR), the degradation level (W2) and the presence/absence of insects (PRES). One characteristics of the data set is that it is largely unbalanced.

a) Assuming that the ANOVA hypothesis are verified, define a factorial model taking into account the factors, variety and presence of insects. Based on the analysis answer the following questions justifying the response:

- Is the flour degradation affected by the wheat type and the presence of insects?
- Is the flour degradation affected by the wheat type?
- Is the flour degradation affected by the presence of insects?
- Compare the results obtained by considering the type I and type III sums of squares. Does the factors order have any influence in the results?
- In this case, which is the role played by the interaction? Could we avoid the interaction term?

b) Do you think that the ANOVA hypothesis are verified? Justify your answer theoretically and by means of plots.

c) In the case of the two factor additive model, repeat the sections a) and b) with the exception of the ones that do not keep having sense.

1.4 Analysis of covariance (ANCOVA)

14) The objective is to compare the punctuations (p) of two pedagogical methodologies (m) as a function of the coefficient of intelligence (c) with the data contained in the file **"comp line.csv"**. Perform the parameter estimation as well as the hypothesis tests to see if the parameters are significantly different from zero or not. Moreover, perform the appropriate tests to answer the following questions:

a) Can we consider that the two lines are not statistically different?

b) Can the two lines be considered parallel lines?

c) Can the two intercepts be considered not statistically different?

d) For each one of the following values of the coefficient of intelligence c: 90, 105 y 120, which differences do it exists in the punctuation for each one of the methodologies?

Do again the exercise answering the same questions, but now with the punctuations that appear in column pp.

15) One wants to compare the evolution in time of the Vitamin C level of an orange juice, as a function of the type of container and the conservation temperature. Different combinations of containers and temperatures give place to what is known as *conservation method*. Three conservation methods have been considered: **"*a*", "*b*"** and **"*c*"**. For each conservation method, and during 12 weeks, the vitamin C level of two units of orange juice was analyzed. The data are in the file ***vitc.csv*** . Its structure is as follows: first column corresponds to the ***Treatment*** (treat), that in this case is the conservation method, second column corresponds to week after packaging where the Vitamin C analysis has been performed. Finally, the third column contains the vitamin C level observed (***VitC***).

It is assumed that the Vitamin C level evolves following the exponential function:

$$VitC = \alpha_i e^{-\beta_i \cdot week},$$

with $\alpha_i > 0$ and $\beta_i > 0$, and that the model parameters may depend on the conservation method, indicated by the subscript i.

Using a significance level of 5%, answer the following questions:

a) Assuming that at the moment of packaging all the juices have the same Vitamin C level, define an appropriate linear model to see if the three conservation methods lose the vitamin C in a similar way. That is to test if statistically $\beta_1 = \beta_2 = \beta_3$ or not. So,

- Compute the β_i estimations.
- Perform the test to check if the three β_i statistically different or not?

b) Before to study the evolution of the vitamin C level, it is convenient to check if at the initial moment the groups are not statistically different, this test is known as *white test*. Define a linear model appropriate to check if at the moment of packaging, the juices of the three treatments had the same vitamin C level or not. From this model:

- for each treatment, estimate the vitamin C level at the packaging moment, that is at $week = 0$.
- are the vitamin C levels statistically different at week zero?

c) Check if the linear model hypothesis are satisfied.

16) In the file PRACOVAR.csv it appears simulated data of an experiment with

- one factor (Factor)
- one continuous independent variable (explanatory variable) (X)
- several response (dependent) variables (Y1, Y2, ..., Y8).

For each one of the dependent variables and separately:

a) Describe a real situation that agrees with the experimental data. Explain how do you propose to collect the data. Perform the following descriptive statistics:
   - the tables of the total number of observations, sample means and sample standard deviations of each variable and covariate for each level of the factor and in general.
   - the scattered plot jointly with the regression line for each factor level.

b) Propose a factorial ANCOVA model (ANCOVA with interaction), and after fitting it, answer the following questions:
   - Is there any significant variable?
   - Is the factor significant?
   - Is the continuous covariate significant?
   - Do the factor and the continuous covariate interact? (Are the lines parallel?)
   - Is your model appropriate? Decide based on the plot of the residuals as a function of the predictions.

c) Write down an ANCOVA model without interaction and answer the following questions:
   - Is there any significant variable?
   - Is the factor significant?
   - Is the continuous covariate significant?
   - Is the model appropriate? Decide based on the residual versus predicted plot (perform this plot distinguishing and without distinguishing the factor levels).
   - Which model do you think is more appropriate?

2 Generalized Linear Models

2.1 General formulation

17) In what follows it appear three different variables jointly with its corresponding probability distribution.

- $Y_1 \sim \Gamma(y; \alpha, \beta)$ that is $f_{Y_1}(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y},\ y \in (0, \infty),\ \alpha > 0\ , \beta > 0$
- $Y_2 \sim B(y; n, \pi)$ that is $f_{Y_2}(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y},\ n \in \mathbb{N}^+\ (known)\ ,\ y \in \{0, 1, \dots n\}\ , \pi \in (0, 1)$
- $Y_3 \sim NB(y; \rho, \pi)$ that is $f_{Y_3}(y) = \frac{\Gamma(y+\rho)}{y!\Gamma(\rho)} \pi^y (1-\pi)^\rho,\ y \in \mathbb{N},\ \rho > 0\ , \pi \in (0, 1)$

a) Check if the probability distributions are exponential families.

b) In the case it is possible, write them in the required form to be used in GLM and deduce:

- the canonical parameters
- the canonical link
- the variance function

2.2 Normal, Gamma and other continuous models

18) The milk diary production of cows it is known to change as a function of the birth day. The production is recorded in the variable ***PROD***, and the days from the birth day in the variable ***Days***. It is known that the production (**PROD**) follows a Gamma distribution, and that it is a potencial-exponencial function of Days. That is, that it has sense to assume that

$$\mathbb{E}[PROD] = e^{\alpha + \beta Days + \gamma log(Days)} = \alpha' \cdot Days^{\beta'} \cdot e^{-\gamma Days},$$

and moreover, that the variance of the variable **PROD** rather than being constant, increases as a function of its mean. The data corresponding to this experiment appear in the ***Diaryp.csv.*** file.

a) Fit the data using linear and non linear models. Compute the parameter estimations, the expected values and perform the residual analysis. Do it for each one of the models specified in what follows:

(1) LogNormal model, that is a linear regression model with $log\,(PROD)$ as a response variable.

(2) Normal model, that is assume that the response follows a normal distribution but the model is not linear.

b) GLM with the *log* as a link function. Given that

$$\mu = e^{\alpha+\beta\cdot Days+\gamma log(Days)} \Longleftrightarrow log\,(\mu) = \alpha + \beta\cdot Days + \gamma log\,(Days)\,,$$

it is possible to fit the data assuming the logarithm as a link. Estimate the model parameters, the mean values and perform the residual analysis. Do it in the next cases:

(1) Assuming a quadratic variance function $var\,(\mu) = \mu^2$ (. *Gamma distribution.*) Important to observe that this is equivalent to apply quasi-likelihood with this variance function.

(2) Assuming a constance variance function $var\,(\mu) = constant$ (. *Normal distribution* .) Important to observe that this is equivalent to apply quasi-likelihood with this variance function.

(3) (Optional) Consider a variance function of the type: $var\,(\mu) = \mu$ (*Poisso*n). Observe that we will have warnings when the data are not positive integer values. an alternative is again quasi-likelihood with this variance function.

c) Compare the results obtained in the different models.

19) The conversion index related to the piglet fattering is defined as the following ratio: *consume/fattering* (by fattering one understands the increment of weight). The experimenters are interested in to study the conversion index as a function of the formula used to fatten the piglets. They want to consider five different formulas corresponding to the same feed but with different doses of sweetener. The data appear int he file: ***CI.csv***. The response variable is the conversion index ***CI*** and the explanatory variable is a factor that corresponds to the sweetener dose ***Sweet***.

These data have the particularity that one could also define the response variable as the *fattering/consume* ratio, and it is reasonable to think that both variables follow the same type of distribution. The Inverse Gaussian distribution is the continuous probability distribution that has this property, that is that if Y follows an inverse Gaussian distribution, the random variable Y^{-1} also follows an inverse Gaussian distribution, with different parameters. Fit the following GLM models and write the conclusions obtained from your fittings.

a) Assuming Inverse Gaussian distribution with canonical link: $1/\mu^2$

b) Assuming Inverse Gaussian distribution with link identity.

c) (optional) Assuming a quasi-likelihood model with identity link and variance function: $var(\mu) = \mu^3$.

d) Compare the results obtained with the different models.

20) One has the results of an experiment focused on modeling the area covered from different plants in slopes located in highways. Two species of plants are considered. Moreover it has also been recorded if some compost has been added or not. The response variable is the ***AREA*** and the two factors considered are ***SPECIES*** and ***COMPOST***. The data appear in the file ***area.csv.***

a) Fit the data using the ANOVA technique. Estimate the parameters and perform the multiple comparisons and the residual analysis. Do it in the following two situations:

(1) Assuming a normal distribution and homoscedasticity.

(2) Assuming a normal distribution for the logarithm of the area (that is equivalent to assume a logNormal for the area) and homoscedasticity.

b) Fit the data using GLM. Obtain the parameter estimations, perform the multiple comparisons and the residual analysis. Do it in the following three situations:

(1) Assuming Normality, link=***identity*** and var=***constant.***

(2) Assuming a Gamma distribution with link=**identity** and var=mu^2**.**

c) Fit the data by means of quasi-likelihood assuming link=**identity** and var=μ^2**.**

d) Compare the result obtained in the different implemented models.

21) This exercise is the continuation of the **15**) where we had that:

One wants to compare the evolution in time of the Vitamin C level of an orange juice, as a function of: the type of container and the conservation temperature.

To that end, three conservation methods where considered: ***"a", "b"*** *and* ***"c"****.*

For each conservation method, and during 12 weeks, two units of orange juice where analyzed.

The data are in the file ***vitc.csv****. Its structure is as follows: first column corresponds to* ***Treat****: conservation method, second column corresponds to* ***Week****: and it indicates the time after packaging, the third column indicates corresponds to* ***VitC****: level of vitamin C that has been observed.*

It is supposed that the Vitamin C level evolves following the exponential function:

$$VitC = \alpha_i e^{-\beta_i \cdot week},$$

with $\alpha_i > 0$ *and* $\beta_i > 0$ *, and that these parameters may depend on the conservation method, indicated by the subscript* i*.*

Assuming that in the moment of packaging may exists differences between the levels of Vitamin C, and using a significance level equal to 5%, answer the following questions:

a) Define a generalized linear model with the **"gamma"** family, useful to check if the treatments loss Vitamin C at the same velocity, that is if $\beta_1 = \beta_2 = \beta_3$ or not, and also to see if the three values of α_i are or not statistically equivalent. From this model,

- Estimate α_i. Are they statistically different?
- Estimate β_i. Are they statistically different?

b) Define a ***LogNormal*** generalized linear model useful to check if the treatments loss Vitamin C at the same velocity, that is if $\beta_1 = \beta_2 = \beta_3$ or not, and also to see if the three values of α_i are or not statistically equivalent. From this model,

- Estimate α_i. Are they statistically different?
- Estimate β_i. Are they statistically different?

c) Define a generalized linear model with the **"normal"** family, useful to check if the treatments loss Vitamin C at the same velocity, that is if $\beta_1 = \beta_2 = \beta_3$ or not, and also to see if the three values of α_i are or not statistically equivalent. From this model,

- Estimate α_i. Are they statistically different?
- Estimate β_i. Are they statistically different?

d) (optional) Define a quasi-likelihood model as a function of **"mu"**, useful to check if the treatments loss Vitamin C at the same velocity, that is if $\beta_1 = \beta_2 = \beta_3$ or not, and also to see if the three values of α_i are or not statistically equivalent. From this model,

- Estimate α_i. Are they statistically different?
- Estimate β_i. Are they statistically different?

e) Justify which of the four models that you have fitted better verify the model hypothesis and gives place to the bests fits.

2.3 Binomial and Poisson models

22) In the file ***clusters.csv*** we have two variables: *Cancers* and *distance*.

- Cancers: number of cancers detected in a given hospital in a year.
- Distance: distance from the hospital to a nuclear power station.

The objective is to know if it exists a relationship between the the proximity to a nuclear power station and the incidence of cancer in general. Explain in this setting which are the response variable, the explanatory variable and which is their type.

a) Plot the data.

b) Fit the appropriate GLM to explain the *Cancers* as a function of the *Distance*. Assume that $Cancers \sim Poisson\,(\mu)$ where μ depends on the *Distance* in the way is explained in the following function: $\mu = e^{\alpha+\beta \cdot Distance}$.

(1) Which probability distribution *family* and which *link* function do you need to consider?

(2) Is the relationship between *Cancers* and *Distance* significant at a 5% level ?

(3) Does it exists overdispersion?

c) Repeat the analysis of point b) assuming that μ depends on *Distance* by means of the function: $\mu = \alpha + \beta \cdot Distance$.

(1) Which probability distribution *family* and *link* function do you need to use?

(2) Is the relationship between *Cancers* and *Distance* significant at a 5% level ?

(3) Does it exists overdispersion?

(4) Can it appear any problem using this second model?

d) Optional. The same as b) but assuming that it may be overdispersion or underdispersion.

(1) Which probability distribution *family* and which *link* function do you need to use?

(2) Is the relationship between *Cancers* and *Distance* significant at a 5% level ?

23) In a given study the objective is to test the effectivity of different doses of a given insecticide. The variables are: the insecticide dose, ***DOSE***, the total number of insects that have received a particular dose, ***T***, and the number of insects that died with each particular dose, ***DIED***. The data appear in the file ***insecticide.csv***. It is known that it has sense to assume a model in which the probability of dying π_i at dose i is modeled by means of the following the function: $\pi = \Phi\left(\alpha + \beta \cdot log\left(DOSE_i\right)\right)$, where Φ is the cumulative distribution function of a $N(0,1)$ distribution.

a) Perform the data analysis using linear regression models. Estimate the parameters, the mean vector and perform the corresponding residual analysis.

Apply the transformation $\Phi^{-1}\left(\frac{DIED}{T}\right)$ and fit the data again with a linear model, that is assuming normality and homocedasticity. Problems with the values: (DIED/T)=0 o 1.

b) Perform the analysis using generalized linear models. Estimate the parameters, compute the mean vector and perform the residual analysis. Do it in the two following cases:

(1) Assume a binomial distribution with link=**probit**.

(2) Assume a binomial distribution with link **logit**, **cloglog**, **log**,.....

(3) Perform a comparative analysis of the results obtained in the two previous sections.

24) The number of flowers produced by the plant Argyranthemums (crisantemus) grown in four different types of substratum is recorded for each substratum. The file ***flowers.csv.*** contains the variables: number of flowers produced ***FLOWERS*** and type of substratum ***SUBSTRAT***. Observe that it has sense to assume that the variable *Flowers* follows a Poisson distribution for each substratum, and thus, as a consequence, the variances will also depend on the substratum.

a) Fit the data using a linear models. That is using the ANOVA technique. Perform the parameter estimation, compute the expected values, and perform the multiple comparisons and residual analysis. Do it in the following two assumptions:

(1) Assuming normality and homoscedasticity for the response variable.

(2) Apply the transformation: $\sqrt{FLOWERS + 3/8}$ and assume normality and homoscedasticity for the transformed response variable. Compare the results with the ones obtained before.

b) Fit the data using a GLM in the two following situations:

(1) Assume a Poisson distribution for the response variable with the corresponding canonical link.

(2) Assume a Poisson distribution for the response with link equal to the Identity.

In each case, analyze if there is or not overdispersion (underdispersion), compute the parameter estimations, the predicted mean values, perform multiple comparison and perform the residual analysis. Compare the results obtained in the last two sections.

25) In a germination study performed in packs of 30 seeds, the seeds were submitted to a pre-treatment ***PRE*** and to a different temperatures ***TEMP***. After a given period of time, the number of seeds that germinated for any pre-treatment and temperature were rerecorded ***GN***. The data appear in the file: ***seeds.csv***. We denote by ***TN*** the total number of seeds that, in this case, is equal to 30.

a) Fit the data using linear models, ANOVA, for each one of the next situations. In each case obtain: the parameter estimations, the expected values and perform the multiple comparison for each temperature. Also perform a residual analysis to see if the model is or not appropriate and if the linear model hypothesis are satisfied.

(1) Assume normality and homoscedasticity. Check by means of the residuals that the stochasticity hypothesis is not very clear, which invalidates the obtained results.

(2) Apply the transformation $\arcsin\left(\sqrt{\frac{GN+3/8}{TN+3/4}}\right)$ to the response variable and fit the data assuming normality and homoscedasticity for the new response variable. Check by means of the residual analysis the homoscedasticity hypothesis, and compare the results obtained in the multiple comparisons with the ones obtained previously.

b) Fit the data using GLM. Analyze the overdispersion (underdispersion). For each temperature, estimate the mean values and perform the multiple comparison of the pre-treatment and the residual analysis. Do it in the following situations:

(1) Assuming a binomial distribution with canonical link (**logit**). Compare the results with the ones obtained in a).

(2) Assuming a binomial distribution with three different links: **probit**, **log**, and **identity**. Compare the results with the ones obtained in b)(1).

26) The number of species in a given area has been recorded (***Species***). The areas considered have different biomass levels which are recorded in the continuous variable ***Biomass***. It also has different values for the floor pH, recorded in the variable ***pH***, that is assumed to be categorical with three different levels: *high*, *mid* and *low*. Assuming that for a given value of the biomass and a given level of pH the variable *Species* follows a Poisson distribution, one wants to see if both variables have a significant effect on the number of species at a 5% level. To that end,

a) Show the data graphically.

b) Fit the appropriate GLM using the canonical link function.

(1) Which effects are significant?

- For a mean value of the biomass, Which differences do it exists between the three levels of pH?
- For a given *pH*, do it exists differences in the *Species* as a function of the *Biomass*?

(2) Does it exist over or underdispersion?

(3) Check if the model may be considered a good model for this set of data.

c) Compute the predicted number of *Species:*

(1) For each pH, which is the expression of $\mathbb{E}(Species)$ as a function of the *Biomass*?

(2) Plot the data jointly with the predicted number of species for each level of *pH*.

(3) For each level of *pH*, plot the predicted interval of Species at 95% .

2.4 Advanced exercises

27) The Botrytis is a fungus that causes several diseases in plants and specially in vineyards. We have measured the inhibition produced by some fungicide in this fungus. The inhibition is a continuous variable that takes values between zero and one. Nevertheless, the experimental results may give some negative values as a consequence of the the way of obtaining this measure. Several fungicide doses have been considered and the corresponding inhibition value has been measured. The data appear in file: ***botrytis.csv***. The regression model appropriate to fit this situation is the following:

$$\mathbb{E}[INHIB] = \frac{\alpha \cdot DOSE^{\beta}}{1 + \alpha \cdot DOSE^{\beta}}$$

Observe that if $\mu = \mathbb{E}[INHIB]$, the logarithm of the inverse of μ is

$$log\,(DOSE) = -\frac{log\,(\alpha)}{\beta} + \frac{1}{\beta} log\left(\frac{\mu}{1-\mu}\right)$$

which is equivalent to say that

$$\log\left(\frac{\mu}{1-\mu}\right) = \log(\alpha) + \beta \cdot \log(DOSE) = \alpha' + \beta \cdot log(DOSE) \tag{2.4.1}$$

Fit the data with the following instructions:

a) Perform a linear regression. Estimate the parameters, the mean vector and the residual analysis after applying the transformation $z = log\left(\frac{INHIB}{1-INHIB}\right)$. This means to fit a linear regression assuming z as a response variable and $log(DOSI)$ as a covariate. Observe that the point such that $INHIB < 0$ can not be taken into account and that there exist an influential point.

b) Using no linear regression with the original data (without transforming).

c) (1) Fit the GLM with link=*logit* corresponding to model (2.4.1). Compute the parameter estimations, the mean vector and analyze the residuals. Assume the Normal family, which means that the variance function is constant.

(2) Assume a quasi-likelihood model with variance function $var(\mu) = 1 + \mu$, that need to be defined by us.

In this case, given that $q(y;\mu) = \int \frac{y-\mu}{1+\mu} d\mu = (1+y) \cdot log(1+\mu) - \mu$, the deviance is equal to:

$$dev = 2 \cdot (q(y;y) - q(y;\mu)) = 2 \cdot (1+y) \cdot log\left(\frac{1+y}{1+\mu}\right) - 2 \cdot (y-\mu)$$

The instructions to do that are the following:

```
varf <- function(mu) 1+mu
valid <- function(mu) all(mu > -1)
dev <- function(y,mu,wt) 2*wt*((1+y)*log(ifelse(y>=1,1,(1+y)/(1+mu)))-(y-mu))
init <- expression({
    n <- rep.int(1,nobs)
    mustart <- pmax(0.001, pmin(0.999, y))})
‘1+mu‘<-list(varfun=varf,validmu=valid,dev.resids=dev,initialize=init)
```

(3) The same analysis as in c.1) and c.2) but assuming that:

$$E[INHIB] = \frac{1}{2} \cdot \left(1 + \frac{\alpha + \beta log(DOSE)}{\sqrt{1 + (\alpha + \beta log(DOSE))^2}}\right),$$

which requires the following link function: $link = \frac{\mu - 1/2}{\sqrt{\mu(1-\mu)}}$, that will be defined by us and denote by “sigmo”.

The following instructions are necessary:

```
linkfun <- function(mu) (mu-1/2)/sqrt(mu*(1-mu))
linkinv <- function(eta) 1/2*(1+eta/sqrt(1+eta^2))
mu.eta <- function(eta) 1/2/(1+eta^2)^(3/2)
sigmo <- list(linkfun = linkfun, linkinv = linkinv, mu.eta = mu.eta)
```

28) In the city of Barcelona, an experimental study is being held to avoid the hydric stress suffered for the plane trees. It consists on applying the given treatment to the ground and to measure several variables in the banana tree. One suspects that probably in the first years of treatment no changes will be appreciated. The variables that have been measures at the initial point are the following: the tree perimeter ***PER0*** and the treatment ***TREAT***. In the next year, they have also measured: the number of leaves, ***LEAVES***, the number of principal branches, ***BRANCH***, and the number of secondary branches ***SEC***. The data appear in the file ***platanus.csv***.

Fit the data using GLM. Compute the parameter estimations, the mean values, perform the multiple comparisons, analyze if there exist any tendency, and do the residual analysis. Do it in the following cases.

a) First we want to see if there is any effect of *TREAT* over *LEAVES* taking into account *PER0* in a potential way. To that end, we will use link=log and explanatory variable *log(PER0)*. Do it assuming:

   (1) A Poisson model.

   (2) A quasi Poisson model.

   (3) Compare the results obtained.

b) The same as in a) but with *BRANCH*, instead of *LEAVES*.Do it assuming:

   (1) A Poisson model.

   (2) A quasi Poisson model.

   (3) Compare the results obtained.

c) Fit the linear regression model $LEAVES \sim BRANCH+SEC$, using GLM and the identity link. Do it assuming:

   (1) A Poisson model.

   (2) A quasi Poisson model.

   (3) Compare the results obtained.

d) Assuming that BRANCH is known in advance, and that the number of leaves for each tree instead of following a Poisson(λ) distribution with the same λ for all trees, follows a Poisson($\lambda 1$) where λ_1 is constant for all the branches, one has as a consequence that *LEAVES* follows a $Poisson\,(BRANCH \cdot \lambda_1)$ distribution. This situation may be adjusted considering that *BRANCH* is an offset variable, or using weights. Fit the following model:

(1) Quasi-Poisson with logarithmic link. That is:

$$E[LEAVES] = BRANCH \cdot \lambda_1,$$

where $\lambda_1 = e^{\alpha+\beta \cdot SEC}$. This gives place to the quasi-Poisson model with logarithmic link:

$$LEAVES \sim offset(log(BRANCH)) + log(SEC).$$

Check that the all the hypothesis test are significant and that the residuals are reasonable but there exists overdispersion.

(2) Do the same as in d)(1) but using weights. In this case we consider the variable **LEAVES/BRANCH**, and we take as a weight **BRANCH**. Fit a quasi-Poisson model with logarithmic link and weights=**BRANCH**, assuming

$$LEAVES/BRANCH \sim log(SEC)$$

(3) To model $E[LEAVES] = BRANCH \cdot \lambda_1$, *with* $\lambda_1 = \alpha + \beta \cdot SEC$, *it is not possible to use the logarithmic link. So, use the quasi-Poisson model with identity link and weights=****BRANCH****.*

$$LEAVES/BRANCH \sim SEC$$