# Basic Statistical Analysis in Life and Environmental Sciences

Chapter 3 - Binomial Models

Rodrigo Labouriau [1]

2022

[1] Applied Statistics Laboratory, Department of Mathematics, Aarhus University.

# Contents

# Chapter 3

# Binomial models

R.Labouriau [1]

Draft - Please do not circulate. [2]

This chapter covers a range of statistical models based on the binomial distribution. First we introduce a very simple example where the binomial distribution is used to estimate and compare proportions. The complexity of the models will increase along the chapter, culminating with a non-standard multiple regression binomial model. We will treat in the way classical models used in bio-statistics like: one and two-ways classification binomial models, logistic, probit and complementary log-log binomial regressions and the whole class of dose-response models.

## 3.1 Binomial models with discrete explanatory variables

The following example will serve as a motivation for introducing the binomial models with discrete explanatory variables and will be retaken and enlarged along of this and the next section.

[1] Applied Statistics Laboratory, Department of Mathematics, Aarhus University.

[2] Last revised: February, 2022. 

### 3.1.1 Simple binomial experiments

**Example 1 (Seed germination)** *In this example we study a data arising from an experiment on seed germination. Five treatments, corresponding to five different levels of amount of available water, are reported. For each of these watering levels four identical boxes were sown with* 100 *seeds each. The numbers of germinated seeds after two weeks were recorded and are displayed in the table bellow.*

| *Watering Level* | | | | |
|---|---|---|---|---|
| *1* | *2* | *3* | *4* | *5* |
| *22* | *41* | *66* | *82* | *79* |
| *25* | *46* | *72* | *73* | *68* |
| *27* | *59* | *51* | *73* | *74* |
| *23* | *38* | *78* | *84* | *70* |

*The question here is whether the watering level affects the proportion of germination. We will develop step by step a mathematical model for this experiment following the sequence below: first we model what happens in each single box; then this model is extended to describe separately the results of each of the five watering levels; and finally we consider a model that takes into account* **simultaneously** *the data with the five watering levels.*

*The seed germination in one box can be modelled in the following way. First we identify a basic binomial trial with two possible results (success or failure). Here the basic binomial trial corresponds to observing a single seed, which can germinate (success) or not (failure). Let us assume that the probability of each seed in the box to germinate is the same, say $\rho$, and that the events describing whether each seed germinated or not are independent. Under these assumptions*[3]*, the number of germinated seeds in each box can be described by a binomial distribution, as mentioned in chapter 2 and studied bellow.* □

[3]That might be unrealistic in some cases, but are used very often.

Motivated by the last example we introduce the binomial distributions. These distributions are appropriate to model counts of dichotomous (or binary) experiments and will be the basis for all the models considered in this chapter.

As we saw in chapter 2 a *basic binary trial* is a simple experiment in which there are two possible outcomes termed "success" and "failure". The choice of what is success and what is failure is completely arbitrary and will not essentially affect the models we will construct. What is considered "success" is just a matter of convention. An experiment consisting in repeating independently $n$ times a binomial trial with fixed probability of success $\rho$ is called a *binomial experiment* with $n$ repetitions and probability (of success) $\rho$. Sometimes we just refer to such experiment simply by a "binomial experiment". The number of successes occurred in a binomial experiment with $n$ repetitions and probability $\rho$ is a random quantity following a probability distribution called the *binomial distribution* with $n$ repetitions and probability $\rho$. We will briefly study this distribution next.

Let us consider a random variable $Y$ representing the results of a binomial experiment with $n$ repetitions and probability $\rho$. Clearly, $Y$ can take only the values $0, 1, \ldots, n$. The law of probability of $Y$ is specified by assigning a probability for each of the possible values that $Y$ can assume. It is easy to show that the probability that $Y$ takes the value $n$ is

$$P\left([Y = n]\right) = \rho^n \,, \tag{3.1}$$

because the event $[Y = n]$ (i.e. the event "$Y$ takes the value $n$") is equivalent to observing $n$ independent success, each of them with probability $\rho$. The probability that $Y$ takes the value 0 is

$$P\left([Y = 0]\right) = (1 - \rho)^n \,, \tag{3.2}$$

because $[Y = 0]$ is equivalent to $n$ independent failures, each of them with probability $1 - \rho$. We show next that the probability that $Y$ takes the value 1 is

$$P\left([Y = 1]\right) = n\ \rho\,(1 - \rho)^{n-1} \,. \tag{3.3}$$

First note that $Y$ takes the value 1 if we observe a success in the first trial and failure in the other $n-1$ trials. The probability of this result is $\rho(1-\rho)^{n-1}$, because the probability of one success is $p$ and the probability of $n-1$ failures is $(1-\rho)^{n-1}$. But $Y$ can takes the value 1 also if we observe success in the second trial and failure in all the other trials (including also the first trial). Again, this result has probability $\rho(1-\rho)^{n-1}$. Proceeding in this way we can show that $Y$ can take the value 1 in $n$ different ways, each of them corresponding to observing a success in one of the $n$ trials. Furthermore, the probability of each of these results is $\rho(1-\rho)^{n-1}$. We conclude that the probability of $Y$ takes the value 1 is $n$ times $\rho(1-\rho)^{n-1}$, as stated in formula (3.3).

The general formula for specifying the probability function of $Y$ is

$$P(Y=y) = \left( \begin{array}{c} n \\ y \end{array} \right) \rho^y (1-\rho)^{n-y} \text{ where } y = 0, \ldots, n \; . \tag{3.4}$$

The equation above can be proved by imitating the argument used to justify the equation (3.3), as done below. If we observe successes in the first $y$ trials and failure in the remaining $n-y$ trials ($y$ is an integer number between 0 and $n$), then $Y$ takes the value $y$. The probability of this result is $\rho^y(1-\rho)^{n-y}$. Permuting the position of the $n$ basic experiments in the previous result (ie, $y$ successes followed by $n-y$ failures) we obtain the

$$\left( \begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y!(n-y)!} \tag{3.5}$$

possible results for which $Y$ takes the value $y$.

We can state then a formal definition of the binomial distribution. A random variable $Y$ taking values in $\{0, \ldots, n\}$ with probability function given by (3.4) is said to be *binomially distributed*. In this case the distribution of $Y$ is called the *binomial distribution* and we write

$$Y \sim Bi(n, \rho) \, .$$

The number $\rho$ is sometimes referred as the *probability parameter*.

A binomially distributed random variable $Y \sim Bi(n, \rho)$ has mean and the variance given by

$$\mathrm{E}(Y) = n\rho \tag{3.6}$$

and

$$\mathrm{Var}(Y) = n\rho(1-\rho)\,. \tag{3.7}$$

[4]

**Example 2 (Seed germination, continuation)** *Let us use the notions and the notation introduced above to summarize our previous discussion on what happens in one box. Representing the number of germinated seeds in one box by the random variable $Y$, we only said up to now that $Y$ is binomially distributed with* 100 *repetitions and unspecified probability parameter $\rho$. In short $Y \sim Bi(100, \rho)$.* □

[4] Optional exercise: Prove that.

### 3.1.2 One-way classification binomial models

Up to now we considered the simple situation where independent binomial distributed random variables *with the same probability parameter* observed. We turn now to the situation where we observe independent binomial distributed variable with probability parameter not necessarily constant, but depending on a classification variable taking a given number of values. This situation is illustrated in the following example.

**Example 3 (Seed germination, cont.)** *Let us retake the example 1 given at the very beginning of this chapter. From the previous discussions of the example, it is reasonable to describe the number of germinated seeds in the four boxes submitted to the watering level* 1 *as a sample from a binomial distribution with* 100 *repetitions and a certain fixed probability parameter (i.e. germination probability). That is, according to our construction, we observed four independent random variables,* $Y_{11}, Y_{12}, Y_{13}$ *and* $Y_{14}$*, corresponding to the four repetitions. Note that we used two subindexes for labelling the results, the first indicating the watering level and the second the repetition. We assumed that*

$$Y_{11} \sim Bi(100, \rho), \ldots, Y_{14} \sim Bi(100, \rho),$$

*where* $\rho$ *was a parameter to be estimated.*

*The same argument used to construct a model for the number of germinated seeds in the boxes submitted to the watering level 1 can be applied to construct a statistical model describing the number of germinated seeds in the boxes submitted to each watering level. That is, for each fixed value of watering level, say* $w$ *(with* $w = 1, 2, 3, 4$ *and* $5$*), we observed four independent random variables,* $Y_{w1}, Y_{w2}, Y_{w3}$ *and* $Y_{w4}$*, corresponding to the four repetitions (i.e. the first subindex indicates the watering level and the second the repetition). Moreover, we assumed that*

$$Y_{w1} \sim Bi(100, \rho_w), \ldots, Y_{w4} \sim Bi(100, \rho_w),$$

*where $\rho_w$ is the germination probability of seeds submitted to the watering level $w$.*

*A statistical model for the experiment is specified by saying that the results from the $r$th box (repetition, $r = 1, \ldots, 4$) of the $w$th watering level ($w = 1, \ldots, 5$) is a realization of a random variable $Y_{wr}$ with distribution given by*

$$Y_{wr} \sim Bi(100, \rho_w)\,. \tag{3.8}$$

*Moreover, the random variables $Y_{11}, \ldots, Y_{14}, Y_{21}, \ldots, Y_{54}$ are assumed to be independent. This scenario can be represented by the following scheme:*

| | *Watering Level* | | | | |
|---|---|---|---|---|---|
| *Rep.* | *1* | *2* | *3* | *4* | *5* |
| *1* | $Y_{11} \sim Bi(100, p_1)$ | $Y_{21} \sim Bi(100, p_2)$ | $Y_{31} \sim Bi(100, p_3)$ | $Y_{41} \sim Bi(100, p_4)$ | $Y_{51} \sim Bi(100, p_5)$ |
| *2* | $Y_{12} \sim Bi(100, p_1)$ | $Y_{22} \sim Bi(100, p_2)$ | $Y_{32} \sim Bi(100, p_3)$ | $Y_{42} \sim Bi(100, p_4)$ | $Y_{52} \sim Bi(100, p_5)$ |
| *3* | $Y_{13} \sim Bi(100, p_1)$ | $Y_{23} \sim Bi(100, p_2)$ | $Y_{33} \sim Bi(100, p_3)$ | $Y_{43} \sim Bi(100, p_4)$ | $Y_{53} \sim Bi(100, p_5)$ |
| *4* | $Y_{14} \sim Bi(100, p_1)$ | $Y_{23} \sim Bi(100, p_2)$ | $Y_{34} \sim Bi(100, p_3)$ | $Y_{43} \sim Bi(100, p_4)$ | $Y_{54} \sim Bi(100, p_5)$ |

□

**The general formulation of the binomial one-way classification model**

The example above illustrates a typical situation encountered in practice: samples of binomial variables, with probability parameter depending on a classification variable (usually called treatments) are observed. We give bellow a more precise definition and set the necessary notation.

Suppose that we have a classification variable $T$ taking $t$ different values. The classification variable $T$, called a *factor*, will indicate the different possibilities of a "treatment" and we refer to those values as *levels of the factor* For each level $\tau$ of the classification variable $T$, we observe $r_\tau$ independent binomial distributed random variables with the same probability parameter, but not necessarily with the same number of repetitions. We denote by $Y_{\tau r}$ the value of the binomial random variable corresponding to the $r$th repetition at the $\tau$th level of $T$. The results can be arranged in a matrix, as bellow,

$$\begin{aligned} &Y_{11}, \quad \ldots \quad , Y_{1r_1} \\ &Y_{21}, \quad \ldots \quad , Y_{2r_2} \end{aligned} \tag{3.9}$$

$$\begin{array}{ccc} \vdots & & \vdots \\ Y_{t1}, & \ldots & , Y_{tr_t} \end{array},$$

where each line corresponds to a level of the classification variable $T$. We assume that the random variables above (i.e. in (3.9)) are independent and binomial distributed in the following way,

$$Y_{\tau r} \sim Bi(n_{\tau r}, \rho_\tau) \text{ for } \tau = 1, \ldots, t \text{ and for } r = 1, \ldots, r_\tau \ .$$

Note that we allow the number of repetitions of the binomial distribution to depend on both, the level $\tau$ of $T$ and on $r$. On the other hand, the probability parameter depends only on the level $\tau$ of the classification variable $T$.

The model described above will be referred as the *one-way binomial model*. This name comes from a straightforward analogy with the classical one-way analysis of variance model. Note that we consider the numbers of repetitions, $n_{\tau r}$, of the binomial random variables, $Y_{\tau r}$, as known. Therefore, with the notation used here, the one-way binomial model is parametrized by $t$ parameters (the number of levels of the classification variable $T$).

### Maximum likelihood estimation under the binomial one-way classification model*

[5] The maximum likelihood estimation in the one-way binomial model is very simple. Indeed, the likelihood function for the parameters based on the observed values $y_{11}, \ldots, y_{1r_1}, \ldots, y_{t1}, \ldots, y_{tr_t}$ can be written, due to independency, as the product

$$L(\rho_1, \ldots, \rho_t) = \prod_{\tau=1}^{t} \prod_{r=1}^{r_\tau} \left( \begin{array}{c} n_{\tau r} \\ y_{\tau r} \end{array} \right) \rho_\tau^{y_{\tau r}} (1-\rho_\tau)^{n_{\tau r} - y_{\tau r}} \ = \ \prod_{\tau=1}^{t} L(\rho_\tau),$$

where

$$L(p_\tau) = \left( \begin{array}{c} n_{\tau r} \\ y_{\tau r} \end{array} \right) \rho_\tau^{y_{\tau r}} (1-\rho_\tau)^{n_{\tau r} - y_{\tau r}} \ ,$$

[5] Optional reading.

coincides with the likelihood of a model where only the observations where the classification variable $T$ is equal to $\tau$ are taken into account. Taking logarithms we obtain the log-likelihood becomes

$$l(p_1, \ldots, p_t) = \sum_{\tau=1}^{t} l(p_\tau)\,,$$

where $l(p_\tau) = \log\{L(p_\tau)\}$ is the log-likelihood of a model containing only the observations at the level $\tau$ of $T$. Differentiating $l(p_1, \ldots, p_t)$ with respect to each of the $t$ parameters and equating it to zero, yields $t$ equations, each of them corresponding to one of the levels of $T$. The solution of these equations, which give the maximum likelihood estimators, was already calculated in the last section and is given by

$$\hat{p}_\tau = \frac{\sum_{r=1}^{r_\tau} y_{\tau r}}{\sum_{r=1}^{r_\tau} n_{\tau r}}\ .$$

**Example 4 (Seed germination, cont.)** *We used an one-way binomial model for modelling example 1. Here $T$ is a variable indicating the five watering levels (i.e. $t = 5$), for each level of watering we have 4 repetitions (i.e. $r_1 = \ldots = r_5 = 4$ ) and the numbers of binary trials defining each binomial distribution involved in the model are all equal to 100 (i.e. for $\tau = 1, \ldots, 5$ and $r = 1, \ldots, 4$, $n_{\tau r} = 100$).*

*The maximum likelihood estimators for the proportion of germinated seeds for each level of watering are:*

$$\begin{array}{rcrcl}
\hat{p}_1 & = & 97/400 & = & \textit{0.2425} \\
\hat{p}_2 & = & 184/400 & = & \textit{0.4600} \\
\hat{p}_3 & = & 267/400 & = & \textit{0.6675} \\
\hat{p}_4 & = & 312/400 & = & \textit{0.7800} \\
\hat{p}_1 & = & 291/400 & = & \textit{0.7275}
\end{array}$$

□

**Comparing the probability parameter of two or more binomial distributed variables and the likelihood ratio test**

Note that under the one-way binomial model we *allow* the probability parameter to take different values for the different levels of the classifying variable, however they might take the same values as well. We turn now to the question of deciding whether the probability parameters indeed take different values or whether they could be collapsed in one single value representing **the** probability parameter. If the levels of the classification variable $T$ represent different treatments, the question addressed here is whether the treatments affect the distribution of the response variable. We will use the likelihood ratio test, introduced below, to answer this question.

The idea behind the likelihood ratio test is the following. First we identify two models: In the first model, termed here the "large model", we allow the probability parameters $p_1, \ldots, p_t$ to take different values. In the second model, referred as the "reduced model", the probability parameters are assumed to take a common value $p$, i.e. $p_1 = \ldots = p_t = p$. The "large model" has $t$ parameters and therefore is more complex than the "reduced model", which has only one parameter. The likelihood ratio test is designed to test whether it is possible to reduce the "large" and more complex model to the simpler "reduced" model. The idea is that such a reduction is reasonable when the "reduced model" fits the data as well as the "large model". This is done in the following way: The ratio of the likelihood function of the "reduced model" and the likelihood function of the "large model", both evaluated at their respective maxima, measures the discrepancy between the two models. Values of this ratio close to 1 indicate that the two models are not "in disagreement", on the other hand, large values of this ratio point to "disagreement" between the two models. Equivalently, one might look at the logarithm of the ratio of the two likelihood, which is the log-likelihood of the "large model" minus the log-likelihood of the "reduced model". This positive quantity can be used to test whether there are differences among the

values of the probability parameters. That is, consider the quantity

$$\Lambda = 2\left\{l_R(\hat{p}) - l_L(\hat{p}_1, \ldots, \hat{p}_t)\right\} = 2\left\{l_R - l_L\right\} \quad ,$$

where $l_R(\hat{p}) = l_R$ and $l_L(\hat{p}_1, \ldots, \hat{p}_t) = l_L$ are the log-likelihood functions of the "reduced" and the "large model" evaluated at their maxima ($\hat{p}$ and $(\hat{p}_1, \ldots, \hat{p}_t)$ ), respectively. The quantity $\Lambda$ is called the *log-likelihood ratio statistic*.

It can be shown that, if $p_1 = \ldots = p_t$, then $\Lambda$ is approximately chi-square distributed (for values of $n$ large enough). The number of degrees of freedom $d$ of the referred chi-square distribution is given by the difference between the number of parameters of the "larger model" , $d_l$, minus the number of parameters of the "reduced model", $d_r$, i.e. $d = d_l - d_r$. Therefore, the quantity $\Lambda$ can be used to test the null hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_t,$$

at a level of significance $\alpha$, by using the rejection rule

$$\text{" Reject } H_0 \text{ when } \Lambda > \chi^2_d(1-\alpha) \text{ "}.$$

Here $\chi^2_d(1-\alpha)$ is the $(1-\alpha)$- quantil of a chi-square distribution with $d$ degrees of freedom.

**Example 5 (Seed germination, cont.)** *The log-likelihood (apart from a common constant) of the "large" and the "reduced model" evaluated at their maxima for the one-way model for the seed germination experiment are -1196.9843 and -1363.4059, respectively. That is*

$$\Lambda = 2(-1196.9843 + 1363.4059) = 2.166.4216 = 332.8432.$$

*Since the large model has* 5 *parameters and the reduced model has* 1 *parameter (i.e.* $d_l = 5$ *and* $d_r = 1$*),* $\Lambda$ *is distributed according to a chi-square distribution with* 4 *degrees of freedom (i.e.* $d = 4$*), under the null hypothesis* $H_0 : p_1 = p_2 = p_3 = p_4 = p_5$*. The probability of a chi-square distributed*

*random variable with* 4 *degrees of freedom taking values equal or superior* 332.8432 *is smaller than* 0.0001*, therefore we reject the null hypothesis of equality of the germination proportion for the different watering levels,* **and conclude that the data available supply evidence that the probability of germination is not the same for all the watering levels**.

□

### Checking homogeneity and the notion of deviance

One of the basic assumptions of the one-way binomial model is that the observations related with the same level of the classification variable are realizations of binomial random variables with *the same* probability parameter. This will be called the *homogeneity assumption* of the one-way binomial model. We will show bellow how to test the homogeneity assumption by using the likelihood ratio test.

Let us define a new model that allows the probability parameters to take different values for the different repetitions at the same level of the classification variable. This model attributes one different probability parameter for each observation and is called the *saturated model*. Now we compare the saturated model with the one-way binomial model (the one that attributes the same value of the probability parameter for the observations with the same classification) by using the likelihood ratio test. This is equivalent to test the null hypothesis

$H_0$ : "The probability parameters associated with observations with the same level of the classification variable are all equal".

The log-likelihood ratio statistic for testing the hypothesis above is given by

$$\Lambda = 2\left\{l_L - l_S\right\}, \tag{3.10}$$

where $l_L$ and $l_S$ are the log-likelihood functions of the one-way binomial model and the log-likelihood of the saturated, both evaluated at their maxima, respectively. The log-likelihood ratio statistic $\Lambda$ is approximately chi-square

distributed with degrees of freedom given by the difference of the number of observations minus the number of levels of the classification variable (i.e. the number of parameters of the saturated model minus the number of parameters of the one-way binomial model).

The log-likelihood ratio statistic $\Lambda$ given by (3.10) is called the *deviance* of the one-way binomial model. The deviance plays a fundamental rule in generalized linear models, as it will be clear from the rest of these notes. A general informal interpretation of the deviance is as follows: Since the saturated model has the same number of parameters and observations, it has the best achievable fit to the current data with the models we are using (*i.e.*, binomial distributions in the case in discussion). If we assume that at least one of the models in the class of models we are using, fits well the data, then the saturated model certainly does also. However, the saturated model is not an interesting model, since it is as complex as the raw data. A reduction of the saturated model is therefore desirable. Note that the deviance is the log-likelihood statistics of a likelihood ratio test for checking whether the current model is equivalent to the saturated model. If the current model is equivalent to the saturated model, then we may reduce the saturated model to the current model, and, since the saturated model fits the data well, we have an evidence that the current model fits the data well also (provided at least one model in the class of models we are using fits the data). This argument is completely informal, and it is advisable to always check the interpretation of the deviance in the specific model in use. Here the deviance is the test statistic used to verify the homogeneity assumption, and therefore has a very clear interpretation, but unfortunately this will not always be the case (we will return to this point many times in these notes).

**Example 6 (Seed germination, cont.)** *The saturated model in the example 1 is a model that attributes one germination probability to each box. More precisely, if the results from the $j$th box (repetition, $j = 1, \ldots, 4$) of the $t$'th watering level ($t = 1, \ldots, 5$), are represented by $Y_{tj}$, then the saturated model*

*assumes that*

$$Y_{tj} \sim Bi(100, p_{tj}) \quad , \qquad \textit{for } t = 1, \ldots, 5 \textit{ and } j = 1, \ldots 4 \ . \tag{3.11}$$

*In contrast, under the one-way binomial model we have*

$$Y_{tj} \sim Bi(100, p_t) \quad , \qquad \textit{for } t = 1, \ldots, 5 \textit{ and } j = 1, \ldots 4 \ . \tag{3.12}$$

*Note that in (3.11) (the saturated model) the germination probability $p_{tj}$ depends on both $t$ and $j$, and in (3.12) (the one-way binomial model) the germination probability $p_t$ depends only on $t$.*

*The log-likelihood evaluated at their maxima for the one-way binomial model (given by (3.12) and for the saturated model (given by (3.11) are* $1196.9843$ *and* $1177.5676$*, respectively. Therefore the deviance is given by* $2(1196.9843 - 1177.5676) = 38.8334$*. Since there are* $20$ *observations and* $5$ *parameters in this one-way binomial model, its deviance is approximately chi-square distributed with* $15$ *degrees of freedom. The probability of a chi-square distributed random variable with* $15$ *degrees of freedom takes values equal or superior to* $38.8385$ *is approximately* $0.000677432$*, indicating that the homogeneity assumption does not hold in this example! (note that this invalidates the test and estimation procedures used previously!). Indeed, an inspection of the data shows that the 3th box submitted to watering level 3 (51% germinated seeds, when the mean from this water level is 66%) and the 3th box submitted to watering level 2 (59% germinated seeds, when the mean from watering level 2 is 46%) contribute very much to the hight value of the deviance. Indeed, if these two observations are eliminated from the data the deviance becomes 15.3796 now with 13 degrees of freedom. The probability of observing a value equal or superior than 15.3796 out of a chi-square distribution with 13 degrees of freedom is 0.28426. That is the test of homogeneity has a p-value of 0.28426, therefore we do not have evidence to reject the hypothesis of homogeneity with the new data set.*

□

The discussion of the example above raises an interesting use of the deviance: it can be used in the one-way binomial model to detect the presence of lack of homogeneity. Hight values of the deviance indicate that the homogeneity assumption fails, but does not inform why it fails. The homogeneity assumption can fail because we have different populations or we are not controlling an important factor in the experiment. Another reason for failing the homogeneity assumption is that there was an error in the determination of the value of one (or a few) observation(s). In this case a single (or a few) observation(s) will contribute to inflate the value of the deviance. In fact, it is possible to calculate the deviance as a sum over the observations. Each parcel of this sum is the contribution that this observation gave to the deviance. These parcels are called the *deviance residuals* and can be calculated in most of the software for generalized linear models.

**Example 7 (Seed germination, cont.)** *The deviance residuals for the example 1 are listed below together with the number of germinated seeds and the expected proportion under the one-way binomial model.*

| *box* | *water level* | *germ. seeds* | *predicted value* | *deviance residual* |
|---|---|---|---|---|
| *1* | *1* | *22* | *0.2425* | *-0.53077* |
| *2* | *1* | *25* | *0.2425* | *0.17439* |
| *3* | *1* | *27* | *0.2425* | *0.63384* |
| *4* | *1* | *23* | *0.2425* | *-0.29340* |
| *1* | *2* | *41* | *0.4600* | *-1.00678* |
| *2* | *2* | *46* | *0.4600* | *0.00000* |
| *3* | *2* | *59* | *0.4600* | **2.60499** |
| *4* | *2* | *38* | *0.4600* | *-1.61569* |
| *1* | *3* | *66* | *0.6675* | *-0.15890* |
| *2* | *3* | *72* | *0.6675* | *1.13088* |
| *3* | *3* | *51* | *0.6675* | **-3.24806** |
| *4* | *3* | *78* | *0.6675* | **2.47491** |
| *1* | *4* | *82* | *0.7800* | *0.98872* |
| *2* | *4* | *73* | *0.7800* | *-1.17749* |
| *3* | *4* | *73* | *0.7800* | *-1.17749* |
| *4* | *4* | *84* | *0.7800* | *1.50320* |
| *1* | *5* | *79* | *0.7275* | *1.44218* |
| *2* | *5* | *68* | *0.7275* | *-1.04918* |
| *3* | *5* | *74* | *0.7275* | *0.28212* |
| *4* | *5* | *70* | *0.7275* | *-0.61149* |

*Note that 3th boxes of watering level 2 and 3 give the larger contribution to the deviance. They present also a large discrepancies between the observed number of germinated seeds and the expected proportions.*

□

### 3.1.3 Two-ways binomial classification models

We study now a model suitable for the situation where two classification variables are supposed to affect the response of a binomial distributed variable.

Let us consider first two examples.

**Example 8 (Cancer incidence)** *This example refers to the occurrence of a certain type of cancer. The table bellow shows the incidence of the cancer in two towns, classified into six age categories. The age categories are: from 15 to 24, from 25 to 34, from 35 to 44, from 45 to 54, from 55 to 64 and from 65 to 74 years. The two towns will be referred as: "town 0" and "town 1". The two classification variables in this example will be referred as "age" and "town". The table presents also the population of each town, classified also according to the age.*

| *Town 0* | | | *Town 1* | | |
|---|---|---|---|---|---|
| *Age categories* | *Number of cases* | *Population (P)* | *Age categories* | *Number of cases* | *Population (P)* |
| *15-24* | *1* | *172675* | *15-24* | *4* | *181343* |
| *25-34* | *16* | *123065* | *25-34* | *38* | *146207* |
| *35-44* | *30* | *96216* | *35-44* | *119* | *121374* |
| *45-54* | *71* | *92051* | *45-54* | *221* | *111353* |
| *55-64* | *102* | *72159* | *55-64* | *259* | *83004* |
| *65-74* | *130* | *54722* | *65-74* | *310* | *55932* |

*We will denote the number of cases of cancer and the population of the town $t$ at the age category $a$ by $X_{ta}$ and $N_{ta}$, respectively.*

*These data can be modelled by assuming the number of cases $X_{ta}$ in the town $t$ at the age $a$ binomially distributed with a certain probability parameter (the probability of having the cancer at the respective age in the respective town) and with a number of trials equal to the population $N_{ta}$. That is, for each town $t$ and age category $a$,*

$$X_{ta} \sim Bi(N_{ta}, p_{ta}) \ .$$

*Note that we assumed here that the probabilities $p_{ta}$ of having cancer depend on both the town and the age category. This model has one parameter (the* 12 *probability parameters) for each observation (the* 12 *combinations of town*

*and age category). Therefore it is called the* saturated model. *The idea is to study the possibility of simplifying this basic model, i.e. construct a model with less parameters.*

□

**Example 9 (Seed germination, extended)** *The example 1 presented in the beginning of this chapter was indeed only partial results of an experiment. The seed germination experiment described there was performed in two different ways: covering the box or not. We refer to that as the box treatment. For each of the box treatments 5 watering levels were used. For each combination of box treatment and watering level 4 repetitions were observed. The data analyzed in the example 1 just the results for the covered box. The table bellow presents the results of the whole experiment. Note that one of the repetitions of the watering level 5 of the uncovered boxes is missing (this was indicated by using the symbol ".", which indicates that these value was not observed).*

| *Box treatment* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Covered* | | | | | *Not Covered* | | | | |
| *Watering level* | | | | | *Watering level* | | | | |
| *1* | *2* | *3* | *4* | *5* | *1* | *2* | *3* | *4* | *5* |
| *22* | *41* | *66* | *82* | *79* | *45* | *65* | *81* | *55* | *31* |
| *25* | *46* | *72* | *73* | *68* | *41* | *80* | *73* | *51* | *36* |
| *27* | *59* | *51* | *73* | *74* | *42* | *79* | *74* | *40* | *45* |
| *23* | *38* | *78* | *84* | *70* | *43* | *77* | *76* | *62* | . |

*Clearly, we have in this example two classification variables (the box treatment and the watering level) and a response that can be reasonably modelled by using the binomial distribution. Here the* saturated model, *specifies that for each watering level $w$, each box treatment $t$, and each repetition $r$, the number of germinated seeds (out of 100 seeds placed in the box) is binomially distributed as specified bellow:*

$$Y_{wtr} \sim Bi(100, p_{wtr})\,.$$

*A simpler and more interesting model assumes that the probability parameters are equal for the observations with the same watering level and the same box treatment, i.e. for each watering level $w$, each box treatment $t$, and each repetition $r$,*

$$Y_{wtr} \sim Bi(100, p_{wt}) \,.$$

*This model is called the full two-ways binomial model.*

□

The examples 8 and 9 above illustrate a situation were we have two classifying variables (with a certain number of levels) possibly affecting the probability parameter of a binomially distributed variable. We call this model the *two-ways classification binomial model* or simply the *two-ways binomial model*. Let us introduce some notation for the two-way classification model. Suppose that there are two classification variables, $A$ and $B$, taking $a$ and $b$ values respectively. For each value $\alpha$ ($\alpha = 1, \ldots, a$) of $A$ and each value $\beta$ ($\beta = 1, \ldots, b$) of $B$ we observe $R_{\alpha\beta}$ independent realizations of a binomially distributed variables with the same probability parameter. That is, for $\alpha = 1, \ldots, a$, $\beta = 1, \ldots, b$ and $r = 1, \ldots, R_{\alpha\beta}$, we observe the random variable $Y_{\alpha\beta r}$ distributed as

$$Y_{\alpha\beta r} \sim Bi(n_{\alpha\beta r}, p_{\alpha\beta}) \,. \tag{3.13}$$

This model has $a.b$ parameters and will be referred as the *full two-ways binomial model* or the *two-ways binomial model with interactions.* [6]

The maximum likelihood estimate for the probability parameters of the two-ways binomial model are the observed proportions, i.e for $\alpha = 1, \ldots, a$ and $\beta = 1, \ldots, b$, the maximum likelihood estimate $\hat{p}_{\alpha\beta}$ of $p_{\alpha\beta}$ is

$$\hat{p}_{\alpha\beta} = \frac{\sum_{r=1}^{R_{\alpha\beta}} Y_{\alpha\beta r}}{\sum_{r=1}^{R_{\alpha\beta}} n_{\alpha\beta r}} \,. \tag{3.14}$$

[6] Optional exercise: Please, identify the elements of this general definition in the examples 8 and 9. More precisely, specify that one of the classification variables is $A$ and the other is $B$, then find out what is "$a$, $b$, $R_{\alpha\beta}$, $n_{\alpha\beta r}$ and $p_{\alpha\beta}$".

The argument to this is just a variation of the procedure we have already done in the one-way binomial model and is left as an exercise. [7]

We assumed that the observations corresponding to the same levels of the classification variables are realizations of binomial variables with *the same* probability parameter. This is called the *homogeneity assumption* and is analogous to the homogeneity assumption used in the one-way binomial model. As in the case of the one-way binomial model, this assumption can be tested, provided there are repetitions. The idea is essentially the same: define a saturated model and use the likelihood ratio test for testing the possibility of reducing the saturated model to the full two-way classification model. The details are given below.

The *saturated model* is a statistical model that attributes one specific (not necessarily equal) probability parameter for each observation. This can be expressed precisely in the context of the two-ways binomial model as follows. For $\alpha = 1, \ldots, a$, $\beta = 1, \ldots, b$ and $r = 1, \ldots, R_{\alpha\beta}$,

$$Y_{\alpha\beta r} \sim Bi(n_{\alpha\beta r}, p_{\alpha\beta r}) \ . \tag{3.15}$$

Note that the difference between (3.13) and (3.15) is that in the first the probability parameters does not depend on the subindex $r$ and in the second they do depend. We will use the saturated model as a reference model for testing the homogeneity assumption. For that we will use a likelihood ratio test. The log-likelihood statistic for this test is called the deviance. More precisely, the *deviance* is defined as

$$D = 2(l_F - l_S) \ , \tag{3.16}$$

where $l_F$ and $l_S$ are respectively the log-likelihood functions of the full two-way binomial and the saturated model, both evaluated at their respective

[7] Optional exercise: Use the expression (3.14) to give the maximum likelihood estimate for the parameters of the full model for the examples 8 and 9. Give an argument to show that the maximum likelihood estimate for $p_{\alpha\beta}$ is like given in (3.14) (hint: you do not need to explicitly calculate the score function and solve the score equation; try to use an argument using the results we have already obtained before).

maxima. According to the theory of likelihood ratio test, the deviance is approximately distributed according to a chi-square distribution with the number of degrees of freedom given by the difference of the number of parameters of the two models, that is $\{\sum_\alpha^a \sum_\beta^b R_{\alpha\beta}\} - a.b.$ [8]

In the case were there is only one repetition for each combination of the classification variables $A$ and $B$ (as in the example 8), then the full model coincides with the saturated model and the test described above is meaningless. Indeed the test cannot be done, since the deviance provided by (3.16) is zero. [9]

We consider next some simpler alternative models to the full two-ways binomial model. First we can think that the probability parameter changes independently according to the classification with respect to each classification variables. In mathematical terms this situation can be described in the following way. Suppose that for $\alpha = 1, \ldots, a$, $\beta = 1, \ldots, b$ and $r = 1, \ldots, R_{\alpha\beta}$, $Y_{\alpha\beta r}$ distributed as

$$Y_{\alpha\beta r} \sim Bi(n_{\alpha\beta r}, p_{\alpha\beta}) \ , \tag{3.17}$$

where

$$p_{\alpha\beta} = A_\alpha + B_\beta \ . \tag{3.18}$$

That is, the probability of success $p_{\alpha\beta}$ can be then expressed as a sum of two quantities: one depending only on the classification according to the classification variable $A$ (namely $A_\alpha$) and one depending only on the classification according to the classification variable $B$ (namely $B_\beta$). Here $A_1, \ldots, A_a$, $B_1, \ldots, B_b$ are parameters (we have then $a + b$ parameters and $a.b$ observations). This model is called the *additive two-ways binomial model*.

[8] Optional exercise: Please, describe the test referred here in details. That is, write explicitly the null hypothesis, the alternative hypothesis and the rejection rule (i.e. a rule saying explicitly when to reject the null hypothesis) for the likelihood ratio test described here. Hint: Just mimic what we did in the one-way binomial model.

[9] Optional exercise: Please, try to write what would be the null and the alternative hypothesis of a test like the test for the homogeneity assumption.

**Example 10 (Seed germination, extended)** *We describe the example 9 in a suitable way for specifying the additive model. The full two-ways binomial model states that for each watering level $w$, each box treatment $t$, and each repetition $r$,*

$$Y_{wtr} \sim Bi(100, p_{wt})\,. \tag{3.19}$$

*The additive two-ways binomial model assumes that the probability parameters used in (3.19) can be written in the form*

$$p_{wt} = W_w + T_t \ \text{ , for } w = 1, \ldots, 5 \text{ and } t = 1, 2, \tag{3.20}$$

*where $W_1, \ldots, W_5, T_1, T_2$ are parameters. Note that (3.20) implies a restriction in the model, as compared to the full two-ways model. Indeed, the additive model has $5 + 1 = 6$ parameters and the full model has $5.2 = 10$ parameters.*

□

[10]

There are two important hypothesis tests that can be done with the additive two-ways binomial model: we can test the model reduction from the full model to the additive model, or we can test the model reduction directly from the saturated model to the additive model. Both tests can be done by using a suitable likelihood ratio test, but they have very different interpretations. We discuss both tests below.

Testing the model reduction from the full model to the additive model corresponds to test the special pattern in the probability parameter defined by the additive model, *given the full model*. That, is assuming the full model, we test the reduction to the additive model. This test can be done by looking to the log-likelihood ratio statistic of a likelihood ratio test, given by

$$\Lambda = 2(l_A - l_F)\ ,$$

[10] Optional exercise: Please, describe the additive two-ways binomial model for the example 8 in similar way as we did for the example 9 in example 10.

where $l_F$ and $l_A$ are respectively the log-likelihood functions of the full two-ways binomial and the additive two-ways binomial model, both evaluated at their respective maxima. The quantity $\Lambda$ is approximately distributed according to a chi-square distribution with $a.b-(a+b)$ degrees of freedom.

On the other hand, testing the model reduction from the saturated model to the additive corresponds to test simultaneously the homogeneity assumption and the pattern imposed by the restrictions of the additive two-ways binomial model. This test can be done by a likelihood ratio test using the following log-likelihood ratio statistic

$$D_S = 2(l_A - l_S) \ . \tag{3.21}$$

Here $l_A$ and $l_S$ are respectively the log-likelihood functions of the additive two-way binomial and the saturated model, both evaluated at their respective maxima. The quantity $D$ defined in (3.21) is called the *deviance* of the additive two-way binomial model. The deviance is approximately distributed according to a chi-square distribution with the number of degrees of freedom given by the difference of the number of parameters of the two models, that is $\{\sum_\alpha^a \sum_\beta^b R_{\alpha\beta}\} - (a+b)$.

The deviance of the full and the additive two-ways binomial models, together with their respective related degrees of freedom, are supplied by most of the modern statistical software that can treat generalized linear models. This information is enough to test the model reduction from the full to the additive model. To see that, note that the difference of degrees of freedom associated with the two models gives precisely the difference of number of parameters the two models. Moreover,

$$\begin{aligned} D_S - D &= 2(l_A - l_S) - 2(l_F - l_S) \\ &= 2(l_A - l_S - l_F + l_S) = 2(l_A - l_F) = \Lambda \ , \end{aligned}$$

i.e., the log-likelihood ratio statistic $\Lambda$ is equal to the the deviance of the full minus the deviance of the additive model. We then construct the following table which is analogue to the classic ANOVA table.

Table of deviance

| Source | DF | Deviance | $\chi^2$ |
|---|---|---|---|
| Full model | $DF_F$ | $D$ | - |
| Additive model | $DF_A$ | $D_A$ | $D - D_A = \Lambda$ |

Here $DF_F$ and $DF_A$ are the degrees of freedom associated with the deviance of the full and the additive model (i.e the difference between the number of observations and the number of parameters of the respective models). This table is used to summarise the results of a statistical analysis and will be extended when we consider further models in this section. We will call this table (and its extensions) the *table of deviance*.

**Example 11 (Seed germination, extended, cont.)** *The table of deviance for the example 9*

*Table of Deviance*

| *Source* | *DF* | *Deviance* | $\chi^2$ |
|---|---|---|---|
| *Full model* | *29* | *63.3093* | - |
| *Additive model* | *33* | *325.4153* | *262.11* |

*Note that* $\Lambda = 262.11$ *is chi-square distributed with 4 degrees of freedom under the additive model. Since this value is too high, we reject the null hypothesis of equality of the additive and the full model and conclude that the full model cannot be reduced to the additive model. Figure 3.1 illustrates the absence of additivity detected in the test above. 3.1.*

□

In the last example we could not simplify further the full two-ways binomial model. An analogous procedure would yield the same conclusion to the case of the case of the example 8. That is, the two classification factors do not act additively on the probability parameters. However, as we will show, it is possible to simplify the full model in the example 8! The idea is

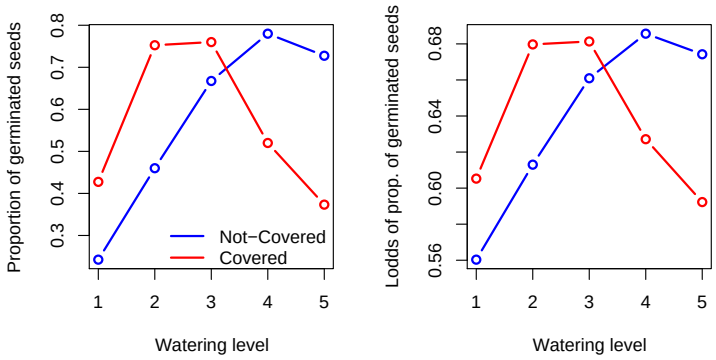


Figure 3.1: Proportion and logistic transformed proportion of seed germination for covered and not covered box and different levels of watering.

to transform (i.e. apply a certain function to) the probability parameters. In some cases one can have a good surprise by doing that and obtain that the two classification variables act additively in the new transformed parameters, yielding a substantial simplification of the model. We will define next a model that will give this simplification to the example 8, but not to the example 9.

The *additive logistic two-way binomial model* is defined as follows. For $\alpha = 1, \ldots, a$, $\beta = 1, \ldots, b$ and $r = 1, \ldots, R_{\alpha\beta}$, $Y_{\alpha\beta r}$ distributed as

$$Y_{\alpha\beta r} \sim Bi(n_{\alpha\beta r}, p_{\alpha\beta}) \ , \tag{3.22}$$

where

$$logit(p_{\alpha\beta}) = \log\left(\frac{p_{\alpha\beta}}{1-p_{\alpha\beta}}\right) = A_\alpha + B_\beta \ . \tag{3.23}$$

Here $A_1, \ldots, A_a$, $B_1, \ldots, B_b$ are parameters. Note that, although we used the same notation, these parameters have a completely different interpretation when compared to the the parameters used in (3.18) to define the additive two-ways binomial model. The interpretation of the *logit* transformation used above is the following: given an event with probability $p$ the odds of this events are $p/(1-p)$, that is, the probability $p$ of the event happens divide by the probability $(1-p)$ that the event does not happen. Large values of the odds indicate that the event is likely to happen. The lods are given by the logarithm of the odds, and have a similar interpretation as the odds (i.e. they are just the odds expressed in another scale). The equation (3.23) says that the effect of the classification variables are additive in the lods scale instead of in the probability scale as before). The function "$logit(\ \cdot\ ) = \log\left\{\frac{(\ \cdot\ )}{1-(\ \cdot\ )}\right\}$" is called the *link function* of the model. Other functions could be used (but this would result in other models).

The way of estimating the parameters, defining deviance and testing reductions to nested models is completely similar to the way we proceeded in the previous models.

**Example 12 (Cancer incidence, cont.)** *In the example 8, using the additive logistic two-way binomial model we obtain the following table of deviance:*

*Table of Deviance*

| *Source* | *DF* | *Deviance* | $\chi^2$ |
|---|---|---|---|
| *Full model* | *0* | *0* | - |
| *Additive model* | *5* | *3.21* | *3.21* |

*The probability of a chi-square distributed random variable with 5 degrees of freedom assumes a value greater or equal 3.21 is 0.6672. Therefore we have no evidences for rejecting the reduction from the full model to the additive logistic two-way binomial model. This confirms what we see in figure 3.2 where the logistic transformed incidences are shown for the two towns and the different age categories. The two connected "curves" (connecting the observations in successive age categories made in the same town) are parallel. This illustrates the kind of restriction we do by assuming the additive model. Note that this parallelism is not observed in the pure additive two-ways binomial model, as illustrated in figure 3.2.*

□

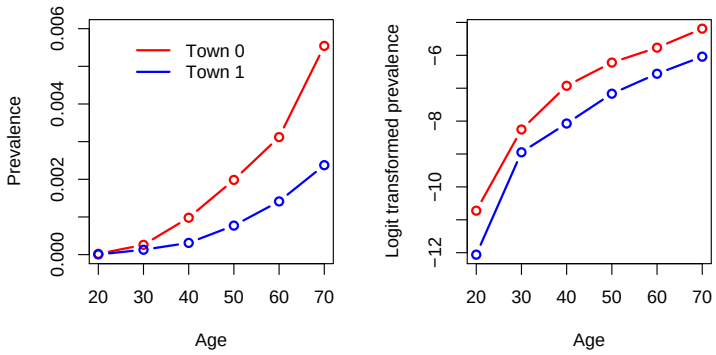


Figure 3.2: Prevalence and logistic transformed prevalence of cancer in the two towns classified by age .

## 3.2 Binomial models with continuous explanatory variables

We considered in the last section models for binomial responses where the explanatory variables taking a finite number of values. In this section we will study models for binomial responses where the explanatory variables are continuous. The following example illustrates this kind of models.

**Example 13 (Mice convulsion)** *The table below reports partial result of an experiment with mice (from Hemmingsen and Krogh, 1926). Several doses of insulin were applied to a number of mice. The main questions here is whether the different amounts of insulin given affect the proportion of animals with convulsion.*

| *Dose* | *Number with convulsion* | *Total number observed* |
|---|---|---|
| *8.5* | *14* | *37* |
| *10.5* | *18* | *40* |
| *13.0* | *21* | *37* |
| *18.0* | *23* | *31* |
| *21.0* | *30* | *37* |
| *28.0* | *27* | *30* |

*Denote the number of mice for which we applied the dose $D$ of insulin ($D = 8.5, 10.5, \ldots, 28.0$) by $N_D$ and the number of animals presenting convulsions by $Y_D$. It is natural to assume that the number of mice with convulsion, for which a certain doses $D$ of insulin was applied, is binomially distributed with $N_D$ repetitions and probability parameter (i.e. probability of having convulsions) depending (at least in principle) on the insulin doses, say $p_D$. More precisely, we assume that for $D = 8.5, 10.5, \ldots, 28.0$,*

$$Y_D \sim Bi(N_D, p_D)\,.$$

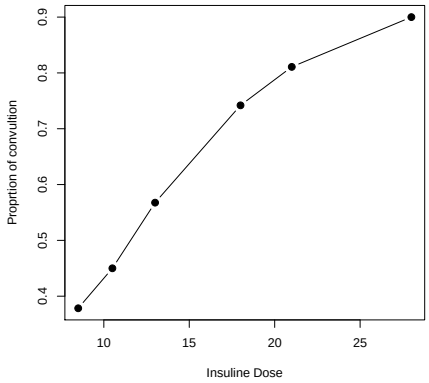


Figure 3.3: Proportion of mice with convulsion against the doses of insulin. The circles represent the observed proportions, the smooth curve drawn with a standard technique for drawing smooth curves (kernel estimator of a non-parametric regression, i.e. no specific form is assumed *a priori*).

*The proportion of mice with convulsion under the several doses of insulin is represented in Figure 3.3. It is clear from the table and Figure 3.3 that the proportion of mice presenting convulsion increased with the doses of insulin. Moreover, the figure suggests that the proportion of animals presenting convulsion increases in a smooth way as a function of the doses. One way to express this idea is to imagine that there is a smooth function that associates each dose with a probability of having convulsion. We cannot observe all the possible doses (that can be thought as varying continuously in an interval), but only some of them (the ones we used in the experiment). However, if the number of observed doses is large enough, then we will be able to observe some regularity: the points in a plot of the probabilities against the doses will be disposed along a smooth curve (the graph of the function). There is a further complication: we cannot observe probabilities. But we can use the proportion as a first approximation to the probability (we will present better ways to do that latter), keeping in mind that there is a random variation in*

*the proportions. Therefore, the pattern observed in the plot of the observed proportions of animals with convulsion against the doses (figure 3.3) indicates that it is reasonable to model the dependence of the probability of convulsion on the doses by saying that there is a smooth function associating the doses to the probabilities. We will study along this section several attempts to specify such a function.*

*It will be convenient to consider the logarithm of the insulin doses, instead of the doses themselves. The proportion of animals with convulsion plotted against the logarithm of the doses are shown in Figure 3.4 For simplicity*

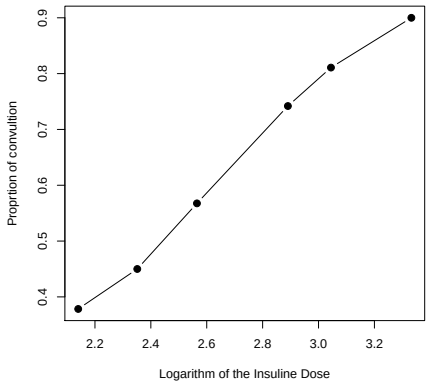


Figure 3.4: Proportion of mice with convulsion against the logarithm of the doses of insulin.

*we refer to the logarithm of the insulin doses as the log-doses and use the symbol* $L = \log(D)$ *to represent this (deterministic) variable. The idea of the models will discuss in the rest of this section is to represent the probability of an animal having convulsion as a continuous and increasing function of the log-doses plus a constant. This function will be called the* response function. *Several choices for the response function will be considered. This idea can be expressed in symbols as, for each possible value* $D$ *of the doses,*

$$p_D = g\left(\alpha + \beta \log(D)\right) = g(\alpha + \beta L), \tag{3.24}$$

*where $g$ is a given function, the response function. Here $\alpha$ and $\beta$ are parameters of of the model that must be estimated. In order to specify the model one has to specify the response function $g$. For example, specifying that the response function is given by the identity function $g(\,\cdot\,) = (\,\cdot\,)$ (i.e. the function that associates each number to it self) gives*

$$p_D = \alpha + \beta L, \tag{3.25}$$

*which says that the probability of convulsion depends linearly on the log-doses.*

*Note that when we specify in (3.24) the dependence of the probability of convulsion $p_D$ on the explanatory variable $L = \log(D)$, nothing is said about the values that $L$ (or equivalently $D$) are taking in the particular data in play. That is, (3.24) should hold for any (reasonable) value of $L$, in particular the values present in the data set in study, but also for any other value of $L$. This is essentially different than the models with classification explanatory variables studied in the last section.*

*The linear model defined above can only be a rough approximation, since the probability $p_D$ must be a number between 0 and 1 and $\alpha + \beta \log(D)$ is not bounded (provided $\beta \neq 0$). A simple inspection of Figure 3.4 should convince the reader that indeed the probabilities $p_D$ do not depend linearly on the log-doses. We need then to consider other candidates for the response function.*

*One natural (and often used in practice) response function for binomial models is the logistic function given by*

$$logist(\,\cdot\,) = \frac{\exp(\,\cdot\,)}{1 + \exp(\,\cdot\,)}.$$

*The logistic function is a S-shaped function ranging between 0 and 1. Using the formula (3.24) with $g$ equal the logistic function yields*

$$p_D = logist(\alpha + \beta L) = \frac{\exp\{\alpha + \beta L\}}{1 + \exp\{\alpha + \beta L\}}. \tag{3.26}$$

*This is the so called* logistic model. *Since the range of the logistic function are the numbers between 0 and 1, the logistic model does not have the problems*

*of range of probabilities presented by the linear model. That is, the right side of (3.25) is unbounded (which shows that the linear model can only hold in a narrow interval of variation of the explanatory variable $L$) and, on the other hand, the right side of (3.26) is always between 0 and 1 (which shows that the logistic model is not incompatible with a relative large range of variation of the explanatory variable $L$).*

*Even though the logistic model is not incompatible with the range of variation of the explanatory variable, it might not reflect the dependence of the proportions of mice with convulsion on the insulin doses in a correct way. One way to verify the adequacy of the logistic response function is to transform the probabilities $p_D$ by the inverse of the logistic function, called the logit function, and check whether the logit transformed probabilities are a linear function of the explanatory variable ($\log(D)$) (see figure 3.5). More precisely, define the logit function by*

$$logit(\,\cdot\,) = \log\left\{\frac{(\,\cdot\,)}{1-(\,\cdot\,)}\right\}.$$

*It is easy to see that the logit function is the inverse of the logistic function. That is, $logit\{logist(p)\} = p$ and $logist\{logit(p)\} = p$.*

*Using the logit function we can re-express the logistic model in the following way:*

$$\begin{aligned} logit(p_D) &= logit\,\{logist(\alpha+\beta L)\} \\ &= \alpha+\beta L\,. \end{aligned}$$

*That is, the logistic model assumes that the logit transformed probabilities of convulsion depend linearly on the explanatory variable $L$. Moreover, since the logit transformed probabilities are the logarithmic transformed odds* [11]*, the lodds, the logistic model assumes that the lodds of convulsion increase*

[11] Recall that the odds of an event with probability $p$ is defined as $p/(1-p)$, that is the probability of the event occurs divided by the probability of the event does not occur. The lodds are just the logarithm of the odds.

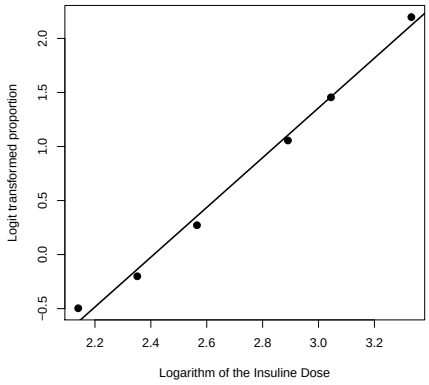


Figure 3.5: Logit transformed proportion of mice with convulsion against the logarithm of the doses of insulin.

*proportionally to the explanatory variable $L$. An alternative interpretation of the logistic model will be discussed when we study the dose-response models.*

*A third model can be obtained by using the response function*

$$f(\,\cdot\,) = 1 - \exp\{-\exp(\,\cdot\,)\}\ .$$

*Like the logistic function, this is a S-shaped function with range of variation between 0 and 1. The model obtained by using this function as response specifies that the dependency of the probabilities of convulsion on $L$ is in the form*

$$p_D = f(\alpha + \beta L) = 1 - \exp\{-\exp(\alpha + \beta L)\}\ .$$

*The inverse of the response function $f$ is*

$$CLL(\,\cdot\,) = \log\{-\log(1 - \,\cdot\,)\}\ .$$

*Therefore, the model can be characterized alternatively by*

$$CLL(p_D) = \alpha + \beta L\,.$$

*The quantity $CLL(p_D)$ is called the complementary log-log transformation of the probability $p_D$, since $1-p_D$ is the complementary of $p_D$, in the sense that $1-p_D$ is the probability of not observing convulsion. The model described above says that the increase in the complementary log-log transformed probability of having convulsion when a certain dose of insulin is applied is proportional to the dose applied. Therefore this model is called the* complementary log-log model. *Figure 3.6 displays a plot of the complementary log-log transformed proportion of mice presenting convulsions against the log-dose. Note that the points in the plot are approximately disposed along a straight line, indicating that the complementary log-log model is reasonable for modelling the present data. We will return to this point latter.*

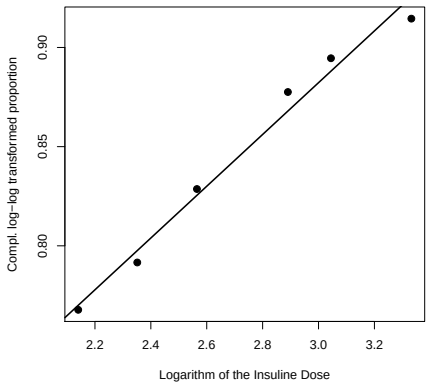


Figure 3.6: Complementary log-log transformed proportion of mice with convulsion against the logarithm of the doses of insulin.

*Many other models could be proposed for modelling the effect of the doses of insulin in the proportion of convulsion. Naturally, some of these many models will describe reasonably the data, and some will not fit at all. It is important to remark that there is no uniformly better model. Indeed, there will be always a range of models that are indistinguishable from the statistical point of view, for a certain data. Choosing a good model between many reasonable*

*models, or sometimes just finding an acceptable model, is an art that often involves knowledge and experience in both statistics and the theory behind the phenomena in study.* □

The example above motivates us to introduce a class of models, called *binomial regression models* for describing the dependence of a binomial distributed variable on a continuous explanatory variable.

More precisely, suppose that we observe $k$ independent binomial distributed variables $Y_1, \ldots, Y_k$ with known numbers of repetitions $n_1, \ldots, n_k$ respectively. This is a typical output of an experiment where one counts the occurrence of a certain phenomenon out of a known number of trials. Moreover, assume that we observe together with each variable $Y_i$ a continuous variable, $L_i$ $(i = 1, \ldots, k)$ that supposedly affects the probability parameter of the binomially distributed variable $Y_i$. Here $L_1, \ldots, L_k$ are assumed to be deterministically known, i.e. they are *not* random variables (!). Typically the variables $L_i$s represent treatments with continuous variation and can be ordered. We call the variable taking values $L_i$ the *explanatory variable*. More precisely, for $i = 1, \ldots, k$,

$$Y_i \sim Bi(n_i, p_i)\,.$$

Additionally, it is assumed that there is a function $g$, called the *response function*, such that

$$p_i = g\left(\alpha + \beta L_i\right). \tag{3.27}$$

It is convenient to introduce the inverse of the response function $g$, denoted $h = g^{-1}$, [12] and called the *link function*. We can alternatively characterize the dependence of the probability parameters $p_i$, for $i = 1, \ldots, k$, as bellow:

$$h(p_i) = \alpha + \beta L_i\,. \tag{3.28}$$

[12] That is, $h(g(\,\cdot\,)) = (\,\cdot\,)$ and $g(h(\,\cdot\,)) = (\,\cdot\,)$. The existence of the inverse is ensured by the strict monotonicity of the response function.

That is, the binomial regression model assumes that the probability parameters transformed through the link function, $h(p_i)$s, depend on the values of the explanatory variable in a linear form.

For completely characterizing the binomial regression model it is necessary to specify the response function or equivalently the link function. The table bellow shows some choices of response functions and the correspondent link functions (i.e. the inverses)

| Model | Response function | Link function |
|---|---|---|
| Linear | $g(\,\cdot\,) = (\,\cdot\,)$ | $h(\,\cdot\,) = (\,\cdot\,)$ |
| Logistic | $g(\,\cdot\,) = \frac{\exp(\,\cdot\,)}{1+\exp(\,\cdot\,)}$ | $h(\,\cdot\,) = \log\left\{\frac{(\,\cdot\,)}{(1-\,\cdot\,)}\right\}$ |
| Probit | $g(\,\cdot\,) = \Phi(\,\cdot\,)$ | $h(\,\cdot\,) = \Phi^{-1}(\,\cdot\,)$ |
| Complementary log-log | $g(\,\cdot\,) = 1 - \exp\{-\exp(\,\cdot\,)\}$ | $h(\,\cdot\,) = \log\{-\log(1-\,\cdot\,)\}$ |

13

The table above lists only some classic alternatives for the response function for binomial regression models. In principle any smooth (at least differentiable with continuous derivative) and strictly monotone (i.e. strictly increasing or decreasing) function can be used as a response function. Most

[13] Here the function $\Phi$ is the accumulative distribution function of the standard normal distribution, i.e. if $X$ is a normally distributed with mean 0 and variance 1, then for each real number $t$, $\Phi(t)$ is the probability of $X$ being less or equal $t$.

of the software for generalized linear model allow the user to specify your own link and response function.

Once defined the binomial regression model we might use the standard likelihood based procedures for studying the model, similarly to the way we studied the binomial models with classification explanatory variables. That is, we might define a likelihood function depending on the models parameters, $\alpha$ and $\beta$, calculate the log-likelihood, calculate the score function by differentiating the log-likelihood with respect to each of the parameter and find the maximum likelihood by equating the score function to zero (or by using some iterative numerical algorithm). We do not need to treat these calculations in details here. What is important here is to realize that all the quantities related to the maximum likelihood estimation that we used before for the binomial models with classification explanatory variables can be calculated here also in an analogous way (but involving some slightly more complicated calculations). We can talk about maximum likelihood estimate of the parameters $\alpha$ and $\beta$, the value of the log-likelihood at the maximum, use likelihood ratio tests for sub models, etc.

Let us discuss now the notion of residuals for the binomial regression models. Roughly speaking the residuals are the differences between what we observe and what we would expect under the current model. The residuals are used to verify the adequacy of the model. The details are as follows: Once estimated the parameters $\alpha$ and $\beta$ by the maximum likelihood estimate, say $\hat{\alpha}$ and $\hat{\beta}$, we may estimate the probability parameter for the $i$th observation (for $i = 1, \ldots k$) by

$$\hat{p}_i \;=\; g\left(\hat{\alpha} + \hat{\beta} L_i\right),$$

and the expected value that this observation would have (under the binomial regression model) by

$$E(Y_i) = n_i\, \hat{p}_i = n_i\, g\left(\hat{\alpha} + \hat{\beta} L_i\right).$$

We can then define the $i$ the *raw residual* by

$$R_i = Y_i - n_i\, g\left(\hat{\alpha} + \hat{\beta} L_i\right) = Y_i - n_i\, \hat{p}_i\,.$$

Note that, for a fixed sample (i.e. given the observed data) $g\left(\hat{\alpha}+\hat{\beta}L_i\right)$ is a constant. Moreover, under the binomial regression model with the parameters $\alpha$ and $\beta$ set at their maximum likelihood estimates, $\hat{\alpha}$ and $\hat{\beta}$, $E(Y_i)=n_i\hat{p}_i$ and $Var(Y_i)=n_i\hat{p}_i(1-\hat{p}_i)$. Hence

$$E(R_i)=0 \quad \text{and} \quad Var(R_i)=n_i\hat{p}_i(1-\hat{p}_i)\,.$$

Next we will standardize the residuals, obtaining a quantity that has constant variance. This is convenient for checking the model. The way of standardizing the raw residuals is to divide them by the square root of their variances. This residuals are called the *Pearson residuals* and are defined in precise terms as, for $i=1,\ldots,k$,

$$P_i=\frac{Y_i-n_i\,\hat{p}_i}{\sqrt{VarY_i}}=\frac{Y_i-n_i\,g\left(\hat{\alpha}+\hat{\beta}L_i\right)}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}\,.$$

Clearly,

$$E(P_i)=0 \quad \text{and} \quad Var(P_i)=1\,. \tag{3.29}$$

We can use the Pearson residuals to verify the model. Indeed, a plot of the Pearson residuals against the predicted values should present a scatter of points showing no clear patterns, indicating that the variance of the raw residuals behaves as we might expect under the binomial regression model. Moreover, the plot of the Pearson residuals against the explanatory variables should not show any kind of pattern. The advantage of the Pearson residuals over the raw residuals is that Pearson residuals are automatically made smaller in the places were the variance would be expected to be larger due to the proprieties of the binomial distribution. Therefore the Pearson residuals would have less tendency to show false patterns due to the large variance.

**Example 14 (Mice convulsion, cont.)** *The plot of the Pearson residuals against the log-dose for the binomial regression models presented in example 13 are presented in the Figures: 3.7 for the linear model, 3.8 for the logistic model and 3.9 for the complementary log-log model.*

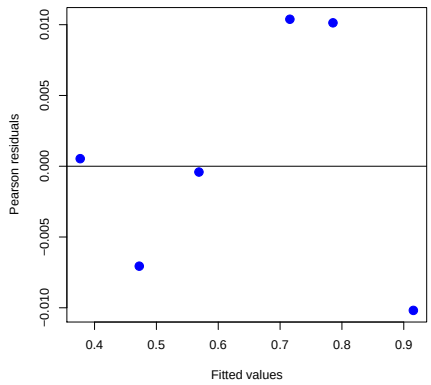


Figure 3.7: Pearson residuals against the log-dose for the linear model.

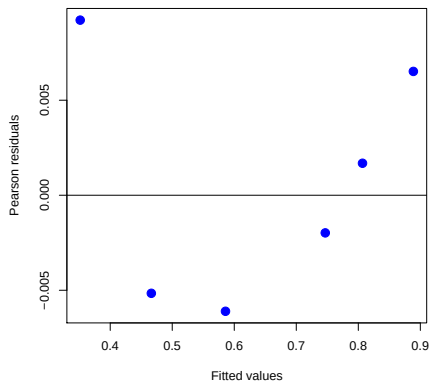


Figure 3.8: Pearson residuals against the log-dose for the logistic model.

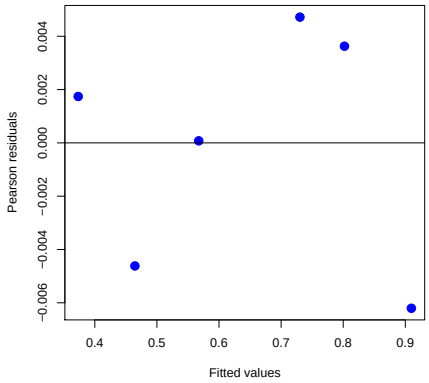


Figure 3.9: Pearson residuals against the log-dose for the complementary log-log model.

*In both the linear model (Figure 3.7) and the logistic model (Figure 3.8) the residuals present a clear pattern when plotted against the log-dose. In the linear model the "central" observations tend to produce increasing residuals when the log-dose increase. In the logistic model the residuals tend to dispose along an asymmetric U-shaped curve. In the case of the complementary log-log model the reader should make considerable efforts to imagine patterns in the plot of the residuals against the log-dose, even though this is always possible with so few observations. We will use the complementary log-log model in the rest of this section. Indeed the data presented here are part of a larger data set which we will analyze latter in this chapter. There the (at least apparently) absence of patterns in the residuals will allow us to advocate the choice of the complementary log-log model as better model for describing the data.*

□

Let us discuss now the notion of deviance for the binomial regression models. Once estimated the parameters $\alpha$ and $\beta$ by the maximum likeli-

hood estimate, say $\hat{\alpha}$ and $\hat{\beta}$, we may calculate the value of the log-likelihood function evaluated at its maximum, say $l(\hat{\alpha}, \hat{\beta})$. Proceeding in a analogously as we did before, we may define a *saturated model*, which is a model that says that each observation is binomially distributed with its own probability parameter. That is, for $i = 1, \ldots, k$,

$$Y_i \sim Bi(n_i, p_i)\,,$$

and no additional restrictions are imposed. In other words, the saturated model just specifies that the observations are binomially distributed (with their respective known number of repetitions) and nothing more is said. Any model specified by imposing additional restrictions on the probability parameters are particular cases of the saturated model; in particular the binomial regression model which uses the further restriction that $p_i = g\left(\alpha+\beta L_i\right)$ given in (3.27). Since the binomial regression model is contained in the saturated model, it makes sense to define a likelihood ratio test for testing the reduction from the saturated model to the binomial regression model. This test is performed by using the log-likelihood ratio statistic which is defined as twice the difference of the log-likelihood of the binomial regression model and the saturated model. As before, this quantity is called the *deviance*. Comparing the value of the deviance with the quantiles of a chi-square distribution with $k-2$ degrees of freedom ($k$ is the number of observations and 2 is the number of parameters in the binomial regression model) gives the decision rule of for test referred.

Let us describe this test in more formal terms next. Consider the null hypothesis:

$$\text{"}H_0 : Y_i \sim Bi(n_i, p_i), \quad \text{with} \quad p_i = g(\alpha + \beta L_i), \quad \text{for } i = 1, \ldots, k, \text{for some } \alpha \text{ and } \beta\text{"}\,,$$

which is to be tested against the alternative hypothesis:

$$\text{"}H_A : Y_i \sim Bi(n_i, p_i), \quad \text{with} \quad p_i \neq g(\alpha + \beta L_i), \quad \text{for } i = 1, \ldots, k, \text{for any } \alpha \text{ and } \beta\text{"}\,.$$

The interpretation of this test is the following: It is tested whether given that the data is binomially distributed (each observation with their respective number of repetitions), the structure assumed on the set of probability parameters by using the binomial regression model holds or not. The log-likelihood statistic for a likelihood ratio test for the hypotheses system above is,

$$D = 2\left\{l_S - l(\hat{\alpha}, \hat{\beta})\right\},$$

where $l_S$ is the value of the log-likelihood function of the saturated model evaluated at its maximum [14]. The quantity $D$ is called the *deviance* of the binomial regression model. An *asymptotic* (i.e. valid for large samples) test with level $\zeta$ is given by the rule:

"Reject $H_0$ when $D \geq \chi^2_{k-2}(1-\zeta)$",

where $\chi^2_{k-2}(1-\zeta)$ is the $(1-\zeta)$-quantile of a chi-square distribution with $k-2$ degrees of freedom.

We turn now to the question of how to study possible reductions of the binomial regression model. For example, suppose that we want to test whether the parameter $\beta$ in the binomial regression model is equal to zero. In fact this will test whether the explanatory variable affects the probability parameter. We can then define a sub-model of the binomial regression model (i.e. a model that is contained as a particular case of the binomial regression model) obtained by setting $\beta$ equal to 0. That is, define a models given by, for $i = 1, \ldots, k$,

$$Y_i \sim Bi(n_i, p_i), \text{ with } p_i = g(\alpha + 0L_i) = g(\alpha),$$

which is a sub-model of the binomial regression model which assumes that

$$Y_i \sim Bi(n_i, p_i), \text{ with } p_i = g(\alpha + \beta L_i).$$

[14] There is one maximum, believe me!

The idea is to use the likelihood ratio test to test the null hypothesis:

$$\text{"}H_0 : Y_i \sim Bi(n_i, p_i), \quad \text{with} \quad p_i = g(\alpha), \quad \text{for } i = 1, \ldots, k, \text{for some } \alpha\text{"} ,$$

against the alternative:

$$\text{"}H_A : Y_i \sim Bi(n_i, p_i), \quad \text{with} \quad p_i = g(\alpha + \beta L_i), \quad \text{for } i = 1, \ldots, k, \text{for some } \alpha \text{ and } \beta \neq 0\text{"} .$$

Sometimes we write the two hypotheses above in a shorter form as below:

$$\text{"}H_0 : \beta = 0\text{"} ,$$

and

$$\text{"}H_A : \beta \neq 0\text{"} .$$

Here it is implicit the context of the binomial regression model.

The log-likelihood ratio statistic testing $H_0$ against $H_A$ is given by

$$\Lambda = 2\left\{l(\hat{\alpha}, \hat{\beta}) - l(\hat{\alpha})\right\} , \tag{3.30}$$

where $l(\hat{\alpha}, \hat{\beta})$ and $l(\hat{\alpha})$ are the values of the log-likelihood function of the binomial regression model and the sub-model of the binomial regression model (obtained by setting $\beta = 0$) evaluated at their respective maxima. The required (asymptotic) test with level $\zeta$ is given by the rejection rule:

$$\text{"Reject } H_0 \text{ when } \Lambda \geq \chi^2_1(1-\zeta)\text{"} ,$$

where $\chi^2_1(1-\zeta)$ is the $(1-\zeta)$-quantile of a chi-square distribution with 1 degree of freedom.

We can calculate the log-likelihood ratio statistic above alternatively using the notion of deviance as follows. First we can define also the deviance of the sub-model of the binomial regression model as

$$D_0 = 2\left\{l_S - l(\hat{\alpha})\right\} ,$$

where $l_S$ and $l_S - l(\hat{\alpha})$ are the values of the log-likelihood functions of the saturated model and the sub-model of the binomial regression model evaluated at their maxima[15]. The log-likelihood ratio test defined in (3.30) can be calculated, alternatively, as follows:

$$\begin{aligned}
\Lambda &= 2\left\{l(\hat{\alpha},\hat{\beta}) - l(\hat{\alpha})\right\} \\
&\quad \text{(summing and subtracting } l_S \text{ )} \\
&= 2\left\{l(\hat{\alpha},\hat{\beta}) - l_S - l(\hat{\alpha}) + l_S\right\} \\
&\quad \text{(rearranging)} \\
&= 2\left\{l(\hat{\alpha}) - l_S\right\} - 2\left\{l(\hat{\alpha},\hat{\beta}) - l_S\right\} \\
&\quad \text{(using the definition of deviance)} \\
&= D_0 - D\,.
\end{aligned}$$

That is, for testing whether the parameter $\beta$ is equal to zero we use the difference of the deviances of the binomial regression model and the deviance of a sub-model of this model obtained by setting $\beta$ equal to zero. This difference of deviances is compared with the quantiles of a chi-square distribution with the number of degrees of freedom given by the difference of the degrees of freedom associated to these deviances. The mechanics of this test and the interpretation of the deviance are exactly as in the case of the binomial models with classification explanatory variable, even though the likelihood functions, the maximum likelihood functions and the deviances are not equivalent (but are analogous). [16]

**Example 15 (Mice convulsion, cont.)** *We consider next the complementary log-log model for the example 13.The deviance is* 0.1117 *with 4 degrees of freedom (6 observations minus 2 parameters). Then, we have no evidences*

[15]Exercise: Interpret the deviance $D_0$ in terms of a likelihood ratio test as we did for the binomial regression model. That is, write down the relevant null and alternative hypotheses of the pertinent likelihood ratio test, write the log-likelihood ratio statistic of this test and the rejection rule and finally interpret the test.

[16]Exercise: describe a likelihood ratio test for verifying whether the parameter $\alpha$ in the binomial regression model is equal to zero.

*for rejecting a reduction of the saturated model (the one that assumes that the the observations are binomially distributed each of them with their own probability parameters) to the binomial regression model. Note that we have too few observations, so the validity of this asymptotic test is questionable.*

*The maximum likelihood estimator of the parameters of the complementary log-log regression model are:* $\hat{\alpha} = -3.724$ *and* $\hat{\beta} = 1.3745$.

*The deviance of a sub-model of the binomial regression model that assumes the parameter* $\beta$ *equal to zero is* 34.4374 *with 5 degrees of freedom (6 observations minus 1 parameter). That is the log-likelihood ratio statistic for testing the hypothesis that* $\beta$ *is equal to zero is*

$$\Lambda = 34.4374 - 0.1117 = 34.3257\,,$$

*which should be compared with the quantiles of a chi-square distribution with* $5-4=1$ *degree of freedom. Clearly the p-value of this test is very small (less than 0.0001) and therefore we reject the hypothesis that* $\beta$ *is equal to zero at any reasonable level of significance. We come then to the brilliant conclusion that the insulin affects the probability of convulsion. Since the estimate of* $\beta$ *is positive, we conclude that increasing the dose of insulin increases the probability of convulsion, moreover, the model allow us (at least in principle) to estimate the probability of convulsion even when using doses of insulin different of the doses used in the experiment (at least as long we believe on estimation made with only 6 observations). However, one must be careful with extrapolations of regression models (as the one we developed) specially for extrapolations outside the rage of doses used in the experiment.*

□

## 3.3 Exercises

**Exercise 3.1** *In this exercise we will analyse an experiment on the effect of different methods of scarification on the germination of Calotropis procera L. Six different scarification treatments were applied to batches of seeds (1g of seeds or approximately 26 seeds, varying from batch to batch). Each batch was put to germinate on filter paper in a Petry dish. There were 20 batches (or Petry dishes) per treatment, all in all 120 Petry dishes and 3169 seeds! The main question is whether different scarification treatments affect the germination rate. A subsidiary question is whether these data is compatible with the two basic assumptions of the one-way binomial model.* [17]

*Here are some steps for analysing these data. As usual, there are many possible ways to analyse the data, but please follow these steps in this exercise (you will have opportunity to make "free" analyses latter in this course).*

*a- Calculate the proportion of germination for each batch (i.e., each Petri dish). Make a scatter plot representing the proportions against the treatment (here each point of the scatter plot will represent one batch).* [18] *Calculate the proportion of germination for each treatment and represent that in the scatter plot.*

*b- Fit a one-way binomial model that attributes one germination probability for each treatment.* [19] *Calculate the parameter estimates from the model and compare them with the proportions of germinated seeds per treatments. Why do the estimates coincide with the proportions?*

[17] The data frame "Ex3.1" (see the work-page) contains the data for this experiment. There are 3 variables denoted "Treat", "Germ" and "total". "Treat" represents the scarification treatment, "Germ" is the number of germinated seeds and "total" is the total number of seeds in the batch (Petri dish).

[18] Hint to R users: convert the variable Treat to numeric before plotting by using the function as.numeric( ) and convert to factor again after plotting by using the function factor().

[19] Hint: Construct a response matrix using the function cbind(). Use the function glm() with the function parameter "family" set as binomial(link="identity")

c- *Fit a logistic model that attributes one lodds for germination for each treatment.* [20] *Compare the parameter estimates obtained with this model to the parameter estimates obtained in the previous item. Convert the lodds estimated to probabilities and compare with the proportions of germinated seeds per treatments. Calculate the deviance of this model and compare with the previous. How many parameters does this model have?*

d- *Fit a logistic model that attributes the same lodds for all the observations. Calculate the deviance of this model. How many parameters has this model?*

e- *Use the results of the item c and d to make a likelihood ratio test to test the hypothesis of differences of germination rates for different treatments.*

f- *Test the hypothesis of homogeneity.* [21]

g- *Is the assumption of independence of the results between the Petri dishes reasonable? Is the assumption of independency between the germination probabilities for seeds in the same Petri dish reasonable? Discus these two assumptions and whether you can verify them.*

**Exercise 3.2** *The data of this exercises comes from a study of vegetation composition in four fields. Samples of pollen were collected in each field and it was determined the proportion of pollen from graminea plants (grass). The main question was whether the proportion of graminea (flowering) differ from field to field.*

*Details: The experiment was slightly more complex. In fact, two types of pollen traps were placed in each of the four fields, which could in principle yield different counts. The pollen in each trap was washed and fixed in*

[20] Hint to R users: set the parameter family of the function glm as binomial(link="logit")

.

[21] Hint: you do not need to fit a saturated model, since the deviance of the saturated model is known! Use the results of item e.

*microscope glass plates. For each plate 30 microscope fields (random chosen via the microscope table coordinates) were observed for determining the total number of pollen grains and the number of gaminea pollen grains. Here subsidiary questions are whether the type of pollen trap yield different proportions of graminea pollen and the whether the suppositions of a two-ways binomial model are reasonable.*

[22]

a) *Calculate the proportion of graminea pollen for each observation (i.e., counts of a microscope field). Make a scatter plot representing the proportions against the fields, separately for each trap type (here each point of the scatter plot will represent one observation, i.e., microscope fields).* [23] *Calculate the proportion of graminea for each combination of the factors Trap and Field and represent that in the scatter plot also.*

b) *Repeat the item a) using lodds instead of proportions.* [24]

c) *Fit a (full) two-way binomial logistic model* [25] *that attributes one germination probability for each combination of Trap and Field.* [26] *Calculate*

[22]The data frame "Ex3.2" contains the data for this experiment. There are 6 columns denoted "Obs.num", "Trap", "Field", "Repetition", "n.gram" and "n.grains", following the obvious name convention.

[23]Hint to R users: convert the variables Trap and Field to numeric before plotting by using the function as.numeric( ) and convert to factor again after plotting by using the function factor().

[24]Hint to R users: Use the function

$$Logit <- function(p)\{return(log(p/(1-p)))\}$$

to transform probabilities (or proportions) in lodds and the function

$$ILogit <- function(lodds)\{return(exp(lodds)/(1+exp(lodds)))\}$$

to transform lodds in probabilities (or proportions).

[25]Hint: set the parameter family of the function glm as binomial(link="logit") .

[26]Hint: Construct a response matrix using the function cbind().

*the parameter estimates from the model and compare them with the proportions (remember to transform lodds in probabilities). Calculate the deviance of this model.* [27]

*d) Fit an additive logistic model (for the two factors in play). Compare the parameter estimates obtained with this model to the parameter estimated lodds calculated in item b).*
*Convert the lodds estimated to probabilities and compare with the proportions obtained in item a).*

*e) Calculate the deviance of the additive logistic model fit in item d) and compare with the deviance of the model fit in item c). How many parameters each of these model have? Perform a likelihood ratio test to reduce the full model to the additive model. What is the interpretation of the result of this test in terms of the graph drawn in item b)?*

*f) Fit a logistic model that include only the factor Field as explanatory variable. Perform a likelihood ratio test to reduce the additive model (fit in item d)) to this model. What do you conclude from this test?*

*g) Fit a logistic model that include only the factor Trap as explanatory variable. Perform a likelihood ratio test to reduce the additive model (fit in item d)) to this model. What do you conclude from this test?*

*h) Fit a logistic model that include only the intercept as explanatory variable (i.e. not including the factor Field nor the factor Trap). Perform a likelihood ratio test to reduce the additive model (fit in item d)) to this model. What do you conclude from this test?*

*i) Perform a likelihood ratio test comparing the saturated model with the logistic additive model (fit in item d).* [28] *Interpret this test in terms of the homogeneity assumption.*

[27] Hint to R users: Use the function deviance(), if you are in doubt type ?deviance in the prompt of R.

[28] Hint: you don't need to fit the saturated model! But, if you want to do that anyway

*j) End the exercise by drawing conclusions in terms of the basic questions raised in the background description.*

**Exercise 3.3** *The data of this exercises comes from a study of the effect of phosphorous fertilization (i.e. phosphatation, a common practice fertilization in acid tropic soils) on the development of mycorrhiza in manioc plants (i.e. Manihot esculenta, a plant from the family Euphorbiaceae used to alimentation). The aim is to verify if the level of infestation of mycorrhiza is affected by phosphatation and to quantify this effect (if any).*

*Details: An experiment involving 20 manioc fields and seven levels of phosphatation (0, 100, 200, 300, 400, 500, 600 Kg phosphate per ha) was set up. In each combination of field and phosphate dose (i.e. each plot) a range of samples (up to 15, min=9, median=14) were collected and each of them was analysed and determined whether the roots of manioc (present in the sample) contained mycorrhiza or not (the samples were taken in a standardized way, close to manioc plants ...). The number of positive samples in each plot was recorded.*

[29]

*a) Calculate the proportion of (mycorrhiza) positive samples for each plot (i.e. combination of field and P dose). Make a scatter plot representing the proportions against the P dose. Calculate the proportion of (mycorrhiza) positive samples for each P dose and represent them also in the scatter plot.*

*b) Repeat the item a) using lodds instead of proportions.*

*c) Fit a binomial logistic regression model.* [30] *What is the deviance of this model? How many parameters this model has? How do you interpret the parameters of this model?*

---

(just for fun) use a factor defined by the variable Obs.num. You would have to convert this variable to a variable of type factor. Use the function factor().

[29]The data is presented in the data frame "Ex3.3". There are 5 columns denoted "Obs", "field", "P", "mycor", "nsamples", following the obvious name convention.

[30]Hint: set the parameter family of the function glm as binomial(link="logit") .

*d) Fit a binomial one-way model using the P dose as a factor (i.e. a classification variable).* [31] *What is the deviance of this model? Compare this deviance with the deviance of the previous model. How many parameters this model has? Test, using the likelihood ratio test, the reduction of this model to the model previous model. I claim that this test is a test of the adequacy of the form of the regression curve used. Why?*

*e) Compare, using the likelihood ratio test, the model fit in the item d) to the saturated model. I claim that this test verifies the homogeneity hypothesis. Why?* [32]

*f) Compare, using the likelihood ratio test, the model fit in the item c) to the saturated model. I claim that this test verifies simultaneously the homogeneity hypothesis and the assumed (logistic) form of the regression. Why?* [33]

*g) End the exercise by drawing conclusions in terms of the basic questions raised in the background description of this exercise.*

**Exercise 3.4** *The data of this exercises comes from a study of the effect of phosphorous fertilization (i.e., phosphatation, a common practice fertilization in acid tropic soils) on the development of mycorrhiza in of the gender Manihot (including the manioc plants i.e. Manihot esculenta), a gender of the family Euphorbiaceae. The aim is to verify if the level of infestation of mycorrhiza is affected by phosphatation and to quantify this effect (if any) in different for three species: Manihot esculenta (species 1), Manihot brasiliensis (species 2) and Manihot glaziovi (species 3).*

*Details: An experiment involving 20 manioc fields and seven levels of phosphatation (0, 100, 200, 300, 400, 500, 600 Kg phosphate per ha) and the three species was set up. In each combination of field, species and phosphate*

[31] Hint: Convert P to a factor using the function factor().
[32] Hint: Formulate the null and the alternative hypothesis.
[33] Hint: Combine the conclusions of item e) and f).

*dose (i.e. each plot) a range of samples (up to 15 samples per plot) were collected and each of them was analysed and determined whether the roots of manioc (present in the sample) contained mycorrhiza or not (the samples were taken in a standardized way, close to manioc plants ...). The number of positive samples in each plot was recorded.* [34]

a) *Calculate the proportion of (mycorrhiza) positive samples for each plot (i.e. combination of field, species and P dose). Make a scatter plot representing the proportions against the P dose for each of the species (you might superpose the plots and use different colours for the different species). Calculate the proportion of (mycorrhiza) positive samples for each combination of species and P dose and represent them also in the scatter plot.*

b) *Repeat the item a) using lodds instead of proportions.*

c) *(optional) Follow the steps of exercise 3.3 separately for each species (you have probably done for the species 1 already in the exercise 3.3). Fit a binomial logistic regression model.*

d) *In the next 5 steps (from step e) to i) )we restrict the analyses to species 1 and 2 (leaving the species 3 for the moment out of the analyses). So, prepare the data leaving the observations corresponding to the species 3 out.* [35]

e) *Fit a model that uses one regression per species.* [36]

[34] The data is presented in the data frame "Ex3.4". There are 6 columns denoted "Obs", "field", "P", "spec", "mycor", "nsamples", following the obvious name convention.

[35] Hint: detach the file with the data, use the function subset (e.g. restricted data ¡- subset(original.data, spec != 3) and attach the new data.

[36] Hint: Including an interaction of the **factor** spec (defining the species) and the (non-factor) P will define the model required. The call for the function glm will then be something like "glm( response.matrix ~ spec*P, family=binomial ...)".

*f) Fit a model assigning one probability (or lodds) to each combination of P level and species.* [37]

*g) Compare the model defined in item f) to a saturated model using a likelihood ratio test, which is equivalent to test the homogeneity assumption (why?) Then compare the model defined in item e) to the model defined in item f), which is equivalent to test the form of the regression curves used (why?). Now compare the model defined in item e) to a saturated model, which is equivalent to test simultaneously the homogeneity hypothesis and the regression form (why?).*

*h) Fit a model where the regression lines are parallel (in the logistic scale) and compare with the model fit in the item e). What is your conclusion?*

*i) Fit a model containing one single regression curve (common to the species) and compare, using a likelihood ratio test, it to the model fit in f). What do you conclude?*

*j) Repeat the analysis of item e) to item i) now including the three species. What changes in the analysis? Relate your conclusions with the plots made in item b).*

*k) End the exercise by drawing conclusions in terms of the basic questions raised in the background description of this exercise.*

[37] Hint: A call like "glm( response.matrix ~ spec*factor(P), family=binomial ...)" will fit the required model. Note that the difference is that here we converted the numeric variable P to a factor.

**Exercise 3.5** *In this exercise a data on sampled number of weeds is studied. The total number of plants and the number of a weed was counted by sampling (throwing a ring) in four places. In each place two topographic positions Position, 1 = low, 2 = high) were studied separately (there are several repetitions for each position at each place). The question is whether the soil type and/or the topographic position affects the prevalence of weeds.* [38]

**Exercise 3.6** *Here a data on fish mortality is to be studied. Six doses of antibiotics were applied in different tanks of fish suffering from a epidemic bacterial infection. There are two fish ecotypes in the experiment. The number of dead fish out of a given number of fish was recorded. The general question is whether the antibiotic has an effect on the mortality and whether this effect (if any) is the same for the two ecotypes.* [39]

[38]The data-frame "Ex3.5" (contained in the collection of data-frames of the course) contains the of this experiment.

[39]The data-frame "Ex3.6" (contained in the collection of data-frames of the course) contains the of this experiment.