Chapter 8
Regression Models

The aim of regression models is to model the variation of a quantitative response variable y in terms of the variation of one or several explanatory variables $(x_1, \ldots, x_p)^\top$. We have already introduced such models in Chaps. 3 and 7 where linear models were written in (3.50) as

$$y = \mathcal{X}\beta + \varepsilon,$$

where $y(n \times 1)$ is the vector of observation for the response variable, $\mathcal{X}(n \times p)$ is the data matrix of the p explanatory variables and ε are the errors. Linear models are not restricted to handle only linear relationships between y and x. Curvature is allowed by including appropriate higher order terms in the *design* matrix $\mathcal{X}$.

Example 8.1 If y represents response and x_1, x_2 are two factors that explain the variation of y via the quadratic response model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + \varepsilon_i, \ i = 1, \ldots, n. \tag{8.1}$$

This model (8.1) belongs to the class of linear models because it is linear in β. The data matrix $\mathcal{X}$ is:

$$\mathcal{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}^2 & x_{12}^2 & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}^2 & x_{22}^2 & x_{21}x_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}^2 & x_{n2}^2 & x_{n1}x_{n2} \end{pmatrix}$$

For a given value of β, the response surface can be represented in a three-dimensional plot as in Fig. 8.1 where we display $y = 20 + 1x_1 + 2x_2 - 8x_1^2 - 6x_2^2 + 6x_1x_2$, i.e. $\beta = (20, 1, 2, -8, -6, +6)^\top$.

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_8

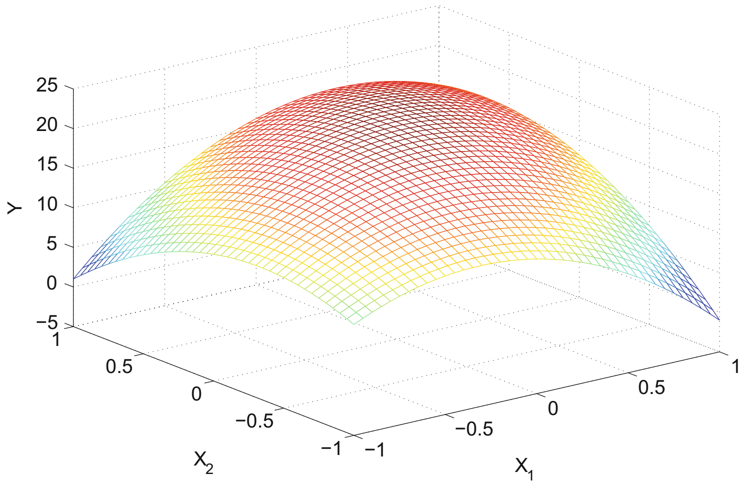

Fig. 8.1 A 3-D response surface `MVAresponsesurface`

Note also that pure non-linear models can sometimes be rewritten as a linear model by choosing an appropriate transformation of the coordinates of the variables. For instance the Cobb–Douglas production function

$$y_i = k\, x_{i1}^{\beta_1}\, x_{i2}^{\beta_2}\, x_{i3}^{\beta_3},$$

where y is the level of the production of a plant and $(x_1, x_2, x_3)^\top$ are three factors of production (e.g. labour, capital and energy), can be transformed into a linear model in the log scale. We have indeed

$$\log y_i = \beta_0 + \beta_1 \log x_{i1} + \beta_2 \log x_{i2} + \beta_3 \log x_{i3},$$

where $\beta_0 = \log k$ and the $\beta_j, j = 1, \ldots, 3$ are the elasticities ($\beta_j = \partial \log y / \partial \log x_j$).

Linear models are flexible and cover a wide class of models. If $\mathcal{X}$ has full rank, they can easily be estimated by least squares $\hat{\beta} = (\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top y$ and linear restrictions on the β's can be tested using the tools developed in Chap. 7.

In Chap. 3, we saw that even qualitative explanatory variables can be used by defining appropriate coding of the nominal values of x. In this chapter, we will extend our toolbox by showing how to code these qualitative factors in a way which allows the introduction of several qualitative factors including the possibility of interactions. This covers more general ANOVA models than those introduced in Chap. 3. This includes the ANCOVA models where qualitative and quantitative variables are both present in the explanatory variables.

When the response variable is qualitative or categorical (for instance, an individual can be employed or unemployed, a company may be bankrupt or not, the opinion of one person relative to a particular issue can be "in favour", "against" or "indifferent to", etc.), linear models have to be adapted to this particular situation. The most useful models for these cases will be presented in the second part of the chapter; this covers the log-linear models for contingency tables (where we analyse the relations between several categorical variables) and the logit model for quantal or binomial responses where we analyse the probability of being in one state as a function of explanatory variables.

8.1 General ANOVA and ANCOVA Models

8.1.1 ANOVA Models

One-Factor Models

In Sect. 3.5, we introduced the example of analysing the effect of one factor (three possible marketing strategies) on the sales of a product (a pullover), see Table 3.2. The standard way to present one factor ANOVA models with p levels is as follows

$$y_{k\ell} = \mu + \alpha_\ell + \varepsilon_{k\ell}, \; k = 1, \dots, n_\ell, \; \text{ and } \; \ell = 1, \dots, p, \tag{8.2}$$

all the $\varepsilon_{k\ell}$ being independent. Here ℓ is the label which indicates the level of the factor and α_ℓ is the effect of the ℓth level: it measures the deviation from μ, the global mean of y, due to this level of the factor. In this notation, we need to impose the restriction $\sum_{\ell=1}^{p} \alpha_\ell = 0$ in order to identify μ as the mean of y. This presentation is equivalent, but slightly different, to the one presented in Chap. 3 (compare with Eq. (3.41)), but it allows for easier extension to the multiple factors case. Note also that here we allow different sample sizes for each level of the factor (an unbalanced design, more general than the balanced design presented in Chap. 3).

To simplify the presentation, assume as in the pullover example that $p = 3$. In this case, one could be tempted to write the model (8.2) under the general form of a linear model by using three indicator variables

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \varepsilon_i,$$

where $x_{i\ell}$ is equal to 1 or 0 according to the ith observation and belongs (or not) to the level ℓ of the factor. In matrix notation and letting, for simplicity, $n_1 = n_2 = n_3 = 2$ we have with $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3)^\top$

$$y = \mathcal{X}\beta + \varepsilon, \tag{8.3}$$

where the design matrix $\mathcal{X}$ is given by:

$$\mathcal{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Unfortunately, this type of coding is not useful because the matrix $\mathcal{X}$ is not of full rank (the sum of each row is equal to the same constant 2) and therefore the matrix $\mathcal{X}^\top \mathcal{X}$ is not invertible. One way to overcome this problem is to change the coding by introducing the additional constraint that the effects add up to zero. There are many ways to achieve this. Noting that $\alpha_3 = -\alpha_1 - \alpha_2$, we do not need to introduce α_3 explicitly in the model. The linear model could indeed be written as

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \varepsilon_i,$$

with a design matrix defined as

$$\mathcal{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix},$$

which automatically implies that $\alpha_3 = -(\alpha_1 + \alpha_2)$. The linear model (8.3) is now correct with $\beta = (\mu, \alpha_1, \alpha_2)^\top$. The least squares estimator $\hat{\beta} = (\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top y$ can be computed providing the estimator of the ANOVA parameters μ and α_ℓ, $\ell = 1, \dots, 3$. Any linear constraint on β can be tested by using the techniques described in Chap. 7. For instance, the null hypothesis of no factor effect $H_0 \ : \ \alpha_1 = \alpha_2 = \alpha_3 = 0$ can be written as $H_0 \ : \ \mathcal{A}\beta = a$, where $\mathcal{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $a = (0 \quad 0)^\top$.

Multiple-Factors Models

The coding above can be extended to more general situations with many qualitative variables (factors) and with the possibility of interactions between the factors. Suppose that in a marketing example, the sales of a product can be explained by two factors: the marketing strategy with three levels (as in the pullover example) but also the location of the shop that may be either in a big shopping centre or in a less commercial location (two levels for this factor). We might also think that there is an

Table 8.1 A two factor ANOVA data set, factor A, three levels of the marketing strategy and factor B, two levels for the location

	B_1	B_2
A_1	18	15
	15	20
		25
		30
A_2	5	10
	8	12
	8	
A_3	10	20
	14	25

The figures represent the resulting sales during the same period

interaction between the two factors: the marketing strategy might have a different effect in a shopping centre than in a small quiet area. To fix the idea the data are collected as in Table 8.1.

The general two factor model with interactions can be written as

$$y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk};\ i = 1, \dots, r,\ j = 1, \dots, s,\ k = 1, \dots, n_{ij} \tag{8.4}$$

where the identification constraints are:

$$\begin{aligned} &\sum_{i=1}^{r} \alpha_i = 0 \text{ and } \sum_{j=1}^{s} \gamma_j = 0 \\ &\sum_{i=1}^{r} (\alpha\gamma)_{ij} = 0,\ j = 1, \dots, s \\ &\sum_{j=1}^{s} (\alpha\gamma)_{ij} = 0,\ i = 1, \dots, r. \end{aligned} \tag{8.5}$$

In our example of Table 8.1 we have $r = 3$ and $s = 2$. The α's measure the effect of the marketing strategy (three levels) and the γ's the effect of the location (two levels). A positive (negative) value of one of these parameters would indicate a favourable (unfavourable) effect on the expected sales; the global average of sales being represented by the parameter μ. The interactions are measured by the parameters $(\alpha\gamma)_{ij}$, $i = 1, \dots, r$, $j = 1, \dots, s$, again identification constraints implies the $(r + s)$ constraints in (8.5) on the interactions terms.

For example, a positive value of $(\alpha\gamma)_{11}$ would indicate that the effect of the sale strategy A_1 (advertisement in local newspaper), if any, is more favourable on the sales in the location B_1 (in a big commercial centre) than in the location B_2 (not a commercial centre) with the relation $(\alpha\gamma)_{11} = -(\alpha\gamma)_{12}$. As another example, a negative value of $(\alpha\gamma)_{31}$ would indicate that the marketing strategy A_3 (luxury presentation in shop windows) has less effect, if any, in location type B_1 than in B_2: again $(\alpha\gamma)_{31} = -(\alpha\gamma)_{32}$, etc.

The nice thing is that it is easy to extend the coding rule for one-factor model to this general situation, in order to present the model a standard linear model with the appropriate design matrix $\mathcal{X}$. To build the columns of $\mathcal{X}$ for the effect of each factor, we will need, as above, $r-1$ (and $s-1$) variables for coding a qualitative variable with r (and s, respectively) levels with the convention defined above in the one-factor case. For the interactions between a r between a r level factor and a s level factor, we will need $(r-1)\times(s-1)$ additional columns that will be obtained by performing the product, element by element, of the corresponding main effect columns. So, at the end, for a full model with all the interactions, we have $\{1+r-1+s-1+(r-1)(s-1)\} = rs$ parameters where the first column of 1's is for the intercept (the constant μ). We illustrate this for our marketing example where $r=3$ and $s=2$. We first describe a model without interactions.

1. *Model without interactions*
   Without the interactions (all the $(\alpha\gamma)_{ij} = 0$) the model could be written with $3 = (r-1)+(s-1)$ coded variables in a simple linear model form as in (8.3), with the matrices:

$$y = \begin{pmatrix} 18\\15\\15\\20\\25\\30\\5\\8\\8\\10\\12\\10\\14\\20\\25 \end{pmatrix}, \quad \mathcal{X} = \begin{pmatrix} 1 & 1 & 0 & 1\\ 1 & 1 & 0 & 1\\ 1 & 1 & 0 & -1\\ 1 & 1 & 0 & -1\\ 1 & 1 & 0 & -1\\ 1 & 1 & 0 & -1\\ 1 & 0 & 1 & 1\\ 1 & 0 & 1 & 1\\ 1 & 0 & 1 & 1\\ 1 & 0 & 1 & -1\\ 1 & 0 & 1 & -1\\ 1 & -1 & -1 & 1\\ 1 & -1 & -1 & 1\\ 1 & -1 & -1 & -1\\ 1 & -1 & -1 & -1 \end{pmatrix},$$

   and $\beta = (\mu, \alpha_1, \alpha_2, \gamma_1)^\top$. Then, $\alpha_3 = -(\alpha_1+\alpha_2)$ and $\gamma_2 = -\gamma_1$.
2. *Model with interactions*
   A model with interaction between A and B is obtained by adding new columns to the design matrix. We need $2 = (r-1)\times(s-1)$ new coding variables which are defined as the product, element-by-element, of the corresponding columns obtained for the main effects. For instance for the interaction parameter $(\alpha\gamma)_{11}$, we multiply the column used for coding α_1 by the column defined for coding γ_1, where the product is element-by-element. The same is done for the parameter

$(\alpha\gamma)_{21}$. No other columns are necessary, since the remaining interactions are derived from the identification constraints (8.5). We obtain

$$\mathcal{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix},$$

with $\beta = (\mu, \alpha_1, \alpha_2, \gamma_1, (\alpha\gamma)_{11}, (\alpha\gamma)_{21})^\top$. The other interactions can indeed be derived from (8.5)

$$\begin{aligned} (\alpha\gamma)_{12} &= -(\alpha\gamma)_{11} \\ (\alpha\gamma)_{22} &= -(\alpha\gamma)_{21} \\ (\alpha\gamma)_{31} &= -\left((\alpha\gamma)_{11} + (\alpha\gamma)_{21}\right) \\ (\alpha\gamma)_{32} &= -(\alpha\gamma)_{31}. \end{aligned}$$

The estimation of β is again simply given by the least squares solution $\hat{\beta} = (\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top y$.

Example 8.2 Let us come back to the marketing data provided by the two-way Table 8.1. The values of $\hat{\beta}$ in the full model, with interactions, are given in Table 8.2. The p-values in the right column are for the individual tests: it appears that the interactions do not provide additional significant explanation of y, but the effect of the two factors seems significant.

Using the techniques of Chap. 7, we can test some reduced model corresponding to linear constraints on the β's. The full model is the model with all the parameters, including all the interactions. The overall fit test H_0 : all the parameters, except μ, are equal to zero, gives the value $F_{\text{observed}} = 6.5772$ with a p-value of 0.0077 for a $F_{5,9}$, so that H_0 is rejected. In this case, the $\text{RSS}_{\text{reduced}} = 735.3333$. So there is some effect by the factors.

Table 8.2 Estimation of the two factors ANOVA model with data from Table 8.1

	$\hat{\beta}$	p-Values
μ	15.25	
α_1	4.25	0.0218
α_2	−6.25	0.0033
γ_1	−3.42	0.0139
$(\alpha\gamma)_{11}$	0.42	0.7922
$(\alpha\gamma)_{21}$	1.42	0.8096
RSS_{full}	158.00	

We then test a less reduced model. We can test if the interaction terms are significantly different to zero. This is a linear constraint on β with

$$\mathcal{A} = \begin{pmatrix} 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \end{pmatrix};\; a = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Under the null we obtain:

$$\hat{\beta}_{H_0} = \begin{pmatrix} 15.3035 \\ 4.0975 \\ -6.0440 \\ -3.2972 \\ 0 \\ 0 \end{pmatrix},$$

and $\mathrm{RSS}_{\mathrm{reduced}} = 181.8019$. The observed value of $F = 0.6779$ which is not significant ($r = 11$, $f = 9$) the p-value $= \mathrm{P}(F_{2,9} \geq 0.6779) = 0.5318$, confirming the absence of interactions.

Now taking the model without the interactions as the full model, we can test if one of the main effects α (marketing strategy) or γ (location) or both are significantly different from zero. We leave this as an exercise for the reader.

8.1.2 ANCOVA Models

ANCOVA (ANalysis of COVAriances) are mixed models where some variables are qualitative and others are quantitative. The same coding of the ANOVA will be used for the qualitative variable. The design matrix $\mathcal{X}$ is completed by the columns for the quantitative explanatory variables x. Interactions between a qualitative variable (a factor with r levels) and a quantitative one x is also possible, this corresponds to situations where the effect of x on the response y is different according to the level of the factor. This is achieved by adding into the design matrix $\mathcal{X}$, a new column obtained by the product, element-by-element, of the quantitative variable with the coded variables for the factor ($r - 1$ interaction variables if the categorical variable has r levels).

For instance consider a simple model where a response y is explained by one explanatory variable x and one factor with two levels (for instance the gender level 1 for men and level 2 for women), we would have in the case $n_1 = n_2 = 3$

$$\mathcal{X} = \begin{pmatrix} 1 & x_1 & 1 & x_1 \\ 1 & x_2 & 1 & x_2 \\ 1 & x_3 & 1 & x_3 \\ 1 & x_4 & -1 & -x_4 \\ 1 & x_5 & -1 & -x_5 \\ 1 & x_6 & -1 & -x_6 \end{pmatrix},$$

with $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^\top$. The intercept and the slope are $(\beta_1 + \beta_3)$ and $(\beta_1 + \beta_4)$ for men and $(\beta_1 - \beta_3)$ and $(\beta_1 - \beta_4)$ for women. This situation is displayed in Fig. 8.2.

Example 8.3 Consider the Car Data provided in Sect. 22.3. We want to analyse the effect of the weight (W), the displacement (D) on the mileage (M). But we would like to test if the origin of the car (the factor C) has some effect on the response and if the effect of the continuous variables is different for the different levels of the factor.

From the regression results in Table 8.3, we observe that only the weight affects the mileage, while the displacement does not. We also consider the origin of the car, however, both the displacement and the factor are not significant. Table 8.4 is for different factor levels.

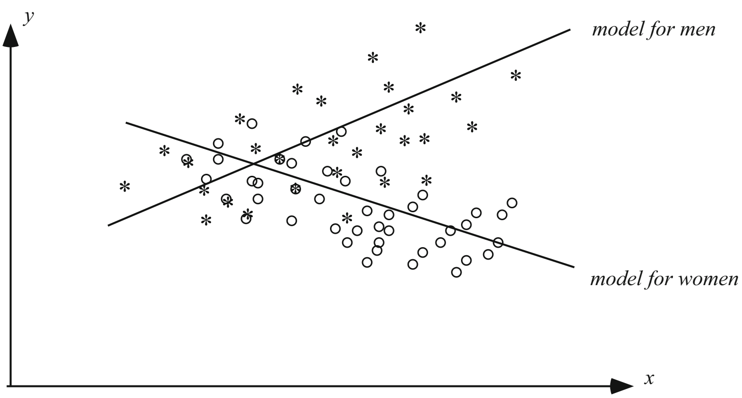

Fig. 8.2 A model with interaction

Table 8.3 Estimation of the effects of weight and displacement on the mileage MVAcareffect

	$\hat{\beta}$	p-Values	$\tilde{\beta}$	p-Values
μ	41.0066	0.0000	43.4031	0.0000
W	−0.0073	0.0000	−0.0074	0.0000
D	0.0118	0.2250	0.0081	0.4140
C			−0.9675	0.1250

Table 8.4 Different factor levels on the response MVAcareffect

	μ	p-Values	W	p-Values	D	p-Values
$c = 1$	40.043	0.0000	−0.0065	0.0000	0.0058	0.3790
$c = 2$	47.557	0.0005	0.0081	0.3666	−0.3582	0.0160
$c = 3$	44.174	0.0002	0.0039	0.7556	−0.2650	0.3031

8.1.3 Boston Housing

In Chaps. 3 and 7, linear models were used to analyse if the variations of the price (the variables were transformed in Sect. 1.9) could be explained by other variables. A reduced model was obtained in Sect. 7.3 with the results shown in Table 7.1, with $r^2 = 0.763$. The model was:

$$X_{14} = \beta_0 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13}$$

One factor (X_4) was coded as a binary variable (1, if the house is close to the Charles River and 0 if it is not). Taking advantage of the ANCOVA models described above, we would like to add to a new factor built from the original quantitative variable X_9 = index of accessibility to radial highways. So we will transform X_4 as being 1 if close to the Charles River and −1 if not, and we will replace X_9 by a new factor coded $X_{15} = 1$ if $X_9 \geq \text{median}(X_9)$ and $X_{15} = -1$ if $X_9 < \text{median}(X_9)$. We also want to consider the interaction of X_4 with X_{12} (proportion of blacks) and the interaction of X_4 with the new factor X_{15}. The results are shown in Table 8.5.

Summary

↪ ANOVA models can be dividend into one-factor models and multiple factor models.

↪ Multiple factor models analyse many qualitative variables and the interactions between them.

Table 8.5 Estimation of the ANCOVA model using the Boston housing data `MVAboshousing`

	$\hat{\beta}$	p-Values	$\tilde{\beta}$	p-Values
β_0	32.27	0.00	27.65	0.00
β_4	1.54	0.00	−3.19	0.32
β_5	−17.59	0.00	−16.50	0.00
β_6	4.27	0.00	4.23	0.00
β_8	−1.13	0.00	−1.10	0.00
β_{10}	0.00	0.97	0.00	0.95
β_{11}	−0.97	0.00	−0.97	0.00
β_{12}	0.01	0.00	0.02	0.01
β_{13}	−0.54	0.00	−0.54	0.00
β_{15}	0.21	0.46	0.23	0.66
β_{4*14}			0.01	0.13
β_{4*15}			0.03	0.95

Summary (continued)
$\hookrightarrow$ ANCOVA models are mixed models with qualitative and quantitative variables, and can also incorporate the interaction between a qualitative and a quantitative variable.

8.2 Categorical Responses

8.2.1 Multinomial Sampling and Contingency Tables

In many applications, the response variable of interest is qualitative or categorical, in the sense that the response can take its nominal value in one of, say, K classes or categories. Often we observe counts y_k, the number of observations in category $k = 1, \ldots, K$. If the total number of observations $n = \sum_{k=1}^{K} y_k$ is fixed and we may assume independence of the observations, we obtain a multinomial sampling process.

If we denote by p_k the probability of observing the kth category with $\sum_{k=1}^{K} p_k = 1$, we have $\mathsf{E}(Y_k) = m_k = np_k$. The likelihood of the sample can then be written as:

$$L = \frac{n!}{\prod_{k=1}^{K} y_k!} \prod_{k=1}^{K} \left(\frac{m_k}{n}\right)^{y_k}. \tag{8.6}$$

In contingency tables, the categories are defined by several qualitative variables. For example in a $(J \times K)$ two-way table, the observations (counts) y_{jk}, $j = 1, \ldots, J$ and $k = 1, \ldots, K$ are reported for row j and column k. Here $n = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk}$. Log-linear models introduce a linear structure on the logarithms of the expected frequencies $m_{jk} = \mathsf{E}(y_{jk}) = np_{jk}$, with $\sum_{j=1}^{J} \sum_{k=1}^{K} p_{jk} = 1$. Log-linear structures on m_{jk} will impose the same structure for the p_{jk}, the estimation of the model will then be obtained by constrained maximum likelihood. Three-way tables $(J \times K \times L)$ may be analysed in the same way.

Sometimes additional information is available on explanatory variables x. In this case, the logit model will be appropriate when the categorical response is binary $(K = 2)$. We will introduce these models when the main response of interest is binary (for instance tables $(2 \times K)$ or $(2 \times K \times L)$). Further, we will show how they can be adapted to the case of contingency tables. Contingency tables are also analysed by multivariate descriptive tools in Chap. 15.

8.2.2 *Log-Linear Models for Contingency Tables*

Two-Way Tables

Consider a $(J \times K)$ two-way table, where y_{jk} is the number of observations having the nominal value j for the first qualitative character and nominal value k for the second character. Since the total number of observations is fixed $n = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk}$, there are $JK - 1$ free cells in the table. The multinomial likelihood can be written as in (8.6)

$$L = \frac{n!}{\prod_{j=1}^{J} \prod_{k=1}^{K} y_{jk}!} \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\frac{m_{jk}}{n} \right)^{y_{jk}}, \tag{8.7}$$

where we now introduce a log-linear structure to analyse the role of the rows and the columns to determine the parameters $m_{jk} = \mathsf{E}(y_{jk})$ (or p_{jk}).

1. *Model without interaction*

   Suppose that there is no interaction between the rows and the columns: this corresponds to the hypothesis of independence between the two qualitative characters. In other words, $p_{jk} = p_j p_k$ for all j, k. This implies the log-linear model:

   $$\log m_{jk} = \mu + \alpha_j + \gamma_k \quad \text{for } j = 1, \ldots, J, \; k = 1, \ldots, K, \tag{8.8}$$

   where, as in ANOVA models for identification purposes $\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \gamma_k = 0$. Using the same coding devices as above, the model can be written as

   $$\log m = \mathcal{X} \beta. \tag{8.9}$$

For a (2×3) table we have:

$$\log m = \begin{pmatrix} \log m_{11} \\ \log m_{12} \\ \log m_{13} \\ \log m_{21} \\ \log m_{22} \\ \log m_{23} \end{pmatrix}, \ \mathcal{X} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

where the first column of $\mathcal{X}$ is for the constant term, the second column is the coded column for the 2-levels row effect and the two last columns are the coded columns for the 3-levels column effect. The estimation is obtained by maximising the log-likelihood which is equivalent to maximising the function $L(\beta)$ in β:

$$L(\beta) = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \log m_{jk}. \tag{8.10}$$

The maximisation is under the constraint $\sum_{j,k} m_{jk} = n$. In summary we have $1 + (J-1) + (K-1) - 1$ free parameters for $JK - 1$ free cells. The number of *degrees of freedom* in the model is the number of free cells minus the number of free parameters. It is given by

$$r = JK - 1 - (J-1) - (K-1) = (J-1)(K-1).$$

In the example above, we have therefore $(3-1) \times (2-1) = 2$ degrees of freedom.

The original parameters of the model can then be estimated as:

$$\begin{aligned} \alpha_1 &= \beta_1 \\ \alpha_2 &= -\beta_1 \\ \gamma_1 &= \beta_2 \\ \gamma_2 &= \beta_3 \\ \gamma_3 &= -(\beta_2 + \beta_3). \end{aligned} \tag{8.11}$$

2. *Model with interactions*
In two-way tables the interactions between the two variables are of interest. This corresponds to the general (full) model

$$\log m_{jk} = \mu + \alpha_j + \gamma_k + (\alpha\gamma)_{jk}, \ j = 1, \ldots, J, \ k = 1, \ldots, K, \tag{8.12}$$

where in addition, we have the $J + K$ restrictions

$$\sum_{k=1}^{K} (\alpha\gamma)_{jk} = 0, \quad \text{for } j = 1, \ldots, J$$
$$\sum_{j=1}^{J} (\alpha\gamma)_{jk} = 0, \quad \text{for } k = 1, \ldots, K \tag{8.13}$$

As in the ANOVA model, the interactions may be coded by adding $(J-1)(K-1)$ columns to $\mathcal{X}$, obtained by the product of the corresponding coded variables. In our example for the (2×3) table the design matrix $\mathcal{X}$ is completed with two more columns:

$$\mathcal{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}.$$

Now the interactions are determined by using (8.13):

$$\begin{aligned}
(\alpha\gamma)_{11} &= \beta_4 \\
(\alpha\gamma)_{12} &= \beta_5 \\
(\alpha\gamma)_{13} &= -\{(\alpha\gamma)_{11} + (\alpha\gamma)_{12}\} = -(\beta_4 + \beta_5) \\
(\alpha\gamma)_{21} &= -(\alpha\gamma)_{11} = -\beta_4 \\
(\alpha\gamma)_{22} &= -(\alpha\gamma)_{12} = -\beta_5 \\
(\alpha\gamma)_{23} &= -(\alpha\gamma)_{13} = \beta_4 + \beta_5
\end{aligned}$$

We have again a log-linear model as in (8.9) and the estimation of β goes through the maximisation in β of $L(\beta)$ given by (8.10) under the same constraint.

The model with all the interaction terms is called the *saturated* model. In this model there are no degrees of freedom, the number of free parameters to be estimated equals the number of free cells. The parameters of interest are the interactions. In particular, we are interested in testing their significance. These issues will be addressed below.

Three-Way Tables

The models presented above for two-way tables can be extended to higher order tables but at a cost of notational complexity. We show how to adapt to three-way tables. This deserves special attention due to the presence of higher-order interactions in the saturated model.

A $(J \times K \times L)$ three-way table may be constructed under multinomial sampling as follows: each of the n observations falls in one, and only one, category of each of three categorical variables having J, K and L modalities respectively. We end up with a three-dimensional table with JKL cells containing the counts $y_{jk\ell}$ where $n = \sum_{j,k,\ell} y_{jk\ell}$. The expected counts depend on the unknown probabilities $p_{jk\ell}$ in the usual way:

$$m_{jk\ell} = n\, p_{jk\ell}, \;\; j = 1, \ldots, J, \; k = 1, \ldots, K, \; \ell = 1, \ldots, L.$$

1. *The saturated model*
   A full saturated log-linear model reads as follows:

$$\begin{aligned}\log m_{jk\ell} &= \mu + \alpha_j + \beta_k + \gamma_\ell + (\alpha\beta)_{jk} + (\alpha\gamma)_{j\ell} + (\beta\gamma)_{k\ell} + (\alpha\beta\gamma)_{jk\ell},\\ & j = 1, \ldots, J, \; k = 1, \ldots, K, \; \ell = 1, \ldots, L. \end{aligned} \tag{8.14}$$

   The restrictions are the following (using the "dot" notation for summation on the corresponding indices):

$$\begin{aligned} &\alpha_{(\bullet)} = \beta_{(\bullet)} = \gamma_{(\bullet)} = 0\\ &(\alpha\beta)_{j\bullet} = (\alpha\gamma)_{j\bullet} = (\beta\gamma)_{k\bullet} = 0\\ &(\alpha\beta)_{\bullet k} = (\alpha\gamma)_{\bullet\ell} = (\beta\gamma)_{\bullet\ell} = 0\\ &(\alpha\beta\gamma)_{jk\bullet} = (\alpha\beta\gamma)_{j\bullet\ell} = (\alpha\beta\gamma)_{\bullet k\ell} = 0 \end{aligned}$$

   The parameters $(\alpha\beta)_{jk}, (\alpha\gamma)_{j\ell}, (\beta\gamma)_{k\ell}$ are called *first-order interactions*. The *second-order interactions* are the parameters $(\alpha\beta\gamma)_{jk\ell}$, they allow to take into account heterogeneities in the interactions between two of the three variables. For instance, let ℓ stand for the two gender categories $(L = 2)$, if we suppose that $(\alpha\beta\gamma)_{jk1} = -(\alpha\beta\gamma)_{jk2} \neq 0$, we mean that the interactions between the variable J and K are not the same for both gender categories.

   The estimation of the parameters of the saturated model are obtained through maximisation of the log-likelihood. In the multinomial sampling scheme, it corresponds to maximising the function:

$$L = \sum_{j,k,\ell} y_{jk\ell} \log m_{jk\ell},$$

   under the constraint $\sum_{j,k,\ell} m_{jk\ell} = n$.

The number of degrees of freedom in the saturated model is again zero. Indeed, the number of free parameters in the model is

$$1 + (J-1) + (K-1) + (L-1) + (J-1)(K-1) + (J-1)(L-1)$$
$$+(K-1)(L-1) + (J-1)(K-1)(L-1) - 1 = JKL - 1.$$

This is indeed equal to the number of free cells in the table and so, there is no degree of freedom.

2. *Hierarchical non-saturated models*
As illustrated above, a saturated model has no degrees of freedom. Non-saturated models correspond to reduced models where some parameters are fixed to be equal to zero. They are thus particular cases of the saturated model (8.14). The *hierarchical* non-saturated models that we will consider here, are models where once a set of parameters is set equal to zero, all the parameters of higher-order containing the same indices are also set equal to zero.

For instance if we suppose $\alpha_1 = 0$, we only consider non-saturated models where also $(\alpha\gamma)_{1\ell} = (\alpha\beta)_{1k} = (\alpha\beta\gamma)_{1k\ell} = 0$ for all values of k and ℓ. If we only suppose that $(\alpha\beta)_{12} = 0$, we also assume that $(\alpha\beta\gamma)_{12\ell} = 0$ for all ℓ.

Hierarchical models have the advantage of being more easily interpretable. Indeed without this hierarchy, the models would be difficult to interpret. What would be, for instance, the meaning of the parameter $(\alpha\beta\gamma)_{12\ell}$, if we know that $(\alpha\beta)_{12} = 0$? The estimation of the non-saturated models will be achieved by the usual way i.e. by maximising the log-likelihood function L as above but under the new constraints of the reduced model.

8.2.3 Testing Issues with Count Data

One of the main practical interests in regression models for contingency tables is to test restrictions on the parameters of a more complete model. These testing ideas are created in the same spirit as in Sect. 3.5 where we tested restrictions in ANOVA models.

In linear models, the test statistics is based on the comparison of the goodness of fit for the full model and for the reduced model. Goodness of fit is measured by the residual sum of squares (RSS). The idea here will be the same here but with a more appropriate measure for goodness of fit. Once a model has been estimated, we can compute the predicted value under that model for each cell of the table. We will denote, as above, the observed value in a cell by y_k and $\hat{m}_k$ will denote the expected value predicted by the model. The goodness of fit may be appreciated by measuring, in some way, the distance between the series of observed and of predicted values.

Two statistics are proposed: the *Pearson chi-square* X^2 and the *Deviance* noted G^2. They are defined as follows:

$$X^2 = \sum_{k=1}^{K} \frac{(y_k - \hat{m}_k)^2}{\hat{m}_k} \tag{8.15}$$

$$G^2 = 2 \sum_{k=1}^{K} y_k \log\left(\frac{y_k}{\hat{m}_k}\right) \tag{8.16}$$

where K is the total number of cells of the table. The deviance is directly related to the log-likelihood ratio statistic and is usually preferred because it can be used to compare nested models as we usually do in this context.

Under the hypothesis that the model used to compute the predicted value is true, both statistics (for large samples) are approximately distributed as a χ^2 variable with degrees of freedom $d.f.$ depending on the model. The $d.f.$ can be computed as follows:

$$d.f. = \#\,\text{free cells} - \#\,\text{free parameters estimated.} \tag{8.17}$$

For saturated models, the fit is perfect: $X^2 = G^2 = 0$ with $d.f. = 0$.

Suppose now that we want to test a reduced model which is a restricted version of a full model. The deviance can then be used as the F statistics in linear regression. The test procedure is straightforward:

$$\begin{aligned} H_0 &: \text{ reduced model with } r \text{ degrees of freedom} \\ H_1 &: \text{ full model with } f \text{ degrees of freedom.} \end{aligned} \tag{8.18}$$

Since, the full model contains more parameters, we expect the deviance to be smaller. We reject the H_0 if this reduction is significant, i.e. if $G^2_{H_0} - G^2_{H_1}$ is large enough. Under H_0 one has:

$$G^2_{H_0} - G^2_{H_1} \sim \chi^2_{r-f}.$$

We reject H_0 if the p-value:

$$\mathrm{P}\left\{\chi^2_{r-f} > \left(G^2_{H_0} - G^2_{H_1}\right)\right\}.$$

is small. Suppose we want to test the independence in a $(J \times K)$ two-way table (no interaction). Here the full model is the saturated one with no degrees of freedom $(f = 0)$ and the restricted model has $r = (J-1)\,(K-1)$ degrees of freedom. We reject H_0 if the p-value of H_0 $\mathrm{P}\{\chi^2_r > \left(G^2_{H_0}\right)\}$ is too small.

This test is equivalent to the Pearson chi-square test for independence in two-way tables ($G^2_{H_0} \approx X^2_{H_0}$ when n is large).

Table 8.6 A three-way contingency table: top table for men and bottom table for women MVAdrug

M	A1	A2	A3	A4	A5
DY	21	32	70	43	19
DN	683	596	705	295	99
F	A1	A2	A3	A4	A5
DY	46	89	169	98	51
DN	738	700	847	336	196

Table 8.7 Coefficient estimates based on the saturated model MVAdrug

	$\hat{\beta}$		$\hat{\beta}$
$\hat{\beta}_0$ intercept	5.0089	$\hat{\beta}_{10}$	0.0205
$\hat{\beta}_1$ gender: M	−0.2867	$\hat{\beta}_{11}$	0.0482
$\hat{\beta}_2$ drug: DY	−1.0660	$\hat{\beta}_{12}$ drug*age	−0.4983
$\hat{\beta}_3$ age	−0.0080	$\hat{\beta}_{13}$	−0.1807
$\hat{\beta}_4$	0.2151	$\hat{\beta}_{14}$	0.0857
$\hat{\beta}_5$	0.6607	$\hat{\beta}_{15}$	0.2766
$\hat{\beta}_6$	−0.0463	$\hat{\beta}_{16}$ gender*drug*age	−0.0134
$\hat{\beta}_7$ gender*drug	−0.1632	$\hat{\beta}_{17}$	−0.0523
$\hat{\beta}_8$ gender*age	0.0713	$\hat{\beta}_{18}$	−0.0112
$\hat{\beta}_9$	−0.0092	$\hat{\beta}_{19}$	−0.0102

Example 8.4 Everitt and Dunn (1998) provide a three-dimensional ($2 \times 2 \times 5$) count table of $n = 5{,}833$ interviewed people. The count were on prescribed psychotropic drugs in the fortnight prior to the interview as a function of age and gender. The data are summarised in Table 8.6, where the categories for the three factors are M for male, F for female, DY for "yes" having taken drugs, DN for "no" not having taking drugs and the five age categories: A1 (16–29), A2 (30–44), A3 (45–64), A4 (65–74), A5 for over 74. The table provides the observed frequencies $y_{jk\ell}$ in each of the cells of the three-way table: where j stands for gender, k for drug and ℓ for age categories. The design matrix $\mathcal{X}$ for the full saturated model can be found in the quantlet MVAdrug.

The saturated model gives the estimates displayed in Table 8.7.

We see for instance that $\hat{\beta}_1 < 0$, so there are fewer men than women in the study, since $\hat{\beta}_7$ is also negative it seems that the tendency of men taking the drug is less important than for women. Also, note that $\hat{\beta}_{12}$ to $\hat{\beta}_{15}$ forms an increasing sequence, so that the age factor seems to increase the tendency to take the drug. Note that in this saturated model, there are no degrees of freedom and the fit is perfect, $\hat{m}_{jk\ell} = y_{jk\ell}$ for all the cells of the table.

The second order interactions have a lower order of magnitude, so we want to test if they are significantly different to zero. We consider a restricted model where

Table 8.8 Coefficients estimates based on the maximum likelihood method MVAdrug-3waysTab

	$\hat{\beta}$		$\hat{\beta}$
$\hat{\beta}_0$ intercept	5.0051	$\hat{\beta}_8$ gender*age	0.0795
$\hat{\beta}_1$ gender: M	−0.2919	$\hat{\beta}_9$	0.0321
$\hat{\beta}_2$ drug: DY	−1.0717	$\hat{\beta}_{10}$	0.0265
$\hat{\beta}_3$ age	−0.0030	$\hat{\beta}_{11}$	0.0534
$\hat{\beta}_4$	0.2358	$\hat{\beta}_{12}$ drug*age	−0.4915
$\hat{\beta}_5$	0.6649	$\hat{\beta}_{13}$	−0.1576
$\hat{\beta}_6$	−0.0425	$\hat{\beta}_{14}$	0.0917
$\hat{\beta}_7$ gender*drug	−0.1734	$\hat{\beta}_{15}$	0.2822

$(\alpha\beta\gamma)_{jk\ell}$ are all set to zero. This can be achieved by testing $H_0: \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = 0$. The maximum likelihood estimators of the restricted model are obtained by deleting the last four columns in the design matrix $\mathcal{X}$. The results are given in Table 8.8.

We have $J = 2$, $K = 2$ and $L = 5$, this makes $JKL - 1 = 19$ free cells. The full model has $f = 0$ degrees of freedom and the reduced model has $r = 4$ degrees of freedom. The G^2 deviance is given by 2.3004; it has 4 degrees of freedom (the chi-square statistics is 2.3745). The p-value of the restricted model is 0.6807, so we do not reject the null hypothesis (the restricted model without 2nd order interaction). In others words, age does not interfere with the interactions between gender and drugs, or equivalently, gender does not interfere in the interactions between age and drugs. The reader can verify that the first order interactions are significant, by taking, for instance, the model without interactions of the second order as the new full model and testing a reduced model where all the first order interactions are all set to zero. MVAdrug3waysTab

8.2.4 *Logit Models*

Logit models are useful to analyse how explanatory variables influence a binary response y. The response y may take the two values 1 and 0 to denote the presence or absence of a certain qualitative trait (a person can be employed or unemployed, a firm can be bankrupt or not, a patient can be affected by a certain disease or not, etc.). Logit models are designed to estimate the probability of $y = 1$ as a logistic function of linear combinations of x. Logit models can be adapted to the analysis of contingency tables where one of the qualitative variables is binary. One obtains the probability of being in one of the two states of this binary variable as a function of the other variables. We concentrate here on $(2 \times K)$ and $(2 \times K \times L)$ tables.

Logit Models for Binary Response

Consider the vector y $(n \times 1)$ of observations on a binary response variable (a value of "1" indicating the presence of a particular qualitative trait and a value of "0", its absence). The logit model makes the assumption that the probability for observing $y_i = 1$ given a particular value of $x_i = (x_{i1}, \dots, x_{ip})^\top$ is given by the logistic function of a "score", a linear combination of x:

$$p(x_i) = \mathrm{P}(y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}. \tag{8.19}$$

This entails the probability of the absence of the trait:

$$1 - p(x_i) = \mathrm{P}(y_i = 0 \mid x_i) = \frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})},$$

which implies

$$\log\left\{\frac{p(x_i)}{1 - p(x_i)}\right\} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{8.20}$$

This indicates that the logit model is equivalent to a log-linear model for the odds ratio $p(x_i)/\{1 - p(x_i)\}$. A positive value of β_j indicates an explanatory variable x_j that will favour the presence of the trait since it improves the odds. A zero value of β_j corresponds to the absence of an effect of this variable on the appearance of the qualitative trait.

For i.i.d observations the likelihood function is:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} \{1 - p(x_i)\}^{1-y_i}.$$

The maximum likelihood estimators of the β's are obtained as the solution of the non-linear maximisation problem $(\hat{\beta}_0, \hat{\beta}) = \arg\max_{\beta_0,\beta} \log L(\beta_0, \beta)$ where

$$\log L(\beta_0, \beta) = \sum_{i=1}^{n} \left[y_i \log p(x_i) + (1 - y_i) \log\{1 - p(x_i)\}\right].$$

The asymptotic theory of the MLE of Chap. 6 (see Theorem 6.3) applies and thus asymptotic inference on β is available (test of hypothesis or confidence intervals).

Example 8.5 In the bankruptcy data set (see Sect. 22.22), we have measures on 5 financial characteristics on 66 banks, 33 among them being bankrupt and the other 33 still being solvent. The logit model can be used to evaluate the probability of

Table 8.9 Probabilities of the bankruptcies with the logit model MVAbankrupt

	$\hat{\beta}$	p-Values
β_0	3.6042	0.0660
β_3	−0.2031	0.0037
β_4	−0.0205	0.0183

Table 8.10 A $(2 \times K)$ contingency table

	1	$\cdots$	k	$\cdots$	K	Total
1	y_{11}	$\cdots$	y_{1k}	$\cdots$	y_{1K}	y_1
2	y_{21}	$\cdots$	y_{2k}	$\cdots$	y_{2K}	y_2
Total	$y_{\bullet 1}$	$\cdots$	$y_{\bullet k}$	$\cdots$	$y_{\bullet K}$	$y_{\bullet} = n$

bankruptcy as a function of these financial ratios. We obtain the results summarised in Table 8.9. We observe that only β_3 and β_4 are significant.

Logit Models for Contingency Tables

The logit model may contain quantitative and qualitative explanatory variables. In the latter case, the variable may be coded according to the rules described in the ANOVA/ANCOVA sections above. This enables a revisit to the contingency tables where one of the variables is binary and is the variable of interest. How can the probability of taking one of the two nominal values be evaluated as a function of the other variables? We keep the notations of Sect. 8.1 and suppose, without loss of generality, that the first variable with $J = 2$ is the binary variable of interest. In the drug Example 8.4, we have a $(2\times2\times5)$ table and one is interested in the probability of taking a drug as a function of age and gender.

$(2 \times K)$ Tables with Binomial Sampling

In Table 8.10 we have displayed the situation. Let p_k be the probability of falling into the first row for the k-th column, $k = 1, \ldots, K$. Since we are mainly interested in the probabilities p_k as a function of k, we suppose here that $y_{\bullet k}$ are fixed for $k = 1, \ldots, K$ (or we work conditionally on the observed value of these column totals), where $y_{\bullet k} = \sum_{j=1}^{J} y_{jk}$. Therefore, we have K independent binomial processes with parameters $(y_{\bullet k}, p_k)$. Since the column variable is nominal we can use an ANOVA model to analyse the effect of the column variable on p_k through the logs of the odds

$$\log\left(\frac{p_k}{1-p_k}\right) = \eta_0 + \eta_k,\ k = 1, \ldots, K, \tag{8.21}$$

where $\sum_{k=1}^{K} \eta_k = 0$. As in the ANOVA models, one of the interests will be to test $H_0 : \eta_1 = \cdots = \eta_K = 0$. The log-linear model for the odds has its equivalent in a logit formulation for p_k

$$p_k = \frac{\exp(\eta_0 + \eta_k)}{1 + \exp(\eta_0 + \eta_k)}, \; k = 1, \ldots, K. \tag{8.22}$$

Note that we can code the RHS of (8.21) as a linear model $\mathcal{X}\theta$, where for instance, for a (2×4) table ($K = 4$) we have:

$$\mathcal{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix}, \; \theta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix},$$

where $\eta_0 = \beta_0, \eta_1 = \beta_1, \eta_2 = \beta_2, \eta_3 = \beta_3$ and $\eta_4 = -(\beta_1 + \beta_2 + \beta_3)$. The logit model for p_k, $k = 1, \ldots, K$ can now be written, with some abuse of notation, as the K-vector

$$p = \frac{\exp(\mathcal{X}\theta)}{1 + \exp(\mathcal{X}\theta)},$$

where the division has to be understood as being element-by-element. The MLE of θ is obtained by maximising the log-likelihood

$$L(\theta) = \sum_{k=1}^{K} \{y_{1k} \log p_k + y_{2k} \log(1 - p_k)\}, \tag{8.23}$$

where the p_k are elements of the K-vector p.

This logit model is a *saturated* model. Indeed the number of free parameters is K, the dimension of θ, and the number of free cells is also equal to K since we consider the column totals $y_{\bullet k}$ as being fixed. So, there are no degrees of freedom in this model. It can be proven that this logit model is equivalent to the saturated model for a table $(2 \times K)$ presented in Sect. 8.2.2 where all the interactions are present in the model. The hypothesis of all interactions $(\alpha\gamma)_{jk}$ being equal to zero (independence case) is equivalent to the hypothesis that the η_k, $k = 1, \ldots, K$ are all equal to zero (no column effect on the probabilities p_k).

The main interest of the logit presentation is its flexibility when the variable defining the column categories is a quantitative variable (age group, number of children, etc.). Indeed, when this is the case, the logit model allows to quantify the effect of the column category by using less parameters and a more flexible relationship than a linear relation. Suppose that we could attach a representative value x_k to each column category for this class (for instance, it could be the median value, or the average value of the class category). We can then choose the following logit model for p_k

$$p_k = \frac{\exp(\eta_0 + \eta_1 x_k)}{1 + \exp(\eta_0 + \eta_1 x_k)}, \ k = 1, \dots, K, \tag{8.24}$$

where we now have only two free parameters for K free cells, so we have $K-2$ degrees of freedom. We could even introduce a quadratic term to allow some curvature effect of x on the odds

$$p_k = \frac{\exp(\eta_0 + \eta_1 x_k + \eta_2 x_k^2)}{1 + \exp(\eta_0 + \eta_1 x_k + \eta_2 x_k^2)}, \ k = 1, \dots, K.$$

In this latter case, we would still have $K-3$ degrees of freedom.

We can follow the same idea for a three-way table when we want to model the behaviour of the first binary variable as a function of the two other variables defining the table. In the drug example, one is interested in analysing the tendency of taking a psychotropic drug as a function of the gender category and of the age. Fix the number of observations in each cell $k\ell$ (i.e. $y_{\bullet k\ell}$), so that we have a binomial sampling process with an unknown parameter $p_{k\ell}$ for each cell. As for the two-way case above, we can either use ANOVA-like models for the logarithm of the odds and ANCOVA-like models when one (or both) of the two qualitative variables defining the K and/or L categories is a quantitative variable.

One may study the following ANOVA model for the logarithms of the odds

$$\log\left(\frac{p_{k\ell}}{1-p_{k\ell}}\right) = \mu + \eta_k + \zeta_\ell, \ k = 1, \dots, K, \ \ell = 1, \dots, L,$$

with $\eta = \zeta = 0$. As another example, if x_ℓ is a representative value (like the average age of the group) of the ℓth level of the third categorical variable, one might think of:

$$\log\left(\frac{p_{k\ell}}{1-p_{k\ell}}\right) = \mu + \eta_k + \zeta x_\ell, \ k = 1, \dots, K, \ \ell = 1, \dots, L, \tag{8.25}$$

with the constraint $\eta = 0$. Here also, interactions and the curvature effect for x_ℓ can be introduced, as shown in the following example. Since the cell totals $y_{\bullet k\ell}$ are considered as fixed, the log-likelihood to be maximised is:

$$\sum_{k=1}^{K} \sum_{\ell=1}^{L} \{ y_{1k\ell} \log p_{k\ell} + y_{2k\ell} \log(1 - p_{k\ell}) \}, \tag{8.26}$$

where $p_{k\ell}$ follows the appropriate logistic model.

Example 8.6 Consider again Example 8.4. One is interested in the influence of gender and age on drug prescription. Take the number of observations for each "gender-age group" combination, $y_{\bullet k \ell}$ as fixed. A logit model (8.25) can be used for the odds-ratios of the probability of taking drugs, where the value x_ℓ is the average age of the group. In the linear form it may be written as one of the two following equivalent forms:

$$\log\left(\frac{p}{1-p}\right) = \mathcal{X}\theta,$$

$$p = \frac{\exp(\mathcal{X}\theta)}{1+\exp(\mathcal{X}\theta)},$$

where $\theta = (\beta_0, \beta_1, \beta_2)^\top$ and the design matrix $\mathcal{X}$ is given by

$$\mathcal{X} = \begin{pmatrix} 1.0 & 1.0 & 23.2 \\ 1.0 & 1.0 & 36.5 \\ 1.0 & 1.0 & 54.3 \\ 1.0 & 1.0 & 69.2 \\ 1.0 & 1.0 & 79.5 \\ 1.0 & -1.0 & 23.2 \\ 1.0 & -1.0 & 36.5 \\ 1.0 & -1.0 & 54.3 \\ 1.0 & -1.0 & 69.2 \\ 1.0 & -1.0 & 79.5 \end{pmatrix}$$

The first column of $\mathcal{X}$ is for the intercept, the second is the coded variable for the two gender categories and the last column is the average of the ages for the corresponding age-group. Then we estimate β by maximising the log-likelihood function (8.26). We obtain:

$$\hat{\beta}_0 = -3.5612$$
$$\hat{\beta}_1 = -0.3426$$
$$\hat{\beta}_2 = \quad 0.0280,$$

the intercept for men is $\hat{\beta}_0 + \hat{\beta}_1 = -3.9038$ and for women is $\hat{\beta}_0 - \hat{\beta}_1 = -3.2186$, indicating a gender effect and the common slope for the positive age effect being $\hat{\beta}_2 = 0.0280$. The fit appears to be reasonably good. There are $K \times L = 2 \times 5 = 10$ free cells in the table. A saturated "full" model with ten parameters and a zero degree of freedom would involve a constant (one parameter) plus an effect for gender (one parameter) plus an effect for age (four parameters) and finally the

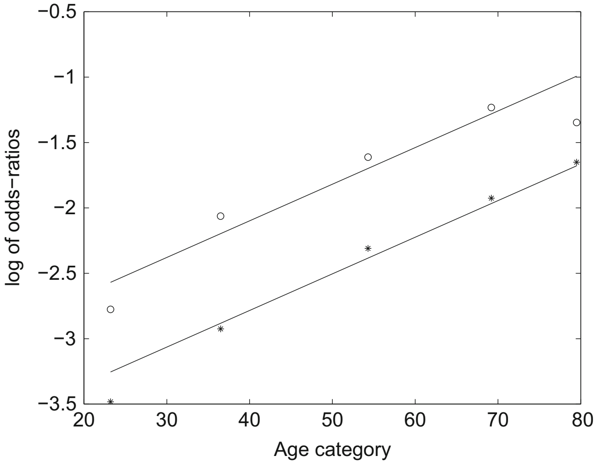

Fig. 8.3 Fit of the log of the odds-ratios for taking drugs: linear model for age effect with a "gender" effect (no interaction). Men are the *stars* and women are the *circles* MVAdruglogistic

interactions between gender and age (four parameters). The model retained above is a "reduced model" with only three parameters that can be tested against the most general saturated model. We obtain the value of the deviance $G^2_{H_0} = 11.5584$ with 7 degrees of freedom ($7 = 10 - 3$), whereas, $G^2_{H_1} = 0$ with no degree of freedom. This gives a p-value $= 0.1160$, so we cannot reject the reduced model.

Figure 8.3 shows how well the model fits the data. It displays the fitted values of the log of the odds-ratios by the linear model for the men and the women along with the log of the odds-ratios computed from the observed corresponding frequencies. It seems that the age effect shows a curvature. So we fit a model introducing the square of the ages. This gives the following design matrix:

$$\mathcal{X} = \begin{pmatrix} 1.0 & 1.0 & 23.2 & 538.24 \\ 1.0 & 1.0 & 36.5 & 1332.25 \\ 1.0 & 1.0 & 54.3 & 2948.49 \\ 1.0 & 1.0 & 69.2 & 4788.64 \\ 1.0 & 1.0 & 79.5 & 6320.25 \\ 1.0 & -1.0 & 23.2 & 538.24 \\ 1.0 & -1.0 & 36.5 & 1332.25 \\ 1.0 & -1.0 & 54.3 & 2948.49 \\ 1.0 & -1.0 & 69.2 & 4788.64 \\ 1.0 & -1.0 & 79.5 & 6320.25 \end{pmatrix}$$

The maximum likelihood estimators are:

$$
\begin{aligned}
\hat{\beta}_0 &= -4.4996 \\
\hat{\beta}_1 &= -0.3457 \\
\hat{\beta}_2 &= 0.0697 \\
\hat{\beta}_3 &= -0.0004.
\end{aligned}
$$

`MVAdruglogistic`

The fit is better for this more flexible alternative, giving a deviance $G^2_{H_1} = 3.3251$ with 6 degrees of freedom $(6 = 10 - 4)$. If we test H_0: no curvature for the age effect against H_1: curvature for the age effect, the reduction of the deviance is $G^2_{H_0} - G^2_{H_1} = 11.5584 - 3.3251 = 8.2333$ with one degree of freedom. The p-value $= 0.0041$, so we reject the reduced model (no curvature) in favour of the more general model with a curvature term.

We know already from Example 8.4 that second order interactions are not significant for this data set (the influence of age on taking a drug is the same for both gender categories), so we can keep this model as a final reasonable model to analyse the probability of taking the drug as a function of the gender and of the age. To summarise this analysis we end up saying that the probability of taking a psychotropic drug can be modelled as (with some abuse of notation)

$$
\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * \text{Sex} + \beta_2 * \text{Age} + \beta_3 * \text{Age}^2. \tag{8.27}
$$

Summary

- $\hookrightarrow$ In contingency tables, the categories are defined by the qualitative variables.
- $\hookrightarrow$ The saturated model has all of the interaction terms, and 0 degree of freedom.
- $\hookrightarrow$ The non-saturated model is a reduced model since it fixes some parameters to be zero.

Summary (continued)
↪ Two statistics to test for the full model and the reduced model are: $$X^2 = \sum_{k=1}^{K} (y_k - \hat{m}_k)^2 / \hat{m}_k$$ $$G^2 = 2 \sum_{k=1}^{K} y_k \ \log (y_k / \hat{m}_k)$$
↪ The logit models allow the column categories to be a quantitative variable, and quantify the effect of the column category by using fewer parameters and incorporating more flexible relationships than just a linear one.
↪ The logit model is equivalent to a log-linear model. $$\log [p(x_i) / \{1 - p(x_i)\}] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

8.3 Exercises

Exercise 8.1 *For the one factor ANOVA model, show that if the model is "balanced" ($n_1 = n_2 = n_3$), we have $\hat{\mu} = \bar{y}$. If the model is not balanced, show that $\bar{y} = \hat{\mu} + n_1\hat{\alpha}_1 + n_2\hat{\alpha}_2 + n_3\hat{\alpha}_3$.*

Exercise 8.2 *Redo the calculations of Example 8.2 and test if the main effects of the marketing strategy and of the location are significant.*

Exercise 8.3 *Redo the calculations of Example 8.3 with the Car Data set.*

Exercise 8.4 *Calculate the prediction interval for "classic blue" pullover sales (Example 3.2) corresponding to price* $= 120$.

Exercise 8.5 *Redo the calculations of the Boston housing example in Sect. 8.1.3*

Exercise 8.6 *We want to analyse the variations in the consumption of packs of cigarettes per month as a function of the brand (A or B), of the price per pack and as a function of the gender of the smoker (M or F). The data are below.*

y	Price	Gender	Brand
30	3.5	M	A
4	4	F	B
20	4.1	F	B
15	3.75	M	A
24	3.25	F	A
11	5	F	B
8	4.1	F	B
9	3.5	M	A
17	4.5	M	B
1	4	F	B
23	3.65	M	A
13	3.5	M	A

1. *In addition to the effects of brand, price and gender, test if there is an interaction between the brand and the price.*
2. *How would the design matrix of a full model with all the interactions between the variables appear? What would be the number of degrees of freedom of such a model?*
3. *We would like to introduce a curvature term for the price variable. How can we proceed? Test if this coefficient is significant.*

Exercise 8.7 *In the drug Example 8.4, test if the first order interactions are significant.*