Modelos Matemáticos e Aplicações – 2020-21

Exercises - Generalized Linear Models

NOTE: The file `dadosMLG.RData` contains the following data frames: `tabaco` (Exercise 1), `ratos` (Exercise 2), `Elisa1` (Exercise 5), `Elisa2` (Exercise 6), `flea.beetles` (Exercise 9) and `sangue` (Exercise 11). The file must be loaded into an `R` session with the `Load Workspace` menu or `load` command.

1. The book by W.N. Venables and B.D. Ripley, *Modern Applied Statistics with S-Plus* (1994, Springer-Verlag), has data from an experiment which studies the resistance of the tobacco budworm *heliothis virescens* to different doses of a toxic substance. Groups of 20 moths of each sex were exposed to different doses of the substance and, after three days, the number of dead (or inactive) individuals in each group was recorded. The results (labelled as deaths) are shown in the following table (where the doses are given in μg).

	Dose					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

   (a) Create a data frame containing the data and suitable to fit models with a Binomial/n random component.

   (b) Draw a scatterplot with the variable `Dose` on the horizontal axis and the *proportion* of deaths for each group of 20 individuals on the vertical axis. Repeat, but now using different colours to represent the data for the individuals of each sex. Comment your results.

   (c) Repeat the previous steps, but now associating the horizontal axis with the values of $\log_2$(Dose). This transformation is justified by the fact that each dosage used in the experiment is twice the previous dosage. Comment.

   (d) Fit a Logistic Regression to the data, ignoring the factor sex and using $\log_2$(Dose) as the numerical predictor. Comment your results. Draw, on the scatterplot from the previous question, the estimated curve for the probability of death, $p(x)$, where x indicates the values of $\log_2$(Dose). Discuss the significance of the estimated parameter value b_1.

   (e) Repeat the previous question, but now using a *Probit* model. What is the dosage corresponding to a 50% probability of death?

   (f) Now fit a generalized linear model with the appropriate random component, but using a complementary log-log link function. Comment your results.

2. In order to study the carcinogenic effects of a toxic product on mice, three different dosages of the toxic substance were administered (0, 0.45 and 0.75 parts per 10 000) to a few hundred mice, during one of two exposure periods (16 or 24 months). At the end of the period of exposure, the mice were checked for tumours. These were the results of the experiment:

		Dosage		
Exposure		0	0.45	0.75
16 months	Mice with tumours	1	3	7
	Mice without tumours	204	301	186
24 months	Mice with tumours	20	98	118
	Mice without tumours	742	790	469

The data are available in the data frame `ratos`. A Generalized Linear Model appropriate for a binary random component was fitted, using the *probit* link function and, as numerical predictors, dosage (`Dose`) and exposure time (`Exposicao`). These were the results:

```
> summary(ratos.probit.var)
Call:
glm(formula = cbind(com, sem) ~ Dose + Exposicao, family = binomial(probit),
    data = ratos)
[...]
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8474     0.3948 -12.279  < 2e-16 ***
Dose          1.4344     0.1397  10.269  < 2e-16 ***
Exposicao     0.1229     0.0163   7.538 4.78e-14 ***
---
    Null deviance: 198.5347  on 5  degrees of freedom
Residual deviance:   1.3381  on 3  degrees of freedom
AIC: 33.594
Number of Fisher Scoring iterations: 4
```

(a) Describe in detail the fitted model, specifying the relation that is assumed between the appearance of tumours and the predictor variables.

(b) Discuss the goodness-of-fit of the model to the data.

(c) Is it possible to further simplify the model without a significant loss in the goodness-of-fit? Justify with a formal test.

(d) Based on the fitted model, answer the following questions:

   i. For a dose of 0.75 parts per 10 000 of the toxic substance, what is the expected proportion of mice with tumours after 36 months of exposure?

   ii. What is the dose associated with 50% of mice with tumours after 24 months of exposure?

In the meantime, an objection is raised, stating that the very small number of different values of the predictors `Dose` and `Exposicao` does not recommend using them as numerical variables. It was decided to fit a new model, with these two predictors considered as factors. Interaction effects between the factors are not envisaged. The fit produced the following results:

```
> summary(ratos.probit.fac)

Call:
glm(formula = cbind(com, sem) ~ as.factor(Dose) + as.factor(Exposicao),
    family = binomial(probit), data = ratos)
[...]
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.9038     0.1561 -18.602  < 2e-16 ***
as.factor(Dose)0.45      0.6880     0.1069   6.435 1.24e-10 ***
as.factor(Dose)0.75      1.0859     0.1081  10.042  < 2e-16 ***
as.factor(Exposicao)24   0.9826     0.1302   7.545 4.52e-14 ***
[...]
    Null deviance: 198.5347  on 5  degrees of freedom
```

```
Residual deviance:   1.0902  on 2  degrees of freedom
AIC: 35.347

Number of Fisher Scoring iterations: 4
```

(e) Describe in detail the model that was fitted. Comment the analogies and the differences between this model and the model that was considered initially.

(f) What is the probability, estimated by the model, that a mouse will have a tumour at the end of 16 months, if it was not exposed to the toxic? How does this estimated probability compare with the relative frequency of tumours in that experimental situation? How does this estimated probability compare with the corresponding probability resulting from the initial model? Discuss.

(g) Is it possible to estimate the probability of mice having tumours when exposed for 36 months, using this model?

(h) Based on the indicators of goodness-of-fit available and taking into account the reservations that were raised regarding the initial model, which of these two models would you pick?

(i) Now fit a third model, considering `Dose` and `Exposicao` as factors, but also allowing for interaction effects. How do you explain the fact that the model deviance, and all the deviance residuals are zero? What are the implications of this fact?

3. The `MASS` package has a data frame called `Traffic`, with results from a study of the impact of police controls of speed limits on Swedish roads, carried out in 1961 (see `help(Traffic)` for more details).

   (a) Fit a log-linear model whose random component is the number of accidents recorded on each day, and with an explanatory factor with only two levels: whether or not the speed limits were being enforced. Interpret the fitted parameter estimates.

   (b) Calculate the mean number of accidents on days with speed limits and the mean number of accidents on days without speed limits. Relate the fitted parameter estimates with the values obtained in the previous question.

   (c) **[Supplementary material]**. Determine the equations of the system that is obtained by equating to zero the partial derivatives of the log-likelihood of the log-linear model for this question. Solve the system. State whether or not the relations observed in the questions above are a coincidence.

   (d) Discuss the comparative advantages of using a log-linear model in this study, as compared to using the classical t-test to compare the mean values of the number of accidents per day in the two populations (with, and without, speed limits).

4. In the `MASS` package there is a site × species contingency table, given in an object called `waders`. The dataset has observed frequencies of 19 species of waders (shorebirds), in 15 different locations along the coast of Southern Africa (Namibia and South Africa).

   (a) Carry out a standard χ^2 independence test for the factors "sites" and "species", using Pearson's statistic (**Note:** The `R` command for this test is `chisq.test`.)

   (b) Create a data frame suited for fitting a GLM to the data, that is, a data frame with the following three columns: the counts, the sites and the species. Use the following `R` command:

```
> limicolas <- data.frame(obs=as.vector(as.matrix(waders)), local=rep(LETTERS[1:15],19),
                                             especie=rep(paste("S",1:19,sep=""),each=15))
```

(c) Consider a log-linear model for the data, with two (additive) explanatory factors: `local` and `especie`. Discuss the details of the model equation. Indicate the expected value, given by the model, for the number of observations of species S14, at site C.

(d) Fit the model given in the previous question and discuss its goodness-of-fit based on the model deviance. Compare the number of observations of species S14, at site C, with the corresponding fitted value. Comment.

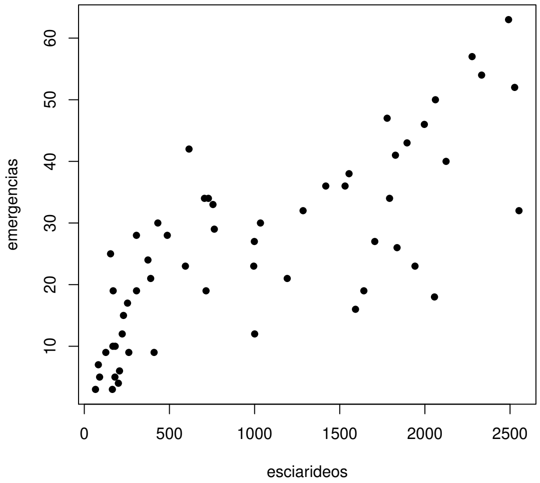
(e) Calculate the sum of squared Pearson residuals for this model. Compare your result with the value of the Pearson statistic from the χ^2 test of the first question. Comment.

(f) Interpret the meaning of the difference of two parameters of the same type of effect, such as for example, $\alpha_4 - \alpha_3$, where α_i denotes the effect of the i-th level of the factor `local`.

(g) Build an (asymptotic) confidence interval for $\alpha_4 - \alpha_3$ and interpret it.

(h) Comment the usefulness of your model, based on the results above.

5. The adult female of a predatory species lays her eggs in a substrate of soil containing oats with fungi, infested with mosquitoes that serve as food for the larvas. The goal is to relate the number of mosquito larvas present in the substrate - variable `esciarideos` - and the number of adults that emerge in the subsequent generation (after feeding as larvas and after pupation) - variable `emergencias`. The number of mosquitoes was calculated by extrapolating the number of larvas observed in a sample to the total substrate volume. The resulting data are in a data frame called `Elisa1` and the relevant scatterplot is shown below.



(a) Draw the scatterplot shown above using `R`.

(b) Do you think that a model for the response variable `emergencias` that assumes a Poisson distribution is suitable?

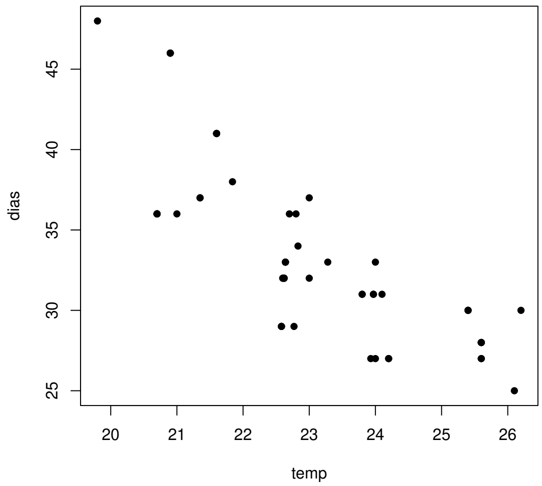
(c) Do you think that the canonical link function for Poisson distributions is a suitable link function?

(d) Fit a log-linear model and discuss your results. The estimate for the parameter β_1 is $b_1 = 0.0005248347$. How can this value be interpreted in the context of this problem?

(e) Draw, on the scatterplot, the curve fitted by the model. Comment it.

(f) Calculate 95% confidence intervals for the model parameters (β_0 and β_1), using the asymptotic theory for maximum likelihood estimators. Discuss it. In particular, state whether, based on these intervals, it can be said that an increase in the number of mosquitoes present in the substrate is associated with an increase in the mean number of adults in the subsequent generation.

6. A pest control study attempts to model, for a given insect species, the relation between the number of days between the moment they are laid and the emergence of new adults (the response variable `dias`) and the environment temperature (the predictor `temp`). The study involved $n=57$ repetitions, given in the data frame `Elisa2`, associated with the following scatterplot:



(a) Fit a log-linear model to the data. In particular,

i. Describe your choices.

ii. How well does the log-linear model fit the trend observable in the scatterplot?

iii. Draw the fitted curve on the scatterplot.

(b) An analyst suggests that, since the response variable `dias` measures time, it is in reality a continuous random variable that is discretized when measured. He suggests that it is possible to make a single modification to the previous GLM: consider that the response variable has a Normal distrbution. Describe this new GLM and, in particular:

i. Explain why this new model is *not* a *Linear* Model.

ii. Write the equation of the fitted curve and draw it on the scatterplot. How can we explain the fact that the fitted curve is different? And how can we explain that it is similar to the previously fitted curve?

iii. Consider the residual deviance associated with this model and discuss the fact that it is substantially different from the deviance of the previous model. In particular, discuss the following statement: "*the model fitted in the previous question is better, because it has a smaller deviance*".

(c) Now fit the *Linear Model* that is most similar to the model in the previous question. In particular,

i. Write the model equation and assumptions. Compare them with those of previous models.

ii. What is the equation of the fitted curve? Draw the curve on the scatterplot.

iii. Study the residuals of this *linear* model and discuss the validity of the model assumptions.

iv. Since a linear model is a specific instance of a GLM, it makes sense to talk about the residual deviance of the model that was now fitted. Calculate it using `R`. Can this value be compared with the value obtained in the previous question, in which a Normal distribution of the random component was also assumed?

7. There are alternative parametrizations for the Gamma density function. The parametrization shown in the slides is:

$$f(y \mid \mu, \nu) \;=\; \frac{\nu^\nu}{\mu^\nu \Gamma(\nu)}\; y^{\nu-1}\; \mathrm{e}^{-\frac{\nu y}{\mu}} \;.$$

In this parametrization, μ is the expected value of the variable and the second parameter, ν, appears in the expression for the variance: $V[Y] = \frac{\mu^2}{\nu}$.

(a) A different parametrization of the Gamma density is:

$$f(y \mid \alpha, \beta) \;=\; \frac{1}{\beta^\alpha \Gamma(\alpha)}\; y^{\alpha-1}\; \mathrm{e}^{-\frac{y}{\beta}} \;.$$

Show that this is the same function, but with new parameters, related by $\mu = \alpha\beta$ and $\nu = \alpha$.

(b) In the book *Probabilidades e Estatística*, by Prof. Bento Murteira (McGraw-Hill Portugal, 1979), a third parametrization of the Gamma density is given:

$$f(y \mid n, \gamma) \;=\; \frac{\gamma^n}{\Gamma(n)}\; y^{n-1}\; \mathrm{e}^{-\gamma\, y} \;.$$

Identify the relations between the parameters in this expression and those of previous parametrizations. Relate the expected value and variance in this parametrization with those of the parametrization used in the classes.

8. Define the following concepts, in the context of Generalized Linear Models:

(a) link function

(b) deviance residual

9. Nineteen beetles of the species *Altica oleracea* and twenty beetles of the species *Altica carduorum* were subjected to morphometric measurements in four variables: the distance from the transversal groove to the posterior border of the pro-torax (variable *TG*), the length of the elytra (variable *Elytra*), the length of the second segment of the antennae (variable *Second.Antenna*) and the length of the third segment of the antennae (variable *Third.Antenna*).

The units of measurement of all variables *except the length of the elytra* are micrometers (the millionth part of the meter, μm). The length of the elytra is given in one hundredths of a millimeter ($10\,\mu m$).

Some of the data collected can be seen below.

```
     Species   TG Elytra Second.Antenna Third.Antenna
1   oleracea  189    245            137           163
2   oleracea  192    260            132           217
3   oleracea  217    276            141           192
4   oleracea  221    299            142           213
(...)
18  oleracea  181    255            146           183
19  oleracea  192    287            141           198
20 carduorum  181    305            184           209
21 carduorum  158    237            133           188
(...)
36 carduorum  192    276            154           209
37 carduorum  181    278            149           235
38 carduorum  175    271            140           192
39 carduorum  197    303            170           205
-----------------------------------------------------
variância   196.888  502.7085      216.0445      341.8313
média       186.8205 279.2308      147.5385      197.8974
```

Haltica oleracea

We seek a model to identify a given species of beetle, that is, we wish to obtain a model that discriminates between the species. Given the difficulty in obtaining precise measurements, due to the animals' small size, it was considered important to have a parsimonious model, that is a model with as few morphomteric predictors as possible.

(a) A Logistic Regression was fitted, initially with the four morphometric variables that are shown. The following results were obtained.

```
Call: glm(formula =  (Species == "carduorum") ~ TG + Elytra + Second.Antenna
   + Third.Antenna, family = binomial, maxit = 50, data=flea.beetles)
Coefficients:
                 Estimate Std. Error   z value Pr(>|z|)
(Intercept)    -6.237e+02  1.869e+06 -3.34e-04        1
TG             -1.162e+01  2.077e+04    -0.001        1
Elytra          5.559e+00  9.735e+03     0.001        1
Second.Antenna  7.634e+00  1.757e+04  4.34e-04        1
Third.Antenna   8.133e-01  1.411e+04  5.77e-05        1

    Null deviance: 5.4040e+01  on 38  degrees of freedom
Residual deviance: 4.7616e-10  on 34  degrees of freedom
AIC: 10          Number of Fisher Scoring iterations: 28
```

i. Describe the fitted model in detail, as a Generalized Linear Model, specifying its three components.

ii. Discuss the model's quality, for the purpose of identifying the species of beetle. How can the fitted model's almost null deviance be explained? Is there a problem of overparametrization?

iii. Interpret the estimated value 7.634 of the parameter associated with the variable *Second.Antenna*.

iv. Based on the available information, do you think it is possible to simplify the model without a significant loss in discriminatory capacity? If so, what is the first predictor that can be excluded from the model, in a backward elimination approach?

(b) A backward elimination stepwise approach was followed, using R's `step` function. Comment the various steps in the algorithm and identify the final model.

```
> step(flea.glm.logit)
Start:  AIC=10
(Species == "carduorum") ~ TG + Elytra + Second.Antenna + Third.Antenna

                 Df Deviance    AIC
- Third.Antenna   1    0.000  8.000
- Second.Antenna  1    0.000  8.000
<none>                 0.000 10.000
- Elytra          1   10.132 18.132
- TG              1   24.686 32.686

Step:  AIC=8
(Species == "carduorum") ~ TG + Elytra + Second.Antenna

                 Df Deviance    AIC
<none>                0.0000  8.000
- Second.Antenna  1   9.8414 15.841
- Elytra          1  16.6409 22.641
- TG              1  29.7719 35.772

Call:  glm(formula = (Species == "carduorum") ~ TG + Elytra +
      Second.Antenna, family = binomial, data = flea.beetles, maxit = 50)

Coefficients:
   (Intercept)              TG          Elytra  Second.Antenna
       -968.93          -19.46            9.37           13.91

Degrees of Freedom: 38 Total (i.e. Null);  35 Residual
Null Deviance:     54.04
Residual Deviance: 3.846e-10  AIC: 8
```

(c) Regardless of your answer in the previous question, it was decided to fit a model with only two predictors. The best model of this kind dropped the measurements relative to the antennaes. Results are shown below.

```
Call: glm(formula = (Species == "carduorum") ~ TG + Elytra,
                    family = binomial, maxit = 50, data=flea.beetles)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  10.1559    12.8285   0.792   0.4286
TG           -0.4271     0.1792  -2.384   0.0171 *
Elytra        0.2505     0.1038   2.413   0.0158 *
---
Null deviance: 54.0398  on 38  degrees of freedom
Residual deviance:  9.8414  on 36  degrees of freedom
AIC: 15.841    Number of Fisher Scoring iterations: 8
```

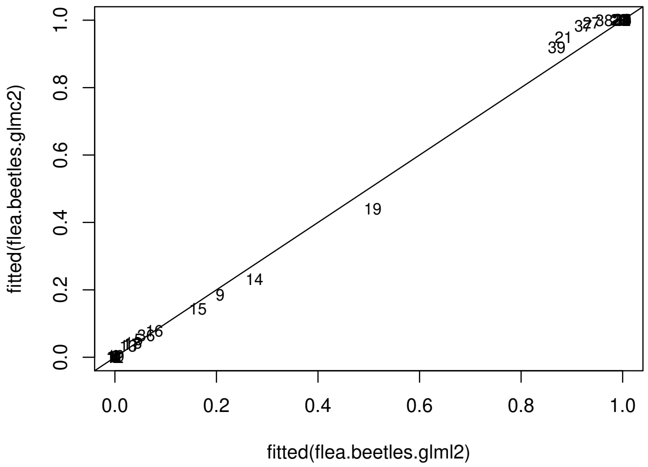
i. Formally test whether this model and the initial model are significantly different.

ii. For each species, what are the probabilities predicted by the model that was now fitted, for a beetle with $TG = 200$ and $Elytra = 250$? What species would you associate with a beetle with those characteristics?

(d) It was then decided to try out a different link function, in particular the complementary log-log link function, using only the two predictors mentioned in question 9c. The results now obtained are the following:

```
Call: glm(formula = (Species == "carduorum") ~ TG + Elytra,
       family = binomial(link = "cloglog"), maxit = 50)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.78272    7.75729   1.003   0.3157
TG          -0.33889    0.13206  -2.566   0.0103 *
Elytra       0.19769    0.07766   2.546   0.0109 *
---
    Null deviance: 54.0398  on 38  degrees of freedom
Residual deviance:  8.7522  on 36  degrees of freedom
AIC: 14.752      Number of Fisher Scoring iterations: 12
```

i. The following plot shows the fitted probabilities for each model, with the probabilities from the complementary log-log model on the vertical axis and the probabilities for the model with the canonical link function on the horizontal axis. Comment your results. In particular, discuss individual 19.



ii. Which two-predictor model do you prefer: this one, or the one discussed in question 9c? Justify your answer.

iii. What is the predicted probability for an individual with observed values $TG = 200$ and $Elytra = 250$? Compare this with the corresponding result for the model in question 9c and comment.

10. Consider again the data in Exercise 1. Fit a *probit* regression model for the probability of death, but now considering in the systematic component not just the numerical variable $\log_2(dose)$, but also the factor `sexo`.

(a) Obtain a single model that may be interpreted as having two different systematic components, $\beta_0 + \beta_1 \log_2(Dose)$, one for males and the other for females, each with its own parameters.

(b) Fit the model indicated in your previous reply to the data and comment. Can we consider this model to be better than the model fitted in Exercise 1?

(c) Now consider a third model, in which the systematic component assumes that the coefficient for the log-dose is the same in both sexes, but a different additive constant may exist. Fit the model and compare its results with those obtained for the previous two models. Discuss.

(d) Which of these three models would you choose? Justify your choice.

11. The book by P. McCullagh and J.A. Nelder, *Generalized Linear Models* (2d. edition, 1989, Chapman & Hall), on pages 300-302, discusses a dataset where the clotting times of blood (in seconds) for normal plasma diluted with nine different concentrations of prothrombin-free plasma (the prothrombin protein is produced in the liver and, when activated - generating thrombine - is associated with the clotting of blood). Two different lots of the activating agent of clotting were used. The observed data are shown below and given in the data frame `sangue`.

	Coagulation time	
Concentration	Lot 1	Lot 2
5	118	69
10	58	35
15	42	26
20	35	21
30	27	18
40	25	16
60	21	13
80	19	12
100	18	12

We seek to study the effects of different concentrations of prothrombine-free plasma on the coagulation times. We begin by ignoring possible lot effects.

(a) Plot coagulation times (`tempo`, on the vertical axis) versus plasma concentrations (`conc.plasma`, horizontal axis), using different symbols and/or colours to represent the observations from each lot. Comment.

(b) It is suggested that the relation between the variables `tempo` and concentration of prothrombine-free plasma (variable `conc.plasma`) follows a hyperbolic-type relation, of the form $tempo = \frac{1}{\beta_0+\beta_1 \cdot conc}$. Produce a suitable graphical representation to visually validate this suggestion. Comment.

(c) After a further visual inspection, it was concluded that the most adequate relation seems to be a hyperbolic-type relation, but on the logarithms of plasma concentration, that is, of the form $tempo = \frac{1}{\beta_0+\beta_1 \ln(conc)}$. Confirm this, by producing a suitable graphical representation.

(d) To fit the model indicated in the previous question, the link function is the reciprocal function, $g(\mu) = \frac{1}{\mu}$, using as a predictor the variable of log-concentrations. But the issue of which distribution should be associated with the response variable *tempo* remains an open issue. Fit two different GLMs, assuming:

   i. that `tempo` has a Normal distribution (Note: In *R*, this assumption corresponds to using the argument `family=gaussian(link=”inverse”)` in the `glm` command);

   ii. that `tempo` has a Gamma distribution (Note: In *R*, this assumption corresponds to using the argument `family=Gamma`, and it is not necessary to specify the link function, since the reciprocal is the canonical link function for a Gamma distribution).

   Draw the curves that correspond to each fitted model on top of the scatterplot of *tempo* (vertical axis) versus log-concentrations of plasma (horizontal axis). Comment.

(e) Compar the resulting fits in the previous question. Comment and indicate which seems to be the better suited for the distribution of *tempo*, taking into account the nature and the values of that response variable, as well as the other available information.

In the following questions, consider the factor `lote`, with its two levels.

(f) Fit models with Normal and Gamma random components, and the reciprocal link function, but now crossing the numerical predictor log-concentration with the factor `lote`.

(g) Interpret the meaning of the resulting parameters, drawing the fitted curves for each lot on the time *vs.* log-concentration plot.

(h) Discuss the quality of the resulting fits.