3.4 Log-linear regression

1. Introduction

The term **log-linear models** is used to describe any model of the form

$$\log \mathbb{E}(\mathrm{Y}_i) = c + \mathbf{x}_i^\top \boldsymbol{\beta} \tag{3.4.1}$$

In this section, we will analyse the introduction of interaction terms and the analogous of the ANOVA for log-linear models.

2. Modelling example

Example:

Let's consider a cross-sectional study of patients with a form of skin cancer called "malignant melanoma". We have 400 patients, with information about the site of the tumour and its histological type. The data are collected into the **melanoma** dataset available in the **dobson** package in **R**.

	Site			
Tumor type	Head & neck	Trunk	Extrem -ities	Total
Hutchinson's melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

The table represents a **contingency table**. Let's call the probability of being in cell (j,k) θ_{jk}. Then, in case of no association

$$\theta_{jk} = \theta_{j.}\theta_{.k} \qquad j = 1, \dots, J \quad \text{and} \quad k = 1, \dots, K \tag{3.4.2}$$

i.e, in the case of **independence**

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \log n + \log \theta_{j.} + \log \theta_{.k} \tag{3.4.3}$$

which can be compared with the **dependent** model, i.e.

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \log n + \log \theta_{jk} \tag{3.4.4}$$

Analogously to the ANOVA model, we should *introduce the factors relative to the single predictors/factors*

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} \tag{3.4.5}$$

where $(\alpha\beta)_{jk}$ represents a coefficient relative to the *interaction term*; therefore, to test for independence we can compare $(3.4.5)$ with

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \mu + \alpha_j + \beta_k \tag{3.4.6}$$

or with the minimal model

$$\log \mathbb{E}(Y)_{jk} = \mu \tag{3.4.7}$$

The specification of log-linear models is **hierarchical**: if the higher-order term (interaction) is included in the model, all the lower-order terms are included as well.

Warning: This means that, in many cases, log-linear models have many parameters: **constraints may be needed!**

While several distributions can be used (**Think:** Can you think of anyone?), **Poisson distributions can be assumed**. Therefore, all standard methods for GLM can be applied (weighted least squares, goodness-of-fit statistics like P^2 and D, Pearson and deviance residuals).

2.1 The saturated model

The **saturated model** is given by

```
library(dobson)

data("melanoma")

ressat.melanoma <- glm(frequency ~ site*type, family=poisson(), data=melanoma)
summary(ressat.melanoma)
```

2.2 The model with no interactions

The model with **no interaction terms** is given by

```
library(dobson)

data("melanoma")

resadd.melanoma <- glm(frequency ~ site + type, family=poisson(), data=melanoma)
summary(resadd.melanoma)
```

2.3 The minimal model

The **minimal** is given by

```
library(dobson)

data("melanoma")

resmin.melanoma <- glm(frequency ~ 1, family=poisson(), data=melanoma)
summary(resmin.melanoma)
```

2.4 Expected frequencies

For the reference category **`type:Hutchinson's melanotic freckle`** on **`site:extremities`** the expected frequencies are

- minimal model: $e^{3.507} = 33.33$
- additive model: $e^{2.9554} = 19.21$
- saturated model: $e^{2.3026} = 10.00$

Note: the expected frequencies for the saturated model correspond to the observed frequencies.

For **`type:indeterminate`** tumours on **`site:head-neck`** the expected frequencies are

- minimal model: $e^{3.507} = 33.33$
- additive model: $e^{2.9554-1.2010+0.499} = 9.520049$
- saturated model: $e^{2.3026+0.7885+1.0296-1.7228} = 11.000$

Again the expected frequencies for the saturated model correspond to the observed frequencies.

For **`type:nodular`** tumours on **`site:trunk`** the expected frequencies are

- minimal model: $e^{3.507} = 33.33$
- additive model: $e^{2.9554-0.7571+1.3020} = e^{3.5003} = 33.12$
- saturated model: $e^{2.3026-1.6094+1.9879+0.8155} = e^{3.4966} = 33.00$