Advanced Statistical Modelling

Dr. S. Jackson

2023-12-07

Contents

Welcome

Welcome to the material for the first term of the module Advanced Statistical Modelling MATH3411 at Durham University. These pages will update as the course progresses, consisting of relevant lecture notes, practical demonstrations (in R), exercise sheets and practical sessions.

I would recommend that you use the html version of these notes (they have been designed for use in this way), however, there is also a pdf version of these notes, which will also be updated as the course progresses.

If you would like to contact me regarding any of the material in this course, then my email address is samuel.e.jackson@durham.ac.uk

Further practical details about the course are covered in the *Introduction* section.

Acknowledgements

Some of the notes on Categorical Data Analysis have been influenced by notes from previous courses given by Prof. P. Craig and Dr. G. Karagiannis. The notes on GLMs are based on previous courses originally designed by Prof. J. Einbeck, and updated by Prof. I. Jermyn, Dr. L. Aslett, Dr. R. Drikvandi and Dr. R. Crossman over the years.

Chapter 1

Motivation, Introduction and Review

1.1 Course Content

In these lectures, we are going to discuss:

Categorical Data Analysis: we will study use of

- *Contingency Tables:* We will learn about the different types of possible associations between two or more categorical variables, before learning how to make inferences about such associations. As part of this, we will discover the importance of different sampling schemes and learn how to make statements about associations between categorical variables utilising odds and odds ratios.
- *Log-Linear Models:* A way of modelling the counts for categorical data represented in contingency tables, with the categorical variables themselves being the predictors. The construction of such models is highly useful for testing and representing (modelling) the associations between the categorical variables themselves.

Generalised Linear Models: a broad class of models permitting

- the response variable y to come from a space that is a subset of the real numbers, e.g. the positive real numbers, or countable subsets such as $\mathbb{N}$ or the set $\{0, 1\}$.
- a larger set $\mathcal{F}$ of functions for making predictive statements about response y based on predictor variables $\boldsymbol{x}$ (not just linear combinations of the predictors). These will be formed by combining the linear functions used in linear models with *nonlinear* maps from $\mathbb{R} \to \mathbb{R}$.
- consideration of general families of probability distributions for response y (not just normal).

We will then go on to develop the standard tools of classical statistics:

1. Estimators for the parameters of the model: we will use maximum likelihood.
2. Sampling distributions for these estimators, either exactly or approximately, and especially their expectations and variances.

3. Statistical tests and confidence intervals; predictions with corresponding prediction intervals.

4. Measures of "goodness-of-fit" for model selection.

1.1.1 Section Overview

In the rest of this introductory chapter, we present a review of some of the most useful material from previous prerequisite courses. At the end of this section, there is a set of references for the course, which will allow you to take your study of Advanced Statistical Modelling deeper than just being restricted to what is contained in these lecture notes alone (brilliant though they are, of course!).

1.2 Statistical Inference II: A Review

This section summarises some of the key topics from Statistical Inference II. Review the corresponding lecture notes (particularly the Chapter on *Likelihood Inference*) for further detail.

1.2.1 Statistical Inference

- A *statistical model* (also known as a family of distributions) is a set of distributions (or densities) M.
- A *parametric model* is any model M that can be parameterised by a finite number of parameters.

1.2.2 Likelihood Inference

1.2.2.1 Maximum Likelihood Estimation

- Suppose data $\boldsymbol{X} = (X_1, ..., X_n)^T$ has joint density $f(\boldsymbol{x}; \theta)$. Given a statistical model parameterised by a fixed and unknown scalar $\theta \in \Omega \subseteq \mathbb{R}$, the likelihood function, $L(\theta)$, is the probability (density) of the observed data considered as a function of θ

$$L(\theta) = L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta) \tag{1.1}$$

  and the log-likelihood function is its logarithm

$$l(\theta) \equiv l(\theta; \boldsymbol{x}) = \log L(\theta) = \log f(\boldsymbol{x}; \theta) \tag{1.2}$$

- The *Maximum Likelihood Estimate (MLE)* $\hat{\theta}$ for θ given the data $\boldsymbol{x}$ is the value of θ which maximises the likelihood over the parameter space

$$\hat{\theta} = \arg\max_{\theta \in \Omega} l(\theta) = \arg\max_{\theta \in \Omega} L(\theta) \tag{1.3}$$

- If $X_1, ..., X_n$ are i.i.d. then

$$L(\theta) = f(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{1.4}$$

$$l(\theta) = \log \prod_{i=1}^{n} f(x_i; \theta) = \sum_{i=1}^{n} \log f(x_i; \theta) \tag{1.5}$$

1.2.2.2 Properties of the MLE

- If $\hat{\theta}$ is the MLE of θ and $\phi = g(\theta)$ is a function of θ, then $\hat{\phi} = g(\hat{\theta})$.
- A statistic $S = S(\boldsymbol{X})$ is a sufficient statistic for θ if the conditional distribution of $\boldsymbol{X}|S = s$ does not depend on θ.
- **Factorisation Theorem:** A statistic $S = S(\boldsymbol{X})$ is sufficient for θ if and only if there exist functions $g(S, \theta)$ and $h(\boldsymbol{X})$ such that for all $\boldsymbol{X}$ and θ:

$$f(\boldsymbol{X}; \theta) = g(S, \theta)\, h(\boldsymbol{X}) \tag{1.6}$$

  where $h(\boldsymbol{X})$ does not depend on θ.

- If $S = S(\boldsymbol{X})$ is a sufficient statistic of $\boldsymbol{x}$ for θ, and the MLE $\hat{\theta}$ of θ exists, then

$$\hat{\theta}(\boldsymbol{x}) = \hat{\theta}(s(\boldsymbol{x})) \tag{1.7}$$

1.2.3 (Frequentist) Hypothesis Testing

1.2.3.1 Hypotheses

- A hypothesis is any statement about the probability model $M = \{f(\boldsymbol{x}|\theta)|\theta \in \Omega\}$ generating the data. When the probability distribution of the model is fixed, then a hypothesis is simply a statement about the parameter θ[1].
- A hypothesis is simple if it completely specifies the associated probability distribution, $f(\boldsymbol{x}|\theta)$. Otherwise, a hypothesis is called composite (or compound).
- The null hypothesis, $\mathcal{H}_0$, is a conservative hypothesis, not to be rejected unless evidence from the data is clear. The alternative hypothesis, $\mathcal{H}_1$, specifies a particular departure from the null hypothesis that is of interest.
- A hypothesis test is a procedure for deciding whether to reject $\mathcal{H}_0$ or not.

1.2.3.2 Hypothesis Test

- We define a test by a critical or rejection region, $\mathcal{R}$, which partitions the sample space $\mathcal{X}$ such that:
  - If $\boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{R}$, then we say that the test rejects $\mathcal{H}_0$.
  - If $\boldsymbol{x} = (x_1, ..., x_n) \notin \mathcal{R}$, then the test fails to reject $\mathcal{H}_0$.

[1]We will here assume that θ is a scalar parameter

- A test statistic, T, is a statistic derived from the sample used in hypothesis testing. The sampling distribution of T under the null hypothesis is known as the null distribution.
- We can define the rejection region in terms of values of the test statistic

$$\mathcal{R} = \{\boldsymbol{x}|T(\boldsymbol{x}) \in \mathcal{T}\} \tag{1.8}$$

for some set of possible values $\mathcal{T}$. Usually, once we find a test statistic, we just ignore the original critical region and just focus on whether or not T is in $\mathcal{T}$ or not.

- Assuming $\mathcal{H}_0$ is true, we can construct the sampling distribution for T, $f_T(t)$ – known as the null distribution for T – and define $\mathcal{R}$ as corresponding to the set of values of T which are "sufficiently unlikely" to occur under $\mathcal{H}_0$ to cause us to reject $\mathcal{H}_0$.

1.2.3.3 Significance, Power and p-values

- For a simple null hypothesis, the significance level of a test is the probability

$$\alpha = P(\text{reject } \mathcal{H}_0|\mathcal{H}_0 \text{ true}) \tag{1.9}$$

Then

$$\mathcal{R} = \{\boldsymbol{x}|P(T(\boldsymbol{x}) \in \mathcal{T}|\mathcal{H}_0) \leq \alpha\} \tag{1.10}$$

- Two errors can arise in hypothesis testing:
  1. Reject $\mathcal{H}_0$ when $\mathcal{H}_0$ true - a type I error, occurs with probability α. If $\mathcal{H}_0$ is simple, this is the usual significance level.
  2. Do not reject $\mathcal{H}_0$ when $\mathcal{H}_1$ true - a type II error, occurs with probability $\boldsymbol{\beta}$.
- Given an observed sample $\boldsymbol{X} = \boldsymbol{x}$ and a simple null hypothesis $\mathcal{H}_0$, the p-value or observed significance is the smallest value of α for which we would reject the null hypothesis $\mathcal{H}_0$ on the basis of the data $\boldsymbol{x}$.
- A p-value is the probability of observing a test statistic T "at least as extreme" as the one observed, t, under the null distribution:

$$p = P(T \text{ "at least as extreme as" } t|\mathcal{H}_0) \tag{1.11}$$

What "extreme" is depends on the nature of the alternative hypothesis $\mathcal{H}_1$ and the p-value quantifies the extent to which the null hypothesis is violated.

- There is a duality of confidence intervals and hypothesis tests[2].
  - A $100(1-\alpha)\%$ confidence region for a parameter θ consists of all of those values of θ_0 for which the hypothesis that $\theta = \theta_0$ will not be rejected at level α.
  - Equivalently, the hypothesis that $\theta = \theta_0$ is accepted if θ_0 lies in the same confidence region.

[2]See the corresponding lecture notes (Theorem 6.3) for the mathematical exposition of this theorem.

1.2.3.4 Likelihood Ratio Tests

- Given a sample $\boldsymbol{X} = (X_1, ..., X_n)$ and the general hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, the distribution of $\boldsymbol{X}$ under $\mathcal{H}_0$ is the null distribution denoted $f_0(X)$ and the distribution under $\mathcal{H}_1$ is the alternative distribution $f_1(X)$.
- Suppose $X \sim f(X|\boldsymbol{\theta})$. The *likelihood ratio (LR)* for comparing the two simple hypotheses $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\mathcal{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ is

$$\lambda(\boldsymbol{X}) = \frac{L(\boldsymbol{\theta}_1; \boldsymbol{X})}{L(\boldsymbol{\theta}_0; \boldsymbol{X})} \tag{1.12}$$

- Suppose $X \sim f(X|\boldsymbol{\theta})$. The *generalised likelihood ratio (GLR)* test of $\mathcal{H}_0 : \boldsymbol{\theta} \in \Omega_0$ against $\mathcal{H}_1 : \boldsymbol{\theta} \in \Omega_1$, where $\Omega_0 \cap \Omega_1 = \emptyset$ and $\Omega_0 \cup \Omega_1 = \Omega$ computes

$$\Lambda(\boldsymbol{X}) = \frac{\max_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta}; \boldsymbol{X})}{\max_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}; \boldsymbol{X})} \tag{1.13}$$

  and rejects $\mathcal{H}_0$ when $\Lambda(\boldsymbol{X}) < c$ where c is chosen such that $P(\Lambda(\boldsymbol{X}) \leq c|\mathcal{H}_0) = \alpha$.
- **Wilks' Theorem:**[3] Consider an i.i.d. sample $\boldsymbol{X} = (X_1, ..., X_n)$ from $f(X_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ contains k unknown parameters. Suppose that $\mathcal{H}_0$ specifies the values for $\nu = k - k_0$ of the parameters, so that k_0 parameters are still unknown. Then, under some regularity conditions and given $\mathcal{H}_0$, as $n \to \infty$:

$$W = -2 \log \Lambda(\boldsymbol{X}) \dot{\sim} \chi^2_\nu \tag{1.14}$$

  Thus for large n a GLR test will reject $\mathcal{H}_0$ at approximate significance level α if

$$-2 \log \Lambda(\boldsymbol{X}) \geq \chi^{2,\star}_{\nu,\alpha} \tag{1.15}$$

1.3 Statistical Modelling II: A Review

This section summarises some of the key topics from Statistical Modelling II. Review the corresponding lecture notes for further detail.

1.3.1 Supervised and Unsupervised Learning

- **Unsupervised Learning:** investigate properties of a data structure $\boldsymbol{Z}$. Sections 2 and 3 may be seen as unsupervised learning.
- **Supervised Learning:** investigate how $\boldsymbol{Z}$ affects reponse variable $\boldsymbol{Y}$. This is the main focus of Section 4 and beyond.

[3] For the purposes of this course, we will use this theorem without proof. A sketch of the proof was outlined in Statistical Inference II. Anyone who is interested in Wilks' original paper on the theorem can consult Wilks [1938].

1.3.2 Linear Models

- The linear model in matrix form is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.16}$$

  where
  - $\boldsymbol{Y}^T = (y_1, ..., y_n)$ is the vector of responses,
  - $\boldsymbol{X}$ is the design matrix,
  - $\boldsymbol{\beta}^T = (\beta_1, ..., \beta_p)$ is the p-dimensional parameter vector,
  - $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ is the vector of errors.
- Taking the i^{th} row of Equation (1.16), one can represent

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \tag{1.17}$$

Note that the predictor variables $x_j, j = 1, ..., p$ (each of which corresponds to a column of $\boldsymbol{X}$) may be transformations or functions of the covariate variables which constitute the data matrix $\boldsymbol{Z}$. In other words, $\boldsymbol{X}$ is not necessarily equal to $[1, \boldsymbol{Z}]$.

1.3.2.1 Assumptions

- (A1): Linearity: $\mathrm{E}[\epsilon_i] = 0$, or $\mathrm{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}^T\boldsymbol{\beta}$
- (A2'): Homoscedasticity and (A2"): Independence[4]: $\mathrm{Cov}[\epsilon_i, \epsilon_j] = \mathcal{I}_{i=j}\sigma^2$
- (A3): Normality: ϵ_i is normally distributed.

In other words:

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I) \tag{1.18}$$

These assumptions can be diagnosed using the diagnostics discussed in Section 1.3.6.

1.3.2.2 Estimation of Model Parameters

- LS estimates $\hat{\boldsymbol{\beta}}$ minimise

$$R(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1.19}$$

  and satisfy

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y}) \tag{1.20}$$

- An unbiased estimate of σ^2, based on known residuals $\hat{\epsilon}_i$, is

$$s^2 = \frac{\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}}{n - p} \tag{1.21}$$

[4] $\mathcal{I}_T$ is the indicator function taking value 1 if T holds and 0 otherwise.

1.3.2.3 Properties of $\hat{\boldsymbol{\beta}}$ and s^2

- $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- $\mathrm{Var}[\boldsymbol{c}^T\hat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}$
- The standard deviation (SD) of $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ as a measure of the precision of the estimate $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ is only useful if the value of σ is known. When it is not known we replace it by its estimate s and one obtains the standard error of $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$:

$$\mathrm{SE}[\boldsymbol{c}^T\hat{\boldsymbol{\beta}}] = s\sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}} \tag{1.22}$$

- $\mathrm{E}[s^2] = \frac{\mathrm{E}[\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}]}{n-p} = \sigma^2$

1.3.2.4 Sampling Distribution

- The *sampling distribution* of a parameter estimator is the probability distribution of the estimator, when drawing repeatedly samples of the same size from the population and observing the value of the estimator for each sample.
- Assuming (A1)-(A3), then

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{1.23}$$

and

$$\boldsymbol{c}^T\hat{\boldsymbol{\beta}} \sim \mathcal{N}_1(\boldsymbol{c}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}) \tag{1.24}$$

- However, σ^2 is usually unknown. We have that

$$\frac{1}{\sigma^2}(n-p)s^2 \sim \chi^2_{n-p} \tag{1.25}$$

Therefore,

$$\frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\mathrm{SE}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}})} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}} \equiv t_{n-p} \tag{1.26}$$

In particular

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-p} \tag{1.27}$$

1.3.2.5 Confidence Intervals

- A $(1-\alpha)$ level confidence interval (CI) for population parameter θ is an interval calculated from sample $\boldsymbol{Y}^T = (y_1, ..., y_n)$ which contains θ with sampling probability $1-\alpha$, in the sense that if we take very many samples, and form an interval from each sample, then proportion $1-\alpha$ will contain θ.
- A $1-\alpha$ CI for $\boldsymbol{c}^T\boldsymbol{\beta}$ is given by

$$\boldsymbol{c}^T\hat{\boldsymbol{\beta}} \mp t_{n-p,\frac{\alpha}{2}}\mathrm{SE}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}) \tag{1.28}$$

1.3.2.6 Hypothesis Testing

- Hypothesis Test: reject
$$\mathcal{H}_0 : \beta_j = \beta_j^0$$
in favour of
$$\mathcal{H}_1 : \beta_j \neq \beta_j^0$$
at significance level α if
$$T = \left| \frac{\hat{\beta}_j - \beta_j^0}{\text{SE}(\hat{\beta}_j)} \right| > t_{n-p,\frac{\alpha}{2}} \tag{1.29}$$
or equivalently if $p < \alpha$, where
$$p = P(|t_{n-p}| > T_{\text{observed}}) \tag{1.30}$$
is the p-value.

1.3.2.7 Prediction

- For a new predictor $\boldsymbol{x}_0^T = (x_{01}, ..., x_{0p})$, the predicted value[5] $\hat{y}_0$ is
$$\widehat{\text{E}[y_0|\boldsymbol{x}_0]} = \hat{y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} + \hat{\epsilon}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \tag{1.31}$$
- We have that[6]
$$\text{Var}[Y_0 - \hat{Y}_0] = \text{Var}[Y_0] + \text{Var}[\hat{Y}_0] = \sigma^2(1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0) \tag{1.32}$$

1.3.3 Factors

- Categorical covariates are called *factors*.
- For inclusion into a linear model, factors need to be *coded*. For factor $\mathcal{A}$ with levels $1, ..., a$, we define
$$x_i^{\mathcal{A}} = \mathcal{I}_{\mathcal{A}=i} \tag{1.33}$$
- Including all the indicators into the linear model, one gets the *unconstrained model*.
$$\text{E}[y|\mathcal{A}] = \beta_0 + \beta_1 x_1^{\mathcal{A}} + ... + \beta_a x^{\mathcal{A}} \tag{1.34}$$
- We need constraints on the parameters to give us a *constrained model*. Popular constraints include:
  - set $\beta_i = 0$ for a single i; now level i takes the role of a reference category represented by the intercept. For example, if we set $\beta_1 = 0$ we have
  $$\text{E}[y|\mathcal{A}] = \beta_0 + \beta_2 x_2^{\mathcal{A}} + ... + \beta_a x^{\mathcal{A}} \tag{1.35}$$
  - The zero-sum constraint: $\sum_{i=1}^{a} \beta_i = 0$.
- Under one of these constraints, X effectively becomes an $n \times a$ matrix with rank a, and so X^TX is now invertible. β becomes an a-vector.

[5] Terminology: The term *fitted value* is typically used for an estimate of a piece of data used to fit the model, and *predicted value* is typically used for an estimate of a piece of data not used to fit the model, but they are mathematically the same.

[6] Recall that Y_0 and $\hat{Y}_0$ are independent since $\hat{Y}_0$ is a linear combination of $y_1, ..., y_n$.

1.3.3.1 Additive Model

- Additive effect of $\mathcal{A}$ and $\mathcal{B}$:

$$y_{ijk} = \mu + \tau_i^{\mathcal{A}} + \tau_i^{\mathcal{B}} + \epsilon_{ijk} \tag{1.36}$$

  where
  - $\tau_i^{\mathcal{A}}$ is the main effect of level i of factor $\mathcal{A}, i = 1, ..., a$
  - $\tau_j^{\mathcal{B}}$ is the main effect of level j of factor $\mathcal{B}, j = 1, ..., b$
  - μ is the intercept term
  - ϵ_{ijk} is the error of the k^{th} replicate of treatment (i, j).

- Constraints: $\tau_1^{\mathcal{A}} = \tau_1^{\mathcal{B}} = 0$.

1.3.3.2 Interaction Model

- We have

$$y_{ijk} = \mu + \tau_i^{\mathcal{A}} + \tau_i^{\mathcal{B}} + \tau_{i,j}^{\mathcal{A},\mathcal{B}} + \epsilon_{ijk} \tag{1.37}$$

  where $\tau_{i,j}^{\mathcal{A},\mathcal{B}}$ is the interaction effect of level i of factor $\mathcal{A}$ with level j of factor $\mathcal{B}$.

- In data there will virtually always be non-zero estimates of $\tau_{i,j}^{\mathcal{A},\mathcal{B}}$, so we often wish to assess whether or not they are significantly different from zero (and thus worth including in the model).

1.3.4 Analysis of Variance (ANOVA)

- We have, assuming (A1)-(A3) and that the model has an intercept,

$$SST = SSR + SSE \tag{1.38}$$

  where
  - $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

- $R^2 = \frac{SSR}{SST}$ is the *coefficient of determination*, with $0 \leq R \leq 1$ since $SSR \leq SST$. This will always increase as we add more terms to the model.

- $R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$ is known as the *adjusted R-squared*. This quantity penalises for large values of p.

- Sequential ANOVA involves comparing m *nested* models $M_1 \subset ... \subset M_m$, with design matrices $X_1, ..., X_m$, where X_j is $n \times p_j$ and X_{j+1} is obtained by adding $p_{j+1} - p_j$ columns to X_j, using a series of partial F-tests.
  Recall that model M_j is nested in M_{j+1} if they can be written as

$$M_j : \quad \mathrm{E}[y|\mathcal{A}] = \beta_1 + \beta_2 x_2^{\mathcal{A}} + ... + \beta_{p_j} x_{p_j} \tag{1.39}$$

$$M_{j+1} : \quad \mathrm{E}[y|\mathcal{A}] = \beta_1 + \beta_2 x_2^{\mathcal{A}} + ... + \beta_{p_j} x_{p_j} + \beta_{p_{j+1}} x_{p_{j+1}} \tag{1.40}$$

1.3.5 Model Selection

- Model selection involves finding a good submodel by selecting a subset of indices I of those terms to be included in the submodel, with D the remainder to be deleted, so that

$$\mathrm{E}[y|x] = \boldsymbol{x}_I^T \beta_I + \boldsymbol{x}_D^T \beta_D \tag{1.41}$$

  and

$$\mathrm{E}_I[y|x] = \boldsymbol{x}_I^T \beta_I \tag{1.42}$$

- A good submodel will have
  - RSS_I "not much larger" than RSS.
  - p_I as small as possible.
- Selection criteria include
  - Mallow's C_I, which seeks to minimise

$$C_I = \frac{RSS_I}{s^2} + 2p_I - n \tag{1.43}$$

  - AIC, which seeks to minimise

$$AIC_I = -2l(\hat{\boldsymbol{\beta}}_I; Y) + 2p_I \tag{1.44}$$

- Selection methods include forward selection, backward elimination and stepwise selection.

1.3.6 Diagnostics

- *Residuals* are used to check linear model assumptions.
- *influential observations* are those cases with a particularly large impact on $\hat{\boldsymbol{\beta}}$, s^2 or both.
- We have that

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{\epsilon} \tag{1.45}$$

  where $\boldsymbol{H}$ is the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the $n \times n$ *hat* matrix.

1.3.6.1 Leverage Values and Studentised Residuals

- $h_i = H_{ii} = (\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i)$ is called the *leverage value* of y_i, and measures the impact of case i onto its own fitted value.
- We have that

$$\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\epsilon} \tag{1.46}$$

  If (A1) - (A3) hold, then $\hat{\boldsymbol{\epsilon}} \sim \mathcal{N}_n(0, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}))$.
- *Studentised residuals* are given by

$$r_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_i}} \tag{1.47}$$

- *Internally studentised residuals* are when all data, including case i, are used to estimate σ, and *externally studentised residuals* are when case i is removed to estimate σ (in which case the estimate s is different for each i).

1.3.6.2 Influence Analysis

- A case i is called *influential* if $\hat{\boldsymbol{\beta}}$ or s^2 change "substantially" when removing it.
- Note that if $h_i \to 1$, $\hat{\epsilon}_i \approx 0$, thus cases with $h_i \approx 1$ may be highly influential, thus the h_i are sometimes called *potential influence* values.
- We can test for influential values by comparing the estimate $\hat{\boldsymbol{\beta}}_{(i)}$, which is the estimate with the i^{th} case omitted, with the fit using all of the data, $\hat{\boldsymbol{\beta}}$.
- An observation is often said to be
  - *potentially influential* if $h_i \geq \frac{2p}{n}$
  - *outlying* if $|r_i| > 2$ (or 3).

1.3.6.3 Diagnostic Plots

- Various plots can aid diagnose the validity of assumptions (A1)-(A3) (see Section 1.3.2.1).
- Plot $\hat{\epsilon}_i$ against x_{ij} and $\hat{y}_i$; $i = 1, ..., n$; $j = 1, ..., p$.
  - if there are "patterns", this indicates violation of (A1).
  - if the spread of $\hat{\epsilon}_i$ varies, this indicates violation of (A2').
- Plot ordered values of $\hat{\epsilon}_i$ (or r_i) versus the corresponding quantities of a standard normal distribution $u_i = \Phi^{-1}\left(\frac{i-0.5}{n}\right)$. Deviation from a straight line indicates non-normality (violation of (A3)).
- For (A2"), we can plot $\hat{\epsilon}_i$ against i (or t_i - the time at which case i is observed). If there is a pattern, this could indicate violation of (A2"), but also possibly indication of violation of (A1) instead/as well.

1.4 Further Review Topics

1.4.1 Types of Variables

It is useful to refresh ourselves about the different types of variables we might encounter. Variables take values in sets with different structures;

- pure sets,
- topological algebras (such as the real numbers $\mathbb{R}$),
- subsets of these sets (such as the positive real numbers $\mathbb{R}^+$).

Different types of variable are described in Figure 1.1.

1.4.2 Poisson Distribution

The Poisson distribution is a simple distribution for count data, for which integer values $x \in \{0, 1, 2, ...\}$ and parameter $\lambda > 0$ has the form

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{1.48}$$

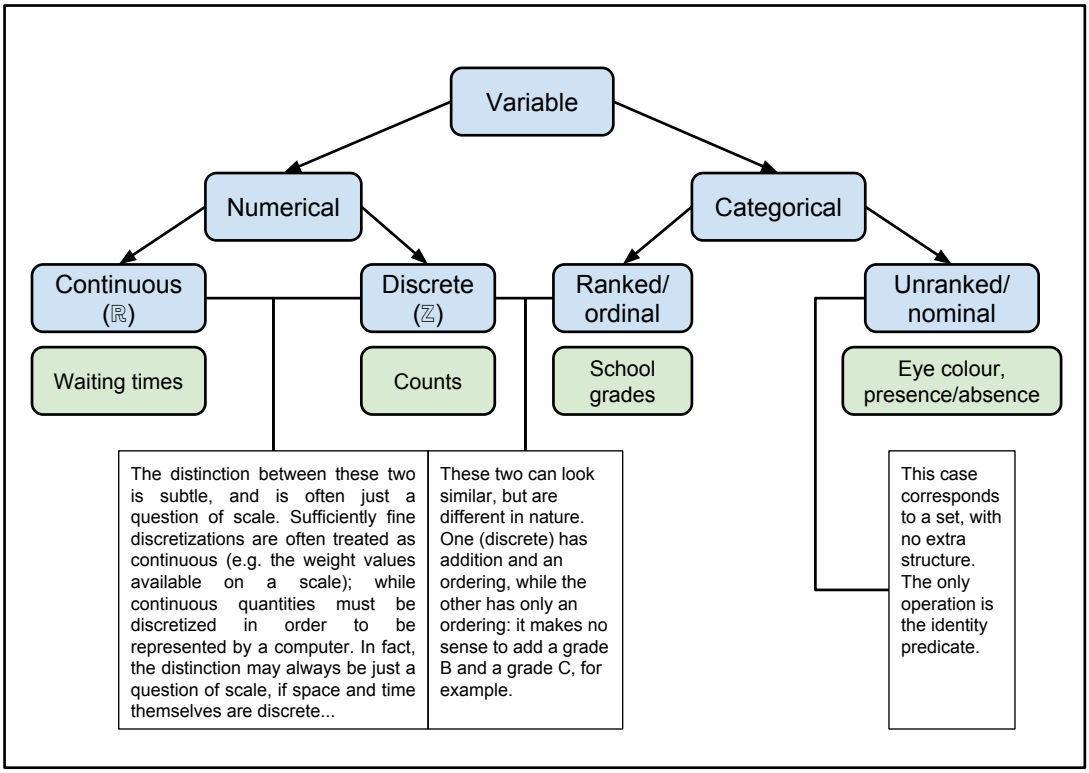


Figure 1.1: Diagram of the different types of variables.

We notate this

$$X \sim Poi(\lambda) \tag{1.49}$$

1.4.2.1 Mean and Variance

The mean and variance of X, distributed according to $Poi(\lambda)$, are given by

$$\mathrm{E}[X] = \mathrm{Var}[X] = \lambda \tag{1.50}$$

1.4.2.2 Sum of Poisson Variables

If $X_1, ..., X_K$ are independent Poisson variables, with parameters $\lambda_1, ..., \lambda_K$ respectively, then their sum

$$\sum_{k=1}^{K} X_k \sim Poi(\sum_{k=1}^{K} \lambda_k) \tag{1.51}$$

1.4.3 Multinomial Distribution

Assume random variables $\boldsymbol{Y} = (Y_1, ..., Y_k)^T$, such that $Y_j \in \{0, ..., n\}$ and $\sum_{j=1}^{k} Y_j = n$. We say that $\boldsymbol{Y}$ follows a Multinomial distribution with parameters

- $n > 0$, the number of trials
- $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)^T$, the success vector of probabilities such that $\pi_i \in (0, 1)$ and $\sum_{j=1}^{k} \pi_j = 1$

if

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{n!}{y_1!...y_k!}\pi_1^{y_1}...\pi_k^{y_k}\mathcal{I}_{\boldsymbol{y}\in\mathcal{Y}} \tag{1.52}$$

where

- $\boldsymbol{y} = (y_1, ..., y_k)^T$
- $\mathcal{Y} = \{\boldsymbol{y} \in \{0, ..., n\}^k | \sum_{j=1}^k y_j = n\}$

We notate this as

$$\boldsymbol{Y} \sim Mult(n, \boldsymbol{\pi}) \tag{1.53}$$

1.4.3.1 Mean and Variance

The expectation is[7]

$$\mathrm{E}[Y_j] = n\pi_j, \qquad j = 1, ..., k \tag{1.54}$$
$$\mathrm{E}[\boldsymbol{Y}] = n\boldsymbol{\pi} \tag{1.55}$$

The covariance structure is[8]

$$\mathrm{Var}[Y_j] = n\pi_j(1 - \pi_j) \tag{1.56}$$
$$\mathrm{Cov}[Y_j, Y_{j'}] = -n\pi_j\pi_{j'} \qquad (j \neq j') \tag{1.57}$$
$$\mathrm{Var}[\boldsymbol{Y}] = n(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T) \tag{1.58}$$

1.4.3.2 Collapsing Categories

Collapsing categories of a Multinomial distribution leads to another Multinomial distribution with fewer categories.
For example, take a multinomial distribution with seven categories $A_1, ..., A_7$ such that

$$(N_1, ..., N_7) \sim Mult(n; (\pi_1, ..., \pi_7)) \tag{1.59}$$

Now, if we combine the categories as follows; $B_1 = \{A_1, A_2, A_3, A_4\}$, $B_2 = A_5$, and $B_3 = \{A_6, A_7\}$, then $(Y_1, Y_2, Y_3) = (N_1+N_2+N_3+N_4, N_5, N_6+N_7)$ follow a multinomial distribution

$$(Y_1, Y_2, Y_3) \sim Mult(n; (\pi_1^*, \pi_2^*, \pi_3^*)) \tag{1.60}$$

with $(\pi_1^*, \pi_2^*, \pi_3^*) = (\pi_1 + \pi_2 + \pi_3 + \pi_4, \pi_5, \pi_6 + \pi_7)$.

1.4.3.3 Distribution of Outcome Subsets

Suppose we are interested in the first $Q < K$ outcomes[9], then

$$(N_1, ..., N_Q) \sim Mult(n_Q; (\pi_1/\pi_Q, ..., \pi_q/\pi_Q)) \tag{1.61}$$

where $n_Q = \sum_{i=1}^q n_i$ and $\pi_Q = \sum_{i=1}^q \pi_i$.

[7] Q1-3 concerns proving the results stated in this section.
[8] Note that $\mathrm{diag}(\boldsymbol{\pi})$ is the $k \times k$ diagonal matrix with diagonal elements taking the values of the vector $\boldsymbol{\pi}$.
[9] or indeed, any collection of $Q < K$ outcomes.

1.4.3.4 Poisson and Multinomial Distributions

The conditional distribution of independent Poisson variables $X_1, ..., X_K$ given that their sum $\sum_{k=1}^{K} X_k = n$ is

$$P((X_1 = n_1, ..., X_K = n_K)|\sum_{k=1}^{K} X_k = n) = \frac{P(X_1 = n_1, ..., X_K = n_K)}{P(\sum_{k=1}^{K} X_K = n)} \tag{1.62}$$

$$= \frac{\prod_{k=1}^{K}(e^{-\lambda_k}\lambda_k^{n_k})/n_k!}{(e^{-\lambda}\lambda^n)/n!} \tag{1.63}$$

where we have used the result of Section 1.4.2.2 in the denominator and defined $\lambda = \sum_{k=1}^{K} \lambda_k$. Since $\sum_{k=1}^{K} n_k = n$, and $\prod_{k=1}^{K} e^{-\lambda_k} = e^{-\sum_{k=1}^{K} \lambda_k} = e^{-\lambda}$, we have that

$$P((X_1 = n_1, ..., X_K = n_K)|\sum_{k=1}^{K} X_k = n) = \frac{n!}{n_1!n_2!...n_K!}\prod_{k=1}^{K}\left(\frac{\lambda_k}{\lambda}\right)^{n_k} \tag{1.64}$$

which is $Mult(n, \boldsymbol{\pi})$, with $\pi_k = \lambda_k/\lambda$, $1 \leq k \leq K$, being the components of $\boldsymbol{\pi}$.

1.4.4 Hypergeometric Distribution

Consider a population of N items, with M of them being of a particular type $\mathcal{A}$ of interest. If a sample of size q is selected from this population, then the number X of type $\mathcal{A}$ items in the sample is modelled by the *hypergeometric distribution*, notated

$$X \sim \mathcal{H}g(N, M, q) \tag{1.65}$$

according to which[10]

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{q-x}}{\binom{N}{q}} \qquad \max(0, q - (N - M)) \leq x \leq \min(q, M) \tag{1.66}$$

1.4.4.1 Mean and Variance

The mean and variance of X, distributed according to the hypergeometic distribution, are

$$\mathrm{E}[X] = \frac{qM}{N} \tag{1.67}$$

$$\mathrm{Var}[X] = \frac{qM(N-q)(N-M)}{N^2(N-1)} \tag{1.68}$$

1.4.5 Chi-Squared Distribution

The central *chi-squared distribution* with n degrees of freedom is defined as the sum of squares of n independent standard normal variables $Z_i \sim \mathcal{N}(0, 1), i = 1, ..., n$. It is denoted by

$$X^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n) \tag{1.69}$$

[10]Note that this is a counting problem. The denominator is the number of ways to sample q items from the total N, and the numerator is the number of ways in which to sample x of the M type $\mathcal{A}$ objects, and $q - x$ of the $N - M$ non-type $\mathcal{A}$ objects.

1.4.5.1 Mean and Variance

If X^2 has the distribution $\chi^2(n)$, then[11]

$$\mathrm{E}[X^2] = n \tag{1.70}$$
$$\mathrm{Var}[X^2] = 2n \tag{1.71}$$

1.4.5.2 Sum of Chi-Squareds

If $X_1^2, ..., X_m^2$ are m independent variables with chi-squared distributions $X_i^2 \sim \chi^2(n_i)$, then[12]

$$\sum_{i=1}^{m} X_i^2 = \chi^2(\sum_{i=1}^{m} n_i) \tag{1.72}$$

1.4.6 Central Limit Theorem

The *Central Limit Theorem (CLT)* tells us that the average $\bar{X}$ of a sum of n i.i.d. variables with mean μ and variance σ^2 will be distributed as:

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \tag{1.73}$$

as $n \to \infty$.

In higher dimensions, the CLT states that the average $\bar{\boldsymbol{X}}$ of n i.i.d. variables with mean vector $\boldsymbol{\mu}$ and variance matrix Σ will be such that:

$$\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{1.74}$$

as $n \to \infty$.

1.4.7 Lagrange Multipliers

We here review the essence of optimisation using *Lagrange multipliers*. For further detail, see the relevant notes from AMVII.

- We want to optimise a function $f(\boldsymbol{x}) \in \mathbb{R}$, with $\boldsymbol{x} \in \mathbb{R}^t$, subject to the vector constraint $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$, with $\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^s$ being s component constraints.
- We construct a function, called the *Lagrange function*, given by

  $$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \lambda_1 g_1(\boldsymbol{x}) + ... + \lambda_s g_s(\boldsymbol{x}) \tag{1.75}$$

  where $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), ..., g_s(\boldsymbol{x}))$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_s)^T$ represents a vector of *Lagrange multipliers*.

Note that the Lagrange function can also be constructed by substracting the constraint functions, that is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - (\lambda_1 g_1(\boldsymbol{x}) + ... + \lambda_s g_s(\boldsymbol{x})) \tag{1.76}$$

It doesn't make any difference to solving the problem, as the solutions for the local optima of the parameters of interest $\boldsymbol{x}$ will be the same[13].

[11] Q1-3c involves proving that the results of this section hold.
[12] Q1-3b involves showing that the result of this section holds.
[13] Indeed, I use the subtraction convention later on.

- To find the points of local optimisation of $f(\boldsymbol{x})$ subject to the equality constraints, we find the stationary points of Lagrange function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$, that is, we solve the following system of equations:

$$\frac{\partial \mathcal{L}}{\partial x_i} = 0 \qquad i = 1, ..., t \tag{1.77}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} = 0 \qquad j = 1, ..., s \tag{1.78}$$

1.5 Acronyms

A reference to some of the acronyms used during the course:

- AIC - Aikaike Information Criterion
- ANOVA - Analysis of Variance
- BIC - Bayesian Information Criterion
- BPD - Bronchopulmonary dysplasia (specific to `bpd` example dataset)
- CDF - Cumulative Distribution Function
- CI - Confidence Interval
- CLT - Central Limit Theorem
- DSSC - Data Science and Statistical Computing
- EDF - Exponential Dispersion Family
- EF - Exponential Family
- GLM - Generalised Linear Model
- GLR - Generalised Likelihood Ratio
- HLLM - Hierarchical Log Linear Model
- i.i.d. - independent and identically distributed
- IWLS - Iterative Weighted Least Squares
- LLM - Log Linear Model
- LR - Likelihood Ratio
- ML - Maximum Likelihood
- MLE - Maximum Likelihood Estimator/Estimation
- PDF - Probability Density Function
- RSS - Residual Sum of Squares
- SSE - Sum of Squares of the Error (equivalent to RSS)
- SSR - Sum of Squares of the Regression
- SST - Sum of Squares Total

1.6 Key References

You are encouraged to read about widely, expanding your knowledge beyond the notes contained here.

Useful references for the first half of this course are the following:

Kateri [2014] - *Contingency Table Analysis - Methods and Implementation Using R*, by Maria Kateri.

Tutz [2012] - *Regression for Categorical Data*, by Gerhard Tutz.

Agresti [2019] - *An Introduction to Categorical Data Analysis*, - by Alan Agresti.

Bilder and Loughin [2015] - *Analysis of Categorical Data with R*, - by Christopher Bilder and Thomas Loughin.

Faraway [2016] - *Extending The Linear Model with R - Generalized Linear, Mixed Effects and Nonparametric Regression Models*, by Julian Faraway.

You can explore the references contained within these books (and these lecture notes) for further information, and take yourselves on a journey of exploration.

Part I

Categorical Data Analysis

Chapter 2

Two-Way Contingency Tables

Contingency tables are data arrays containing counts of observations that are recorded in cross-classifications by a number of discrete factor predictors. They can therefore be seen as summaries of response data where the predictor variables are *discrete factors* or *categorical variables*, and where the responses are *counts*.

2.1 2×2 Tables

In this section we introduce some useful ideas for the more general $I \times J$ tables discussed in Section 2.2 onwards.

- 2×2 tables are very common in biomedical and social sciences. They are particularly useful for binary variables, for example
  - success of a treatment,
  - presence of a characteristic or prognostic factor.

Let X and Y denote two categorical binary response variables. A 2×2 table[1] may be given as shown in Figure 2.1, where

- n_{ij} denotes the observed cell frequency in cell (i, j).
- $n_{i+} = n_{i1} + n_{i2}$ is the marginal frequency for row $i = 1, 2$.
- $n_{+j} = n_{1j} + n_{2j}$ is the marginal frequency for column $j = 1, 2$.
- $n_{++} = n_{1+} + n_{2+} = n_{+1} + n_{+2}$ is the total number of observations in the dataset[2].

2.1.1 Example

Table 2.1 shows a sample of $n_{++} = 3213$ collected in the period 1980-1983 in the St. Louis Epidemiological Catchment Area Survey and cross-classified according to regular smoking habit (rows) and major depressive disorder (columns) (Covey et al. [1990]).

Question

[1] 2×2 tables were studied in Statistical Inference II, but we review them here as they are a good place to start!

[2] In general, a + in place of an index denotes summation over this index.

	$Y=1$	$Y=2$	$Y=.$
$X=1$	n_{11}	n_{12}	n_{1+}
$X=2$	n_{21}	n_{22}	n_{2+}
$X=.$	n_{+1}	n_{+2}	n_{++}

Figure 2.1: Generic 2 x 2 contingency table of counts.

Table 2.1: Survey respondents cross-classified by smoking habit and major depressive disorder.

	Depression: Yes	**No**	**Sum**
Ever Smoked: Yes	144	1729	**1873**
Ever Smoked: No	50	1290	**1340**
Sum	**194**	**3019**	**3213**

- *Is there a relation between cigarette smoking and major depressive disorder?*

2.1.2 Sampling Schemes

Consider the generic 2×2 contingency table in Figure 2.1.
Before being observed, the counts N_{ij} are variables to be sampled from a particular distribution, with n_{ij} then forming a realisation from the sampling scheme.

Important: It is important to understand the implication of different sampling distributions. Each one represents a different sampling scheme or data collection mechanism, these being determined by the experiment being performed (which is likely being driven by a hypothesis to be tested).

Here we provide a brief illustration of what we mean - these will be considered in greater detail in Section 2.3.

2.1.2.1 Poisson Sampling Scheme

Poisson sampling describes the scenario where the dataset (n_{ij}) is collected (sampled) independently in a given spatial, temporal, or other interval.

Then, n_{ij} is a realisation of a Poisson random variable

$$N_{ij} \sim Poi(\lambda_{ij}) \tag{2.1}$$

so that

$$P(N_{ij} = n_{ij}) = \frac{\exp(-\lambda_{ij})\lambda_{ij}^{n_{ij}}}{n_{ij}!} \tag{2.2}$$

where λ_{ij} are the rate parameters.

Table 2.2: Hypothetical data of a crossover trial comparing low and high dose treatments on a sample of 100 patients.

	Low Dose: Success	**Failure**	**Sum**
High Dose: Success	62	18	**80**
High Dose: Failure	8	12	**20**
Sum	**70**	**30**	**100**

2.1.2.1.1 Example Consider the survey data in Table 2.1. We can see that this survey follows a Poisson sampling scheme if it involved questioning people randomly over a fixed timeframe. We do not how many people will be questioned in total beforehand (perhaps as many as possible), and so each compartment now represents an independent Poisson count result (in the sense that we could have counted any of those compartments separately or only and got the same result).

2.1.2.2 Multinomial Sampling Scheme

What if n_{++} is fixed (that is, predetermined)?

Well, then we have that

$$(N_{11}, N_{12}, N_{21}, N_{22}) \sim Mult(n_{++}, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})) \tag{2.3}$$

so that

$$P((N_{11}, N_{12}, N_{21}, N_{22}) = (n_{11}, n_{12}, n_{21}, n_{22}) | N_{++} = n_{++}) = \frac{n_{++}!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{2.4}$$

where $\pi_{ij} = P(X = i, Y = j)$ is the theoretical probability of sampling someone in compartment (i, j) from the represented population. The probability vector $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ is the joint distribution of X and Y.

2.1.2.2.1 Example The data in Table 2.2 represents a hypothetical crossover trial. One fixed sample of patients of size $n_{++} = 100$ is considered and they receive a low dose and then a high dose of a particular treatment, one week apart. A pair of responses is available for each patient and Table 2.2 cross-classifies these responses, reporting the number of treatments for which both treatments were successful, both failed, or one succeeded and the other failed. Notice that each patient has to belong to one of the four compartments, and given that there is a fixed sample of 100 patients, means that the setup of this experiment has resulted in a multinomial sampling procedure.

2.1.2.3 Product Binomial Sampling Scheme

We now assume the marginal row sizes n_{1+}, n_{2+} are fixed. In this case

$$N_{11} \sim Bin(n_{1+}, \pi_{11}) \tag{2.5}$$
$$N_{21} \sim Bin(n_{2+}, \pi_{21}) \tag{2.6}$$

Table 2.3: Hypothetical data from two independent samples of low and high dose treatments.

	Response: Success	**Failure**	**Sum**
Dose: High	41	9	**50**
Dose: Low	37	13	**50**
Sum	**78**	**22**	**100**

so that

$$P(N_{11} = n_{11}|N_{1+} = n_{1+}) = \binom{n_{1+}}{n_{11}} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \tag{2.7}$$

$$P(N_{21} = n_{21}|N_{2+} = n_{2+}) = \binom{n_{2+}}{n_{21}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}} \tag{2.8}$$

and

$$\pi_{12} = 1 - \pi_{11} \tag{2.9}$$

$$\pi_{22} = 1 - \pi_{21} \tag{2.10}$$

$$N_{12} = n_{1+} - N_{11} \tag{2.11}$$

$$N_{22} = n_{2+} - N_{12} \tag{2.12}$$

2.1.2.3.1 Example The data in Table 2.3 are similar to that of Table 2.2, however the experiment is designed differently. Instead of a crossover trial, where each patient is given both treatments, the sample of 100 patients is now divided equally into subsamples of size 50. The patients in each subsample either receive a high or low dose of treatment, and the result is recorded. Note that the row sums are fixed in this case, and so this would correspond to a product multinomial sampling scheme.

2.1.3 Odds and Odds Ratio

2.1.3.1 Odds

Given a generic success probability π_A of event A (a binary response), we are often interested in the *odds* of success of an event A, given by

$$\omega_A = \frac{\pi_A}{1 - \pi_A} \tag{2.13}$$

Interpretation:

$$\begin{aligned}
\omega_A = 1 &\implies \pi_A = 1 - \pi_A &&\implies \text{success is equally likely as failure} \\
\omega_A > 1 &\implies \pi_A > 1 - \pi_A &&\implies \text{success is more likely than failure} \\
\omega_A < 1 &\implies \pi_A < 1 - \pi_A &&\implies \text{success is less likely than failure}
\end{aligned}$$

Question: What does an odds of 2 mean relative to success probability π_A?

2.1.3.2 Odds Ratio

If π_A and π_B are the success probabilities for two events A and B, then their *odds ratio* is defined as

$$r_{AB} = \frac{\omega_A}{\omega_B} = \frac{\pi_A/(1-\pi_A)}{\pi_B/(1-\pi_B)} \tag{2.14}$$

r_{AB} can be interpreted as the (multiplicative) difference in the odds (relative chance) of success to failure between event A and event B. Therefore:

$$\begin{aligned}
r_{AB} = 1 &\implies \omega_A = \omega_B \\
&\implies \text{the relative chance of success to failure of } A \text{ is equal to that of } B \\
r_{AB} > 1 &\implies \omega_A > \omega_B \\
&\implies \text{the relative chance of success to failure of } A \text{ is greater than that of } B \\
r_{AB} < 1 &\implies \omega_A < \omega_B \\
&\implies \text{the relative chance of success to failure of } A \text{ is less than that of } B
\end{aligned}$$

Odds ratios are typically more informative for the comparison of π_A and π_B than looking at their *difference* $\pi_A - \pi_B$ or their *relative risk* π_A/π_B.

2.1.3.3 Odds Ratio for a 2 × 2 Contingency Table

In terms of the joint distribution of a 2×2 contingency table, the odds ratio r_{12} is given by[3]

$$r_{12} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{2.15}$$

and is a (multiplicative) measure of

- the difference in the odds (relative chance) of $Y = 1$ to $Y = 2$ between $X = 1$ and $X = 2$; equivalently
- the difference in the odds (relative chance) of $X = 1$ to $X = 2$ between $Y = 1$ and $Y = 2$.

Therefore:

$$\begin{aligned}
r_{12} = 1 &\implies \pi_{11}/\pi_{12} = \pi_{21}/\pi_{22} \\
&\implies \text{the relative chance of } Y = 1 \text{ to } Y = 2 \text{ is the same given } X = 1 \text{ or } X = 2 \\
&\implies \pi_{11}/\pi_{21} = \pi_{12}/\pi_{22} \\
&\implies \text{the relative chance of } X = 1 \text{ to } X = 2 \text{ is the same given } Y = 1 \text{ or } Y = 2
\end{aligned}$$

$$\begin{aligned}
r_{12} > 1 &\implies \pi_{11}/\pi_{12} > \pi_{21}/\pi_{22} \\
&\implies \text{the relative chance of } Y = 1 \text{ to } Y = 2 \text{ given } X = 1 \text{ is greater than given } X = 2 \\
&\implies \pi_{11}/\pi_{21} > \pi_{12}/\pi_{22} \\
&\implies \text{the relative chance of } X = 1 \text{ to } X = 2 \text{ given } Y = 1 \text{ is greater than given } Y = 2
\end{aligned}$$

[3]Note that the notation r^{12} used here is slightly different to the convention for nominal odds ratios for $I \times J$ tables used later on, where more care is required to clearly distinguish what categories the odds ratio is between.

$$
\begin{aligned}
r_{12} < 1 &\implies \pi_{11}/\pi_{12} < \pi_{21}/\pi_{22} \\
&\implies \text{the relative chance of } Y = 1 \text{ to } Y = 2 \text{ given } X = 1 \text{ is less than given } X = 2 \\
&\implies \pi_{11}/\pi_{21} < \pi_{12}/\pi_{22} \\
&\implies \text{the relative chance of } X = 1 \text{ to } X = 2 \text{ given } Y = 1 \text{ is less than given } Y = 2
\end{aligned}
$$

2.1.3.4 Sample Odds Ratio

The *sample odds ratio* is

$$\hat{r}_{12} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{2.16}$$

where $p_{ij} = n_{ij}/n_{++}$ is the sample proportion in class (i, j).

2.1.3.5 Example

Consider the data in Table 2.3.

- The sample proportions of success for high and low dose of treatment are $p_H = \frac{41}{50}$, and $p_L = \frac{37}{50}$ respectively.
- The overall proportions are $p_{11} = \frac{41}{100}$, and $p_{21} = \frac{37}{100}$.
- The sample odds of success assuming high dose is

$$\frac{p_H}{1 - p_H} = \frac{41/50}{9/50} = \frac{41}{9} \approx 4.556 \tag{2.17}$$

- Incidentally (and unsurprisingly), using the overall proportion values:

$$\frac{p_{11}}{p_{12}} = \frac{41/100}{9/100} = \frac{41}{9} \tag{2.18}$$

  yields the same value.

- The sample odds of success assuming low dose is

$$\frac{p_L}{1 - p_L} = \frac{37}{13} \approx 2.846 \tag{2.19}$$

- The sample odds ratio is given by

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{41 \times 13}{9 \times 37} = \frac{533}{333} \approx 1.601 \tag{2.20}$$

2.2 $I \times J$ Tables

- Let X and Y denote two categorical response variables (also called classifiers).
- X has I categories and Y has J categories.
- Classifications of subjects on both variables have IJ possible combinations.

2.2.1 Table of Counts

The $I \times J$ contingency table[4] of observed frequencies n_{ij} for X and Y is the rectangular table

- which has I rows for categories of X and J columns for categories of Y,
- whose cells represent the IJ possible outcomes of the responses (X, Y),
- which displays n_{ij}, the observed number of outcomes (i, j) in each case.

Such a table is shown as given in Figure 2.2.

	$Y=1$	$Y=2$	$\cdots$	$Y=j$	$\cdots$	$Y=J$	$Y=.$
$X=1$	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1+}
$X=2$	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=i$	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=I$	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I+}
$X=.$	n_{+1}	n_{+2}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n_{++}

Figure 2.2: Generic I x J contingency table of X and Y displaying the observed number of outcomes $(X = i, Y = j)$.

Similar to 2×2 tables, we have that

- n_{ij} denotes the observed number (or counts, frequency) of outcomes (i, j)[5].
- $n_{i+} = \sum_j n_{ij}$ is the marginal frequency of outcomes $X = i$ regardless of the outcome of Y.
- $n_{+j} = \sum_i n_{ij}$ is the marginal frequency of outcomes $Y = j$ regardless of the outcome of X.
- $n_{++} = \sum_{i,j} n_{ij}$ is the observed total number of observations in the dataset.

2.2.2 Table of Theoretical Probabilities

Similar to the 2×2 case, we define

- $\pi_{ij} = P(X = i, Y = j)$ to be the theoretical probability of getting outcome (i, j).

We also define

- $\pi_{i+} = \sum_j \pi_{ij}$ to be the marginal probability of getting an outcome $X = i$ regardless of the outcome of $Y = j$.
- $\pi_{+j} = \sum_i \pi_{ij}$ to be the marginal probability of getting an outcome $Y = j$ regardless of the outcome of $X = i$.
- $\pi_{++} = \sum_{i,j} \pi_{ij} = 1$

One can tabulate the distribution $\pi_{ij} = P(X = i, Y = j)$ of responses (X, Y) as in the classification given in Figure 2.3.

[4]Also called n_{ij} contingency table, 2-way contingency table, or classification table.
[5](i, j) is short for $(X = i, Y = j)$.

	$Y=1$	$Y=2$	$\cdots$	$Y=j$	$\cdots$	$Y=J$	$Y=.$
$X=1$	π_{11}	π_{12}	$\cdots$	π_{1j}	$\cdots$	π_{1J}	π_{1+}
$X=2$	π_{21}	π_{22}	$\cdots$	π_{2j}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=i$	π_{i1}	π_{i2}	$\cdots$	π_{ij}	$\cdots$	π_{iJ}	π_{i+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=I$	π_{I1}	π_{I2}	$\cdots$	π_{Ij}	$\cdots$	π_{IJ}	π_{I+}
$X=.$	π_{+1}	π_{+2}	$\cdots$	π_{+j}	$\cdots$	π_{+J}	$\pi_{++}=1$

Figure 2.3: Generic IJ contingency table of the theoretical probability distribution of responses.

2.2.3 Additional Tables

In a similar manner, a contingency table can display the following quantities

- $p_{ij} = \frac{n_{ij}}{n_{++}}$ - the observed proportion of outcomes (i, j).
- μ_{ij} - the expected number of outcomes (i, j).
- N_{ij} - the variable number of outcomes (i, j) (before the observations are collected).

2.2.4 Conditional Probabilities

Contingency tables can also display the following conditional quantities:

- $\pi_{i|j} = \frac{\pi_{ij}}{\pi_{+j}} = P(X = i|Y = j)$ - the conditional probability of getting an outcome $X = i$ given that the outcome $Y = j$.
- $\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = P(Y = j|X = i)$ - the conditional probability of getting an outcome $Y = j$ given that the outcome $X = i$.
- $p_{i|j} = \frac{p_{ij}}{p_{+j}} = \frac{n_{ij}}{n_{+j}}$ - the observed proportion of the outcomes with $Y = j$ for which $X = i$.
- $p_{j|i} = \frac{p_{ij}}{p_{i+}} = \frac{n_{ij}}{n_{i+}}$ - the observed proportion of the outcomes with $X = i$ for which $Y = j$.

Two schematics of tables representing the conditional probabilities $\pi_{j|i}$, and conditional proportions $p_{j|i}$ of the outcomes $Y = j$, given that $X = i$ (that is, conditioning on the rows) are given by the Tables in Figures 2.4 and 2.5.

	$Y=1$	$Y=2$	$\cdots$	$Y=j$	$\cdots$	$Y=J$	$Y=.$
$X=1$	$\pi_{1\|1}$	$\pi_{2\|1}$	$\cdots$	$\pi_{j\|1}$	$\cdots$	$\pi_{J\|1}$	$\pi_{+\|1}=1$
$X=2$	$\pi_{1\|2}$	$\pi_{2\|2}$	$\cdots$	$\pi_{j\|2}$	$\cdots$	$\pi_{J\|2}$	$\pi_{+\|2}=1$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=i$	$\pi_{1\|i}$	$\pi_{2\|i}$	$\cdots$	$\pi_{j\|i}=\frac{\pi_{ij}}{\pi_{i+}}$	$\cdots$	$\pi_{J\|i}$	$\pi_{+\|i}=1$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=I$	$\pi_{1\|I}$	$\pi_{2\|I}$	$\cdots$	$\pi_{j\|I}$	$\cdots$	$\pi_{J\|I}$	$\pi_{+\|I}=1$

Figure 2.4: A generic IJ table displaying the conditional probabilities of the outcome Y given outcome X.

	$Y=1$	$Y=2$	$\cdots$	$Y=j$	$\cdots$	$Y=J$	$Y=.$
$X=1$	$p_{1\|1}$	$p_{2\|1}$	$\cdots$	$p_{j\|1}$	$\cdots$	$p_{J\|1}$	$p_{+\|1}=1$
$X=2$	$p_{1\|2}$	$p_{2\|2}$	$\cdots$	$p_{j\|2}$	$\cdots$	$p_{J\|2}$	$p_{+\|2}=1$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=i$	$p_{1\|i}$	$p_{2\|i}$	$\cdots$	$p_{j\|i}=\frac{p_{ij}}{p_{i+}}$	$\cdots$	$p_{J\|i}$	$p_{+\|i}=1$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=I$	$p_{1\|I}$	$p_{2\|I}$	$\cdots$	$p_{j\|I}$	$\cdots$	$p_{J\|I}$	$p_{+\|I}=1$

Figure 2.5: A generic IJ table displaying the conditional proportions of the outcome Y given outcome X.

2.3 Sampling Schemes

We here go into a bit more detail on some of the sampling schemes introduced in Section 2.1.2 for 2×2 tables.

Remember that it is important to understand the implication of different sampling distributions. Each one represents a different sampling scheme or data collection mechanism, these being determined by the experiment being performed (which is likely being driven by a hypothesis to be tested).

2.3.1 Poisson Sampling Scheme

As discussed in Section 2.1.2.1, Poisson sampling describes the scenario where the dataset (n_{ij}) is collected (sampled) independently in a given spatial, temporal, or other interval.

Then, n_{ij} is a realisation of a Poisson random variable

$$N_{ij} \sim Poi(\lambda_{ij}) \tag{2.21}$$

so that

$$P(N_{ij} = n_{ij}) = \frac{\exp(-\lambda_{ij})\lambda_{ij}^{n_{ij}}}{n_{ij}!} \tag{2.22}$$

where λ_{ij} are the rate parameters.

2.3.1.1 Mean and Variance

We have that

$$\mathrm{E}[N_{ij}] = \lambda_{ij} \tag{2.23}$$
$$\mathrm{Var}[N_{ij}] = \lambda_{ij} \tag{2.24}$$

for $i = 1, ..., I$ and $j = 1, ..., J$.

2.3.1.2 Likelihood

The likelihood is

$$L(\boldsymbol{\lambda}) = \prod_{i,j} \frac{\exp(-\lambda_{ij})\lambda_{ij}^{n_{ij}}}{n_{ij}!} \propto \exp(-\sum_{i,j} \lambda_{ij}) \prod_{i,j} \lambda_{ij}^{n_{ij}} \tag{2.25}$$

where $\boldsymbol{\lambda} = (\lambda_{11}, \lambda_{12}, ..., \lambda_{IJ})^T$.

2.3.2 Multinomial Sampling Scheme

The Multinomial sampling scheme describes the scenario where the dataset (n_{ij}) is collected (sampled) independently given that the total sample size n_{++} is fixed/predetermined. Then $\boldsymbol{n} = (n_{11}, n_{12}, ..., n_{IJ})^T$ is a realisation of a Multinomial random variable[6]

$$\boldsymbol{N} \sim Mult(n_{++}, \boldsymbol{\pi}) \tag{2.26}$$

where $\boldsymbol{N} = (N_{11}, N_{12}, ..., N_{IJ})^T$ and $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, ..., \pi_{IJ})^T$, so that

$$P(\boldsymbol{N} = \boldsymbol{n} | N_{++} = n_{++}) = \frac{n_{++}!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{2.27}$$

2.3.2.1 Mean and Variance

We have that

$$\mathrm{E}[\boldsymbol{N}] = n_{++}\boldsymbol{\pi} \tag{2.28}$$

$$\mathrm{Var}[\boldsymbol{N}] = n_{++}(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T) \tag{2.29}$$

for $i = 1, ..., I$ and $j = 1, ..., J$.

2.3.2.2 Likelihood

The likelihood is

$$L(\boldsymbol{\pi}) = P(\boldsymbol{N} = \boldsymbol{n} | N_{++} = n_{++}) = \frac{n_{++}!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}} \propto \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{2.30}$$

as n_{++} is predetermined and fixed. Notice that since $\mu_{ij} = n_{++}\pi_{ij}$, we have that

$$L(\boldsymbol{\pi}) = \prod_{i,j} \left(\frac{\mu_{ij}}{n_{++}} \right)^{n_{ij}} \propto \prod_{i,j} \mu_{ij}^{n_{ij}} \tag{2.31}$$

2.3.3 Product Multinomial Sampling Scheme

The Product Multinomial sampling scheme describes the scenario where the sample (n_{ij}) is collected randomly, given that the marginal row sizes n_{i+} are fixed/predetermined for $i = 1, ..., I$. Then $(n_{i1}, ..., n_{iJ})$ is a realisation of the Multinomial random variable

$$(N_{i1}, ..., n_{iJ}) \sim Mult(n_{i+}, \boldsymbol{\pi}_i^*) \tag{2.32}$$

where $\boldsymbol{\pi}_i^\star = (\pi_{1|i}, ..., \pi_{J|i})$ for $i = 1, ..., I$, so that

$$P(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N}_{i+} = \boldsymbol{n}_{i+}) = \prod_i \left[\frac{n_{i+}!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}} \right] \tag{2.33}$$

where $\boldsymbol{N}_{i+} = (N_{1+}, ..., N_{I+})$.

[6] For the avoidance of confusion, note that bold $\boldsymbol{n}$ here represents the vector of counts in each compartment (i, j), and n_{++} denotes total sample size.

2.3.3.1 Likelihood

The likelihood is

$$L(\boldsymbol{\pi}) = \prod_i \left[\frac{n_{i+}!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}} \right] \propto \prod_i \prod_j \pi_{j|i}^{n_{ij}} = \prod_{i,j} \left(\frac{\pi_{ij}}{\pi_{i+}} \right)^{n_{ij}} \propto \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{2.34}$$

since π_{i+} is determined and fixed by the fact that n_{i+} is determined and fixed. Notice that since $\mu_{ij} = n_{++}\pi_{ij}$, we have that

$$L(\boldsymbol{\pi}) \propto \prod_{i,j} \left(\frac{\mu_{ij}}{n_{++}} \right)^{n_{ij}} \propto \prod_{i,j} \mu_{ij}^{n_{ij}} \tag{2.35}$$

2.3.4 Inferential Equivalence

Proposition: The three sampling schemes in Sections 2.3.1, 2.3.2 and 2.3.3 lead to the same MLE.

Proof: By Equations (2.31) and (2.35), it is clear for the Multinomial sampling scheme and the product Multinomial sampling scheme.

Now, consider Equation (2.25). Upon observing the sample, we can condition on the total sample size n_{++}. Then, using the result of Section 1.4.3.4, we have that

$$L(\boldsymbol{\lambda}|n_{++}) = \frac{n_{++}!}{\prod_{i,j}(n_{ij}!)} \prod_{i,j} \left(\frac{\lambda_{ij}}{\lambda} \right)^{n_{ij}} \propto \prod_{i,j} \pi_{ij}^{n_{ij}} \propto \prod_{i,j} \mu_{ij}^{n_{ij}} \tag{2.36}$$

Although we have inferential equivalence in terms of MLEs, the different sampling schemes are still important as they are guided by, and inform us about, the data collection mechanism and experiment being performed (and perhaps any hypothesis being tested).

2.3.5 Fixed Row and Column Sums

A final scenario we have not yet considered is that where sample (n_{ij}) is collected randomly given that the marginal counts n_{i+} and n_{+j} are fixed and predetermined before the experiment was performed. It is actually quite unusual to be in such a situation, and the probability distributions representing possible configurations are more challenging to represent. We will consider specific scenarios only.[7]

2.3.6 What now?

So, we have introduced contingency tables as a means of presenting cross-classification data of two categorical variables. We have explained that different experimental scenarios naturally lead to different sampling schemes.

Now what?

- We may wish to test the homogeneity of Y given X. For example, in Table 2.3, is the distribution of *Response* (Y) the same regardless of *dose level* (X) (Section 2.4.3)?

[7] One such scenario is considered in Q2-12.

- We may wish to test independence (correlation, association) of X and Y. For example, in Table 2.1, is there a correlation between cigarette smoking and major depressive disorder, or are they uncorrelated (Section 2.4.4)?
- If there is an association between two variables, we may wish to understand more about the strength of association, and what elements of the table are driving this conclusion.
- We may wish to understand more about how we can draw conclusions from the data using Odds Ratios (less clear for $I \times J$ tables than 2×2 tables).
- Maybe we need ways of presenting our results to non-experts (visualisation - see Practical 1).
- Maybe we want to do all of the above considering more than two categorical variables (Section 3).

2.4 Chi-Square Test

The *chi-square goodness-of-fit* test can be used to determine whether a given sample is consistent with a (specified) hypothesised distribution. Thus, the term goodness-of-fit implies how well the distributional assumption fits the data. This test was initially developed by Karl Pearson in 1900 for categorical data. We will utilise this test in the context of contingency tables to test for homogeneity and independence[8].

Suppose data consists of a random sample of m independent observations which are classified into k mutually exclusive and exhaustive categories. The number of observations falling into category j, for $j = 1, ..., k$, is called the observed frequency of that category, and denoted by O_j.

We wish to test a statement about the unknown probability distribution $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)$ of the random variable being observed, where π_j is the theoretical probability of an observation in category j. We denote the specified probabilities under our hypothesised distribution as $\boldsymbol{\pi_0} = (\pi_{0,1}, ..., \pi_{0,k})$ Thus, the null and alternative hypotheses are

$$\mathcal{H}_0 : \pi_j = \pi_{0,j}, \qquad j = 1, ..., k \tag{2.37}$$

$$\mathcal{H}_1 : \pi_j \neq \pi_{0,j}, \qquad \text{for any } j = 1, ..., k \tag{2.38}$$

2.4.1 The Chi-Square Test Statistic

The Chi-Square goodness-of-fit test statistic is

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} \tag{2.39}$$

where O_j are the observed frequencies, $E_j = m\pi_{0,j}$ are the expected frequencies, and $m = \sum_{j=1}^{k} O_j$ is the total sample size.

[8]The ideas of these two tests were covered in Statistical Inference II, but we shall go into more theoretical detail here.

The chi-square test statistic is based on the difference between the expected and observed frequencies. The larger the differences between the expected and observed frequencies, the larger the chi-square test statistic value will be, hence providing evidence against the null hypothesis.

2.4.2 Asymptotic Distribution of X^2

Proposition: For large m, we have that, under $\mathcal{H}_0$:

$$X^2 \sim \chi^2_{k-1} \tag{2.40}$$

where k is the number of categories.

Consequence: We reject $\mathcal{H}_0$ at significance level α if

$$X^2 \geq \chi^{2,\star}_{k-1,\alpha} \tag{2.41}$$

where $\chi^{2,\star}_{k-1,\alpha}$ is the $(1-\alpha)$-quantile of the χ^2_{k-1}-distribution.

Proof: Each observation $i = 1, ..., m$ can be viewed as an observation of random variable $\boldsymbol{X}_i$, these being m independent variables drawn from

$$\boldsymbol{X}_i \sim Mult(1, \boldsymbol{\pi}) \tag{2.42}$$

where $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)$ is the vector of probabilities that any randomly selected observation belongs to each category. Each $\boldsymbol{X}_i$ consists of exactly $k-1$ zeroes and a single one, where the one is in the component of the observed category at trial i.

Under this view, we have that

$$E_j = m\pi_j \tag{2.43}$$
$$O_j = m\bar{X}_j \tag{2.44}$$

where $\bar{\boldsymbol{X}} = \frac{1}{m}\sum_{i=1}^m \boldsymbol{X}_i$, so that

$$X^2 = m\sum_{j=1}^k \frac{(\bar{X}_j - \pi_j)^2}{\pi_j} \tag{2.45}$$

We have that (using the multinomial results of Section 1.4.3.1)

$$\text{Var}[X_{ij}] = \pi_j(1-\pi_j) \tag{2.46}$$
$$\text{Cov}[X_{ij}, X_{il}] = -\pi_j\pi_l, \qquad j \neq l \tag{2.47}$$

and that $\boldsymbol{X}_i$ has covariance matrix

$$\Sigma = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_k \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_k \\ \vdots & & \ddots & \vdots \\ -\pi_k\pi_1 & -\pi_k\pi_2 & \cdots & \pi_k(1-\pi_k) \end{pmatrix} \tag{2.48}$$

Since $\mathrm{E}[\boldsymbol{X}_i] = \boldsymbol{\pi}$, the CLT (Section 1.4.6) implies that

$$\sqrt{m}(\bar{\boldsymbol{X}}_m - \boldsymbol{\pi}) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Sigma) \tag{2.49}$$

Note that the sum of the j^{th} row of Σ is $\pi_j - \pi_j(\pi_1 + \cdots + \pi_k) = 0$, hence the sum of the rows of Σ is the zero vector, and thus Σ is not invertible.

Let $\boldsymbol{Y}_i = (X_{i,1}, ..., X_{i,k-1})$, with covariance matrix $\Sigma^\star$ being the upper-left $(k-1) \times (k-1)$ submatrix of Σ. Similarly, let $\boldsymbol{\pi}^\star = (\pi_1, ..., \pi_{k-1})$.

One may verify that $\Sigma^\star$ is invertible and that[9]

$$(\Sigma^\star)^{-1} = \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_k} & \frac{1}{\pi_k} & \cdots & \frac{1}{\pi_k} \\ \frac{1}{\pi_k} & \frac{1}{\pi_2} + \frac{1}{\pi_k} & \cdots & \frac{1}{\pi_k} \\ \vdots & & \ddots & \vdots \\ \frac{1}{\pi_k} & \frac{1}{\pi_k} & \cdots & \frac{1}{\pi_{k-1}} + \frac{1}{\pi_k} \end{pmatrix} \tag{2.50}$$

Furthermore, X^2 can be rewritten as[10]

$$X^2 = m(\bar{\boldsymbol{Y}} - \boldsymbol{\pi}^\star)^T (\Sigma^\star)^{-1} (\bar{\boldsymbol{Y}} - \boldsymbol{\pi}^\star) \tag{2.51}$$

Now, define

$$\boldsymbol{Z}_m = \sqrt{m}(\Sigma^\star)^{-\frac{1}{2}} (\bar{\boldsymbol{Y}} - \boldsymbol{\pi}^\star) \tag{2.52}$$

Then the central limit theorem implies that

$$\boldsymbol{Z}_m \xrightarrow{d} \mathcal{N}_{k-1}(\mathbf{0}, I_{k-1}) \tag{2.53}$$

so that

$$X^2 = \boldsymbol{Z}_m^T \boldsymbol{Z}_m \xrightarrow{d} \chi^2_{k-1} \tag{2.54}$$

2.4.3 Testing Independence

- Variable X and Y are independent if

$$P(X = i, Y = j) = P(X = i)P(Y = j), \qquad i = 1, ..., I \quad j = 1, ..., J \tag{2.55}$$

  that is

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \qquad i = 1, ..., I \quad j = 1, ..., J \tag{2.56}$$

[9] Q2-5a involves showing that Equation (2.50) holds.
[10] Q2-5b involves showing that Equation (2.51) holds.

2.4.3.1 Hypothesis Test

- If we are interested in testing independence[11] between X and Y, then we wish to test with hypotheses as follows:

$$\mathcal{H}_0 : X \text{ and } Y \text{ are not associated.} \tag{2.57}$$
$$\mathcal{H}_1 : X \text{ and } Y \text{ are associated.} \tag{2.58}$$

  which we practically do by testing the following:

$$\mathcal{H}_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \qquad i = 1, ..., I \quad j = 1, ..., J \tag{2.59}$$
$$\mathcal{H}_1 : \mathcal{H}_0 \text{ is not true.} \tag{2.60}$$

- Notice that the Chi-Square test can be applied to the problem of testing $\mathcal{H}_0$ above, where the k mutually exclusive categories are taken to be the IJ cross-classified possible pairings of X and Y.
- Under $\mathcal{H}_0$, we have that $E_{ij} = n_{++}\pi_{ij} = n_{++}\pi_{i+}\pi_{+j}$.
- However, we do not know π_{ij} or π_{i+}, π_{+j}, but we can estimate them using ML.

2.4.3.2 Maximum Likelihood

Assume that the sampling distribution has been specified according to the sampling scheme used. Under an assumed relation between X and Y, ML estimates of $\{\pi_{ij}\}$ can be obtained using the method of Lagrange multipliers (see Section 1.4.7).

2.4.3.2.1 ML - no Independence Consider a Multinomial sampling scheme. We do not assume any independence for X and Y . We need to find the MLE of $\boldsymbol{\pi}$.

- The log likelihood (from Equation (2.30)) is

$$l(\boldsymbol{\pi}) \propto \sum_{i,j} n_{ij} \log(\pi_{ij}) \tag{2.61}$$

- We wish to maximise $l(\boldsymbol{\pi})$ subject to $g(\boldsymbol{\pi}) = \sum_{i,j} \pi_{ij} = 1$.
  Therefore the Lagrange function is

$$\mathcal{L}(\boldsymbol{\pi}, \lambda) = l(\boldsymbol{\pi}) - \lambda\Big(\sum_{i,j} \pi_{ij} - 1\Big) \tag{2.62}$$
$$= \sum_{i,j} n_{ij} \log(\pi_{ij}) - \lambda\Big(\sum_{i,j} \pi_{ij} - 1\Big) \tag{2.63}$$

[11] It is important to be weary of the word *independent* here; if we reject $\mathcal{H}_0$, even correctly, then this evidence of *dependence* (as the opposite of independence) is really only evidence of *association*. One variable may provide information that is relevant to/useful for predicting the other (unknown) variable, but they may not be dependent in the sense of directly depending on each other in the real-world. This is simply the *correlation is not the same as causation* argument. In a sense we could call this test a *test for association*, however, the word *independence* is so well-rooted in the textbooks that I will go along with this misnomer, but encourage you to make sure that you understand that we are really testing for association, despite the name.

- Local optima $\hat{\boldsymbol{\pi}}, \hat{\lambda}$ will satisfy:

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\lambda})}{\partial \pi_{ij}} = 0 \qquad i = 1, ..., I, \quad j = 1, ..., J \tag{2.64}$$

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\lambda})}{\partial \lambda} = 0 \tag{2.65}$$

which implies

$$\frac{n_{ij}}{\hat{\pi}_{ij}} - \hat{\lambda} = 0 \qquad i = 1, ..., I, \quad j = 1, ..., J \tag{2.66}$$

$$\sum_{i,j} \hat{\pi}_{ij} - 1 = 0 \tag{2.67}$$

and hence

$$n_{ij} = \hat{\lambda}\hat{\pi}_{ij} \tag{2.68}$$

$$\implies \sum_{i,j} n_{ij} = \hat{\lambda} \sum_{i,j} \hat{\pi}_{ij} \tag{2.69}$$

$$\implies \hat{\lambda} = n_{++} \tag{2.70}$$

and thus

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_{++}} \tag{2.71}$$

2.4.3.2.2 ML - Assuming Independence Consider a Multinomial sampling scheme, and that X and Y are independent. We need to find the MLE of $\boldsymbol{\pi}$, but now we have that

$$\pi_{ij} = \pi_{i+}\pi_{+j} \tag{2.72}$$

- The log likelihood is

$$l(\boldsymbol{\pi}) \propto \sum_{i,j} n_{ij} \log(\pi_{ij}) \tag{2.73}$$

$$= \sum_{i} n_{i+} \log(\pi_{i+}) + \sum_{j} n_{+j} \log(\pi_{+j}) \tag{2.74}$$

- We can use the method of Lagrange multipliers to show that[12]

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n_{++}} \tag{2.75}$$

$$\hat{\pi}_{+j} = \frac{n_{+j}}{n_{++}} \tag{2.76}$$

2.4.3.3 Chi-Square Test of Independence

The Chi-Square test statistic[13]

$$X^2 = \sum_{i,j=1,1}^{I,J} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{2.77}$$

[12] Q2-6a involves showing this.

[13] Compare this expression with that given by Equation (2.39).

Table 2.4: Cross-classification of a dataset of mushroom samples based on edibility and shape of cap.

	Cap shape: bell	**conical**	**flat**	**knobbed**	**convex**	**Sum**
Edibility: Edible	101	0	399	57	487	**1044**
Edibility: Poisonous	12	1	389	150	427	**979**
Sum	**113**	**1**	**788**	**207**	**914**	**2023**

follows a $\chi^2_{(I-1)(J-1)}$-distribution, where

$$\hat{E}_{ij} = n_{++}\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{++}\frac{n_{i+}n_{+j}}{n_{++}n_{++}} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{2.78}$$

We therefore reject $\mathcal{H}_0$ at significance level α if

$$X^2 \geq \chi^{2,\star}_{(I-1)(J-1),\alpha} \tag{2.79}$$

where $\chi^{2,\star}_{(I-1)(J-1),\alpha}$ is the $1-\alpha$-quantile of the $\chi^2_{(I-1)(J-1)}$-distribution.

2.4.3.3.1 Degrees of Freedom Where do the $(I-1)(J-1)$ degrees of freedom come from? Well, we have a total of IJ cells in the table. If no parameters are under estimation then the degrees of freedom are

$$\text{df} = IJ - 1 \tag{2.80}$$

because the cell probabilities sum to 1, so one cell probability (parameter) is restricted. However, in the case where parameters are under estimation and given the additional constraints

$$\sum_{i=1}^{I}\pi_{i+} = 1 \qquad \text{and} \qquad \sum_{j=1}^{J}\pi_{+j} = 1 \tag{2.81}$$

the total number of unknown parameters to be estimated, under $\mathcal{H}_0$, is

$$s = (I-1) + (J-1) = I + J - 2 \tag{2.82}$$

So, the overall degrees of freedom that remain are

$$\text{df} = IJ - (I+J-2) - 1 = (I-1)(J-1) \tag{2.83}$$

2.4.3.4 Example

Mushroom hunting (otherwise known as *shrooming*) is enjoying new peaks in popularity. You do have to be careful though; whilst some mushrooms are edible delicacies, others are likely to spell certain death, hence it is important to know which ones to avoid!

We are interested in identifying whether there is an association between cap shape and edibility of the mushroom. The dataset of 2023 mushroom samples were cross-classified according to cap shape and edibility, as shown in Contingency Table 2.4.

Table 2.5: Estimated E_{ij} values.

	Cap shape: bell	**conical**	**flat**	**knobbed**	**convex**	**Sum**
Edibility: Edible	58.31537	0.5160652	406.6594	106.8255	471.6836	**1044**
Edibility: Poisonous	54.68463	0.4839348	381.3406	100.1745	442.3164	**979**
Sum	**113.00000**	**1.0000000**	**788.0000**	**207.0000**	**914.0000**	**2023**

Table 2.6: X_{ij}^2 values.

	Cap shape: bell	**conical**	**flat**	**knobbed**	**convex**
Edibility: Edible	31.24352	0.5160652	0.1442649	23.23959	0.4973481
Edibility: Poisonous	33.31791	0.5503290	0.1538432	24.78257	0.5303691

We wish to test

$$\mathcal{H}_0 : \text{No association between cap shape and edibility.} \tag{2.84}$$
$$\mathcal{H}_1 : \text{There exists an association between cap shape and edibility.} \tag{2.85}$$

at the 5% level of significance.

In order to calculate the χ^2-statistic, we need a table of expected values $\hat{E}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$. For example, for cell $(1,1)$, we have

$$\hat{E}_{11} = \frac{n_{1+} \times n_{+1}}{n_{++}} = \frac{1044 \times 113}{2023} = 58.32 \tag{2.86}$$

The table of $\hat{E}_{ij}$ are displayed in Table 2.5[14].

A table of

$$X_{ij}^2 = \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{2.87}$$

is shown in Table 2.6. For example, cell $(1,1)$ of this table was calculated as

$$X_{11}^2 = \frac{(O_{11} - \hat{E}_{11})^2}{\hat{E}_{11}} = \frac{(101 - 58.32...)^2}{58.32...} = 31.24... \tag{2.88}$$

We then have that

$$X^2 = \sum_{i,j} X_{ij}^2 = 114.98 \geq \chi^{2,\star}_{4,0.95} = 9.49 \tag{2.89}$$

The test thus rejects the null hypothesis that there is no association between mushroom cap and edibility.

[14]Recall that $\hat{E}_{ij}$ are estimated E_{ij}-values, because they are being estimated from the data.

Table 2.7: Cross-classification of a dataset of mushroom samples based on edibility and shape of cap.

	Cap shape: bell	flat	knobbed	convex/conical	Sum
Edibility: Edible	101	399	57	487	**1044**
Edibility: Poisonous	12	389	150	428	**979**
Sum	**113**	**788**	**207**	**915**	**2023**

Table 2.8: Estimated E_{ij} values.

	Cap shape: bell	flat	knobbed	convex/conical	Sum
Edibility: Edible	58.31537	406.6594	106.8255	472.1997	**1044**
Edibility: Poisonous	54.68463	381.3406	100.1745	442.8003	**979**
Sum	**113.00000**	**788.0000**	**207.0000**	**915.0000**	**2023**

- But hang on! Are we happy with the test that we have performed? Let's review the table. There are some very small values of $\hat{E}_{ij}$. The chi-square test is based on the *asymptotic behaviour* of X^2 towards a χ^2-distribution under the null hypothesis.
- Small values of $\hat{E}_{ij}$ (which stem from corresponding small n_{i+} or n_{+j} values) cripple this assumption, and have a disproportionally large affect on X^2. Consider the cap shape class $j =$ *conical*. We have 1 sample in this class; this had to belong to one or other of the Y classes for edibility. Regardless of which class this was, the X^2_{ij}-values for $j =$ *conical* would therefore contribute relatively substantially towards X^2, since $\hat{E}_{ij} \in (0, 1)$.
- It is therefore *standard practice to combine categories* with small totals with other categories.
- Without being a mushroom expert, I propose combining *conical* with *convex*, as shown in Table 2.7[15]. Now, in this case it is unlikely to make a huge effect, as we can see that some of the other categories have contributed far more substantially towards the high test statistic value, however, let's repeat the analysis with the combining of cell counts that we should have done.

The table of $\hat{E}_{ij}$-values are displayed in Table 2.8[16].

[15]This is an important point. If this was an analysis aimed at aiding us make important inferences (that is, if this weren't just an example but an exercise in making inferences for actually going mushroom-picking), I should consult a Biologist, or mushroom expert, before making these kinds of decisions about which mushrooms are deemed most similar. Alternatively, if there is no other basis for combining categories, we may combine the smallest categories into a single category *other* (you may have encountered this rule previously). Incidentally, you should not use the analysis presented here as a basis for mushroom-picking - I will not be held responsible for any ill effects of anybody eating dodgy mushrooms!

[16]By estimated E_{ij} values, I mean $\hat{E}_{ij}$, because they are being estimated from the data.

Table 2.9: X_{ij}^2 values.

	Cap shape: bell	**flat**	**knobbed**	**convex/conical**
Edibility: Edible	31.24352	0.1442649	23.23959	0.4638901
Edibility: Poisonous	33.31791	0.1538432	24.78257	0.4946898

The table of

$$X_{ij}^2 = \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{2.90}$$

is shown in Table 2.9.

We now have that

$$X^2 = \sum_{i,j} X_{ij}^2 = 113.84 \geq \chi^{2,\star}_{3,0.95} = 7.81 \tag{2.91}$$

In this case the test still rejects the null hypothesis that there is no association between mushroom cap and edibility.

2.4.3.5 Continuity Correction

In 1934, Yates (Yates [1934]) suggested to correct the Pearson's X^2 test statistic as given by Equation (2.39) as follows

$$X^2 = \sum_{j=1}^{k} \frac{(|O_j - E_j| - 0.5)^2}{E_j} \tag{2.92}$$

Why? In order to reduce the (upwardly biased) approximation error encountered by approximating discrete counts of categorical variables by the continuous chi-square distribution (more pronounced for small sample sizes[17]). The correction is therefore known as *continuity correction*, and reduces the Pearson's X^2 statistic value[18], and consequently increases the corresponding p-value. Yates's correction is by default applied in R when using `chisq.test` (Section 8.1.2), hence it is useful to know what it is!

2.4.4 Testing Homogeneity

- Interest here lies in examining whether the distribution of one variable (Y, say) is homogeneous regardless of the value of the other variable (X).
- The hypotheses under consideration in this case are:

$$\mathcal{H}_0: \quad \pi_{j|1} = \pi_{j|2} = ... = \pi_{j|I} \qquad \text{for } j = 1, ..., J \tag{2.93}$$

$$\mathcal{H}_1: \quad \mathcal{H}_0 \text{ is not true} \tag{2.94}$$

- Note that, if the null hypothesis is true, then combining the I populations together would produce one homogeneous population with regard to the distribution of the

[17] Think about the relative difference in the effect of applying Yates' continuity correction on the X_{ij}^2 values for $j =$ *conical* in table 2.6 compared to the other cap categories.

[18] Yates's correction has been criticised as perhaps being a little too much...

variable Y. In other words, we also have

$$\pi_{j|i} = \pi_{+j} \qquad \text{for all } i = 1, ..., I,\ j = 1, ..., J \tag{2.95}$$

2.4.4.1 Chi-Square Statistic

- Suppose $\pi_{j|i}$ are known. Now, consider the following statistic calculated from the observations in the i^{th} random sample, i.e. row-wise;

$$X_i^2 = \sum_{j=1}^{J} \frac{(n_{ij} - n_{i+}\pi_{j|i})^2}{n_{i+}\pi_{j|i}} = \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.96}$$

since $E_{ij} = n_{i+}\pi_{j|i}$.

- We know that $X_i^2 \sim \chi^2_{J-1}$ when n_{i+} is large, since X_i^2 is just the standard Chi-Square statistic for a random sample of n_{i+} observations from the i^{th} population. So, when we sum the X_i^2 quantities over I we get

$$X^2 = \sum_{i=1}^{I} X_i^2 = \sum_{i,j=1,1}^{I,J} \frac{(n_{ij} - n_{i+}\pi_{j|i})^2}{n_{i+}\pi_{j|i}} \tag{2.97}$$

and as this is the sum of chi-squared distributed random variables (for large samples) we obtain that[19]

$$X^2 \sim \chi^2_{I(J-1)} \tag{2.98}$$

- But... we don't know $\pi_{j|i}$ (recall supposition at the start of section), so we must consider how to estimate them under the null hypothesis $\mathcal{H}_0$. From Equation (2.95), under $\mathcal{H}_0$, we have that

$$\hat{\pi}_{j|i} = \hat{\pi}_{+j} \qquad \text{for all } i = 1, ..., I,\ j = 1, ..., J \tag{2.99}$$

2.4.4.1.1 ML Using Lagrange multipliers, we can show that[20]

$$\hat{\pi}_{+j} = \frac{n_{+j}}{n_{++}} \tag{2.100}$$

2.4.4.1.2 Which means... Given the maximum likelihood results of Section 2.4.4.1.1 and combining them with Equation (2.97) we have that

$$X^2 = \sum_{i,j=1,1}^{I,J} \frac{(n_{ij} - n_{i+}\hat{\pi}_{j|i})^2}{n_{i+}\hat{\pi}_{j|i}} \tag{2.101}$$

$$= \sum_{i,j=1,1}^{I,J} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{2.102}$$

where $\hat{E}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$, which is the same as in the test for independence.

[19] see Section 1.4.5.2

[20] Q2-6b involves showing that Equation (2.100) holds.

2.4.4.1.3 Degrees of Freedom The degrees of freedom are also the same, since when the parameters are assumed known, we have $I(J-1)$ degrees of freedom overall. When we estimate the parameters under $\mathcal{H}_0$ we have J potential outcomes, but the probability of one of those outcomes is constrained. Therefore, under $\mathcal{H}_0$ we estimate $J-1$ parameters (probabilities). Therefore, the degrees of freedom that remain are

$$\text{df} = I(J-1) - (J-1) = (I-1)(J-1) \tag{2.103}$$

2.4.4.1.4 Chi-Square Test of Homogeneity As a result of the above, the *Chi-Square test for independence and homogeneity are identical* and we reject $\mathcal{H}_0$ at significance level α if

$$X^2 \geq \chi^{2,\star}_{(I-1)(J-1),\alpha} \tag{2.104}$$

2.4.4.1.5 Example Following the example in Section 2.4.3.4, the chi-square test for homogeneity rejects the null hypothesis:

$$\mathcal{H}_0 : \pi_{j|1} = ... = \pi_{j|I} \qquad \text{for } j = 1, ..., J. \tag{2.105}$$

at the 0.05 level of significance, in favour of

$$\mathcal{H}_1 : \mathcal{H}_0 \text{ is not true.} \tag{2.106}$$

2.4.5 Relation to Generalised LR Test for Independence

- Recall Wilks' theorem (Section 1.2.3.4), which states that

$$W = 2\Big(\max_{\boldsymbol{\theta}\in\Omega} l(\boldsymbol{\theta}; \boldsymbol{X}) - \max_{\boldsymbol{\theta}_0\in\Omega_0} l(\boldsymbol{\theta}_0; \boldsymbol{X}\Big) \sim \chi^2_\nu \tag{2.107}$$

  for generic parameter $\boldsymbol{\theta} \in \mathbb{R}^k$, $\boldsymbol{\theta}_0 \in \mathbb{R}^{k_0}$, $\Omega_0 \subset \Omega$ and $\nu = k - k_0$.

- In our case, we have that the likelihood ratio statistic G^2 between the general model and that assuming independence is as follows:[21]

$$G^2 = 2\Big(\max_{\boldsymbol{\pi}} l(\boldsymbol{\pi}, \boldsymbol{n}) - \max_{\boldsymbol{\pi}_0} l(\boldsymbol{\pi}_0, \boldsymbol{n})\Big) \tag{2.108}$$

  where $\hat{\boldsymbol{\pi}}_0 = \arg\max_{\boldsymbol{\pi}_0} l(\boldsymbol{\pi}_0, \boldsymbol{n})$ is the MLE under $\mathcal{H}_0$ (as specified in Section 2.4.3.1), and $\hat{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi}} l(\boldsymbol{\pi}; \boldsymbol{n})$ is the MLE with no assumption of independence.

- Based on the likelihood expression of Section 2.3.2.2, we have that

$$G^2 = 2(\log K + \sum_{i,j} n_{ij} \log \hat{\pi}_{ij} - (\log K + \sum_{i,j} n_{ij} \log \hat{\pi}_{0,ij})) \tag{2.109}$$

  where $K = \frac{n_{++}!}{\prod_{i,j} n_{ij}!}$ is a constant.

[21] It is convention to denote this statistic as G^2 (rather than just G).

- From the ML expressions of Section 2.4.3.2, we have that

$$\begin{aligned} G^2 &= 2\Big(\sum_{i,j} n_{ij} \log\Big(\frac{n_{ij}}{n_{++}}\Big) - \sum_{i,j} n_{ij} \log\Big(\frac{n_{i+}}{n_{++}}\frac{n_{+j}}{n_{++}}\Big)\Big) && (2.110)\\ &= 2\sum_{i,j} n_{ij} \log\Big(\frac{n_{ij}n_{++}}{n_{i+}n_{+j}}\Big) && (2.111)\\ &= 2\sum_{i,j} n_{ij} \log\Big(\frac{n_{ij}}{\hat{E}_{ij}}\Big) && (2.112) \end{aligned}$$

  where $\hat{E}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$ as in Section 2.4.3.3.

- From Wilks' theorem we have that

$$G^2 \sim \chi^2_{(I-1)(J-1)} \tag{2.113}$$

  under the null hypothesis of independence between X and Y.

- Note the degrees of freedom. These are established as follows. The general case has $IJ-1$ parameters to estimate (since $\sum_{i,j} \pi_{ij} = 1$), and the restricted case has $(I-1)+(J-1)$. Therefore the degrees of freedom are

$$\nu = k - k_0 = IJ - 1 - ((I-1)+(J-1)) = (I-1)(J-1) \tag{2.114}$$

- Hence the generalised LR hypothesis test is to reject $\mathcal{H}_0$ at significance level α if

$$G^2 \geq \chi^{2,\star}_{(I-1)(J-1),\alpha} \tag{2.115}$$

  where $\chi^{2,\star}_{(I-1)(J-1),\alpha}$ is the $(1-\alpha)$ quantile of the $\chi^2_{(I-1)(J-1)}$-distribution.

2.4.5.1 Conclusion

We can either use the Pearson Chi-Square or generalised LR test statistic to test for independence. We also have that

$$G^2 \approx \sum_{i,j} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = X^2 \tag{2.116}$$

further illustrating the similarity between these two tests. Equation (2.116) can be shown by looking at a second-order Taylor expansion of $x\log(x/x_0)$ around x_0[22].

2.4.5.2 Example

Following from Section 2.4.3.4, and using the data in Tables 2.7 and 2.8, we calculate Table 2.10 of G^2_{ij} values by calculating

$$G^2_{ij} = 2n_{ij} \log \frac{n_{ij}}{\hat{E}_{ij}} \tag{2.117}$$

for each cell (i,j). For example, for cell $(1,3)$, we have that

$$G^2_{13} = 2n_{13} \log \frac{n_{13}}{\hat{E}_{13}} = 2 \times 57 \log \frac{57}{106.82...} = 2 \times (-35.804...) = -71.608... \tag{2.118}$$

Table 2.10: G^2_{ij} values.

	Cap shape: bell	**flat**	**knobbed**	**convex/conical**
Edibility: Edible	110.94946	-15.17365	-71.60858	30.05971
Edibility: Poisonous	-36.40022	15.47166	121.11651	-29.10030

The G^2-statistic is therefore given as

$$G^2 = \sum_{i,j} G^2_{ij} = \sum_{i,j} n_{ij} \log \frac{n_{ij}}{\hat{E}_{ij}} = 110.94... + (-15.17...) + ... = 125.31 \tag{2.119}$$

Since $G^2 \geq \chi^{2,\star}_{3,0.95} = 7.81...$, the test rejects $\mathcal{H}_0$ at the 5% level of significance.

2.4.6 Analysis of Residuals

Upon rejecting the null hypothesis $\mathcal{H}_0$ of independence, or more generally any $\mathcal{H}_0$, interest lies on detecting parts of the contingency table (single cells or whole regions) that contribute more in the value of the goodness-of-fit statistic, i.e., parts of the table that are mainly responsible for the rejection of $\mathcal{H}_0$. One way to do this is by analysing the residuals and detecting large deviations.

2.4.6.1 Basic Residuals

- *Basic* residuals are defined by
$$e_{ij} = n_{ij} - \hat{E}_{ij} \tag{2.120}$$
- We can examine their sign and magnitude.
- Detection of a systematic structure of the signs of the residuals is of special interpretational interest.
- However, the evaluation of the importance of the contribution of a particular cell based on these residuals can be misleading.

2.4.6.2 Standardised Residuals

- *Standardised* residuals are defined by
$$e^{\star}_{ij} = \frac{e_{ij}}{\sqrt{\text{Var}[e_{ij}]}} = \frac{n_{ij} - \hat{E}_{ij}}{\sqrt{\text{Var}[\hat{E}_{ij}]}} \tag{2.121}$$
which measure the difference between n_{ij} and $\hat{E}_{ij}$ in terms of standard errors.
- However, $\text{Var}[\hat{E}_{ij}]$ must be estimated in some way, hence Pearson and Adjusted standardised residuals (Sections 2.4.6.3 and 2.4.6.4 respectively).

[22] Q2-7 concerns showing that Approximation (2.116) holds.

2.4.6.3 Pearson's Residuals

- For Poisson sampling, $\text{Var}[\hat{E}_{ij}] = E_{ij}$, which can be estimated by $\hat{E}_{ij}$.
- This results in *Pearson's* residuals:

$$e_{ij}^P = \frac{n_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}}} \tag{2.122}$$

- Notice that

$$\sum_{i,j} (e_{ij}^P)^2 = X^2 \tag{2.123}$$

  hence they directly measure the contribution towards the X^2 statistic.
- Under multinomial sampling, $\text{Var}[\hat{E}_{ij}]$ is different, hence e_{ij}^P are still asymptotically $\mathcal{N}(0, v_{ij})$ distributed, but $v_{ij} \neq 1$ in this case.
- We would like something invariant of sampling scheme and asymptotically $\mathcal{N}(0, 1)$ distributed. Hence...

2.4.6.4 Adjusted Residuals

- In 1973, Shelby Haberman (Haberman [1973]) proved that under independence and for multinomial sampling, as $n_{++} \to \infty$,

$$v_{ij} = v_{ij}(\boldsymbol{\pi}) = (1 - \pi_{i+})(1 - \pi_{+j}) \tag{2.124}$$

- Haberman suggested use of *Adjusted* standardised residuals[23], which are defined by

$$e_{ij}^s = \frac{e_{ij}^P}{\sqrt{\hat{v}_{ij}}} = \frac{n_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}\hat{v}_{ij}}} \tag{2.125}$$

  where

$$\hat{v}_{ij} = \Big(1 - \frac{n_{i+}}{n_{++}}\Big)\Big(1 - \frac{n_{+j}}{n_{++}}\Big) \tag{2.126}$$

  are maximum likelihood estimates for v_{ij} as $n_{++} \to \infty$.
- Haberman proved that e_{ij}^s are asymptotially standard normal distributed under a multinomial sampling scheme.
- These residuals are commonly used for both Poisson and Multinomial sampling schemes when assessing independence (or causes for deviation from it).

2.4.6.5 Deviance Residuals

Deviance residuals are defined by the square root (with appropriate sign) cell components of the G^2-statistic. They are equal to

$$e_{ij}^d = \text{sign}(n_{ij} - \hat{E}_{ij}) \sqrt{2n_{ij} \Big| \log\Big(\frac{n_{ij}}{\hat{E}_{ij}}\Big) \Big|} \tag{2.127}$$

[23] Some texts refer to these residuals simply as *standardised* residuals since they are the most common choice.

Table 2.11: e_{ij} values for mushroom example.

	Cap shape: bell	**flat**	**knobbed**	**convex/conical**
Edibility: Edible	42.68463	-7.659417	-49.82551	14.8003
Edibility: Poisonous	-42.68463	7.659417	49.82551	-14.8003

Table 2.12: e_{ij}^P values for mushroom example.

	Cap shape: bell	**flat**	**knobbed**	**convex/conical**
Edibility: Edible	5.589590	-0.3798221	-4.820746	0.6810948
Edibility: Poisonous	-5.772167	0.3922285	4.978209	-0.7033419

2.4.6.6 Example: Mushrooms

Let's calculate each of the residuals for cell $(1, 3)$ of the mushrooms data Tables 2.7 and 2.8.

- Basic:
$$e_{13} = n_{13} - \hat{E}_{13} = 57 - 106.82... = -49.82... \tag{2.128}$$

- Pearson:
$$e_{13}^P = \frac{n_{13} - \hat{E}_{13}}{\sqrt{\hat{E}_{13}}} = \frac{57 - 106.82...}{\sqrt{106.82...}} = -4.82... \tag{2.129}$$

- Adjusted standardised:
$$\hat{v}_{13} = \Big(1 - \frac{n_{1+}}{n_{++}}\Big)\Big(1 - \frac{n_{+3}}{n_{++}}\Big) = \Big(1 - \frac{1044}{2023}\Big)\Big(1 - \frac{207}{2023}\Big) = 0.4344... \tag{2.130}$$
hence
$$e_{13}^s = \frac{n_{13} - \hat{E}_{13}}{\sqrt{\hat{E}_{13}\hat{v}_{13}}} = \frac{57 - 106.82...}{\sqrt{106.82 \times 0.4344}} = -7.314... \tag{2.131}$$

- Deviance:
$$e_{13}^d = \text{sign}(n_{13} - \hat{E}_{13})\sqrt{2n_{13}\Big|\log\Big(\frac{n_{13}}{\hat{E}_{13}}\Big)\Big|} = -\sqrt{2 \times 57\Big|\log\Big(\frac{57}{106.82...}\Big)\Big|} = -8.46... \tag{2.132}$$

Residuals for all cells are shown in Tables 2.11 to 2.14 respectively.

Question: What do these residuals tell us?

- Analysis of the basic residuals might suggest that the knobbed mushrooms contribute most towards the X^2 statistic. However, this would be misleading - they have a larger basic residual simply because there was a larger number of knobbed mushrooms than flat mushrooms in our sample overall.

Table 2.13: e_{ij}^s values for mushroom example.

	Cap shape: bell	**flat**	**knobbed**	**convex/conical**
Edibility: Edible	8.269282	-0.6987975	-7.314099	1.322946
Edibility: Poisonous	-8.269282	0.6987975	7.314099	-1.322946

Table 2.14: e_{ij}^d values for mushroom example.

	Cap shape: bell	**flat**	**knobbed**	**convex/conical**
Edibility: Edible	10.53326	-3.895338	-8.462185	5.482674
Edibility: Poisonous	-6.03326	3.933403	11.005295	-5.394469

- In terms of Pearson residuals, we can see that the bell cap-shaped mushrooms contribute most towards the X^2 statistic. This would make sense, as they have the largest discrepancy between observed and expected values in terms of standard deviations, even from just casually observing the relative differences (in terms of proportions) between the values in Tables 2.7 and 2.8.
- The adjusted residuals (unsurprisingly) tell a similar story to the Pearson's residuals in terms of the magnitude of each one relative to the total, and are necessarily symmetric about the binary variable.
- The deviance residuals show the overall contribution towards the G^2 test statistic. You will also notice that calculating $\text{sign}(e_{ij})e_{ij}^2$ for each cell necessarily leads us to the values presented in Table 2.10.

2.5 Odds Ratios

- Decompose the $I \times J$ table into a minimal set of $(I-1)(J-1)$ 2×2 tables able to fully describe the problem in terms of odds ratios.
- However, this decomposition is not unique. We look at some popular choices.
- For nominal classification variables this set of basic 2×2 tables is defined in terms of a reference category, usually the cell (I, J).
- In this case, a 2×2 submatrix given by

$$\begin{pmatrix} \pi_{ij} & \pi_{iJ} \\ \pi_{Ij} & \pi_{IJ} \end{pmatrix} \tag{2.133}$$

  is generated for each of the remaining cells (i, j), $i = 1, ..., I-1$, $j = 1, ..., J-1$.
- You will notice that each 2×2 table has:
  - in its upper diagonal cell, the (i, j) cell of the initial table, $i = 1, ..., I-1$, $j = 1, ..., J-1$,

  - in its lower diagonal cell, the reference cell (I, J).
  - in its non-diagonal cells, the cells of the initial table that share one classification variable index with each diagonal cell, i.e., the cells (i, J) and (I, j).

- Given $\boldsymbol{\pi}$, we could calculate a set of *nominal* odds ratios corresponding to each of the submatrices, given as

$$r_{ij}^{IJ} = \frac{\pi_{ij}/\pi_{Ij}}{\pi_{iJ}/\pi_{IJ}} = \frac{\pi_{ij}/\pi_{iJ}}{\pi_{Ij}/\pi_{IJ}} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{2.134}$$

- How do we interpret the odds ratio here? Well, r_{ij}^{IJ} is a multiplicative measure of:
  - the relative difference between $Y = j$ and $Y = J$ in the odds of $X = i$ to $X = I$;
  - the relative difference between $X = i$ and $X = I$ in the odds of $Y = j$ to $Y = J$.

- The diagonal cells are indicated in the sub- and superscript of the notation. Of course, any cell (r, c) of the table could serve as reference category and the nominal odds ratios are then defined analogously for all $i \neq r$, $j \neq c$.

- As usual, we don't know $\boldsymbol{\pi}$, hence we are often interested in the set of sample odds ratios corresponding to each of the submatrices, given as:

$$\hat{r}_{ij}^{IJ} = \frac{\hat{\pi}_{ij}/\hat{\pi}_{Ij}}{\hat{\pi}_{iJ}/\hat{\pi}_{IJ}} = \frac{n_{ij}n_{IJ}}{n_{Ij}n_{iJ}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{2.135}$$

2.5.0.1 Example: Mushrooms

- Assuming a reference category of $(I, J) = (2, 4)$, we can calculate 3 submatrices for cells $(1, 1)$, $(1, 2)$ and $(1, 3)$ respectively as

$$M_{11}^{24} = \begin{pmatrix} 101 & 487 \\ 12 & 428 \end{pmatrix} \quad M_{12}^{24} = \begin{pmatrix} 399 & 487 \\ 389 & 428 \end{pmatrix} \quad M_{13}^{24} = \begin{pmatrix} 57 & 487 \\ 150 & 428 \end{pmatrix} \tag{2.136}$$

- These submatrices yield sample odds ratios of

$$\hat{r}_{11}^{24} = \frac{101 \times 428}{12 \times 487} = 7.397... \tag{2.137}$$

$$\hat{r}_{12}^{24} = 0.901... \tag{2.138}$$

$$\hat{r}_{13}^{24} = 0.334... \tag{2.139}$$

- This shows that:
  - the relative odds of picking a randomly selected edible ($X = 1$) mushroom (to a poisonous ($X = 2$) mushroom) with a bell-shaped cap ($Y = 1$) from this sample is about 7.4 times greater than the odds of picking an edible ($X = 1$) mushroom (to a poisonous ($X = 2$) mushroom) with a convex/conical-shaped cap ($Y = 4$);
  - the sample relative odds of picking an edible mushroom with a flat cap is about the same as the odds of picking an edible mushroom with a convex/conical shaped cap;

  - the sample relative odds of picking an edible mushroom with a knobbed-shaped cap is about 1/3 of the size of the odds of picking an edible mushroom with a convex/conical shaped cap.

- Comparisons of other categories can be directly obtained from these odds ratios. For example, the odds ratio between bell and flat-capped mushrooms in the odds between edible and poisonous is as follows

$$r_{11}^{22} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\frac{\pi_{11}/\pi_{21}}{\pi_{14}/\pi_{24}}}{\frac{\pi_{12}/\pi_{22}}{\pi_{14}/\pi_{24}}} = \frac{r_{11}^{24}}{r_{12}^{24}} \tag{2.140}$$

  so that the sample odds ratio is given by

$$\hat{r}_{11}^{22} = \frac{\hat{r}_{11}^{24}}{\hat{r}_{12}^{24}} = 7.397.../0.901... = 8.21 \tag{2.141}$$

  which shows that the relative odds of picking a randomly selected edible ($X = 1$) mushroom (to a poisonous ($X = 2$) mushroom) with a bell-shaped cap ($Y = 1$) from this sample is about 8.2 times greater than the odds of picking an edible ($X = 1$) mushroom (to a poisonous ($X = 2$) mushroom) with a flat-shaped cap ($Y = 2$).

2.5.1 Log Odds Ratios, Confidence Intervals and Hypothesis Testing

2.5.1.1 2×2 Tables

- Recall for 2×2 tables, the sample odds ratio is simply[24]

$$\hat{r}_{12} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{2.142}$$

- $\hat{r}_{12}$ is not very well approximated by a normal distribution. In particular, note that $\hat{r}_{12}$ is restricted to being non-negative.
- In comparison, odds ratios are intuitively interpretable on log scale, as we have that

$$\log r = 0 \text{ corresponds to independence.} \tag{2.143}$$
$$\log r > 0 \text{ corresponds to positive odds ratio.} \tag{2.144}$$
$$\log r < 0 \text{ corresponds to negative odds ratio.} \tag{2.145}$$

- Additionally, a positive log odds ratio has symmetric meaning to a negative log odds ratio of equal absolute value. For example;
  - $\log \hat{r}_{12} = 3$ implies that the odds of $X = 1$ to $X = 2$ is e^3 times larger for $Y = 1$ than $Y = 2$.
  - $\log \hat{r}_{12} = -3$ implies that the odds of $X = 1$ to $X = 2$ is e^3 times larger for $Y = 2$ than $Y = 1$.

[24] Note that this would be notated as r_{11}^{22} using the above more general odds ratio notation for an $I \times J$ table.

- Perhaps unsurprisingly then, $\log(\hat{r}_{12})$ is better normally approximated than $\hat{r}_{12}$. In particular, we have that asymptotically

$$\log \hat{r}_{12} \sim \mathcal{N}(\log r_{12}, \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}) \tag{2.146}$$

- A $(1-\alpha)$ confidence interval for r_{12} can be derived by

$$(e^{A(\hat{r},0)}, e^{A(\hat{r},2)}) \tag{2.147}$$

where

$$A(\hat{r}, c) = \log \hat{r} - (1-c) z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \tag{2.148}$$

- Hypotheses about r, like

$$\mathcal{H}_0 : r = r_0 \quad \iff \quad \log r = \log r_0 \tag{2.149}$$

for a known value of r_0, can be tested by the associated Z-test

$$Z = \frac{\log \hat{r} - \log r_0}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \tag{2.150}$$

with $Z \sim \mathcal{N}(0,1)$ under $\mathcal{H}_0$.

2.5.1.1.1 Testing Independence

- Since $r_{12} = 1 \iff \pi_{11}/\pi_{12} = \pi_{21}/\pi_{22}$, the above hypothesis for $r_0 = 1$ is equivalent to the hypothesis of independence or heterogeneity. However, Equation (2.150) is a Wald test and is not equivalent to the X^2 or G^2 tests given by Equations (2.79) and (2.115) respectively.

2.5.1.1.2 Example

- Using the data from Table 2.2, we wish to test the hypothesis:

$$\mathcal{H}_0 : r_{12} = 1 \tag{2.151}$$
$$\mathcal{H}_1 : r_{12} \neq 1 \tag{2.152}$$

- $\hat{r}_{12}$ is given by

$$\hat{r}_{12} = \frac{62 \times 12}{8 \times 18} = 5.166... \tag{2.153}$$

- The Z-statistic is given by

$$Z = \frac{\log \hat{r}_{12}}{\sqrt{\frac{1}{62} + \frac{1}{18} + \frac{1}{8} + \frac{1}{12}}} = 3.103... \tag{2.154}$$

- The test rejects $\mathcal{H}_0$ at the 5% level of significance since Z lies outside of the interval

$$(Z^\star_{0.025}, Z^\star_{0.975}) = (-1.96, 1.96) \tag{2.155}$$

thus providing evidence of an association between high dose and low dose treatment outcome.

Table 2.15: Hypothetical data of a trial comparing the reponse of different doses of a particular treatment.

	Result: Success	**Partial**	**Failure**	**Sum**
Dose: High	47	25	12	**84**
Dose: Medium	36	22	18	**76**
Dose: Low	41	60	55	**156**
Sum	**124**	**107**	**85**	**316**

2.5.1.2 $I \times J$ Tables

- Null hypotheses for $I \times J$ contingency tables can be tested too, with general null hypotheses assuming values for one or more (up to $(I-1)(J-1)$) odds ratio values.
- For example, the independence hypothesis (2.59) is equivalent to the hypothesis that all odds ratios in a minimal set are equal to 1.
- For nominal odds ratios, this would equivalently be

$$r_{ij}^{IJ} = 1, \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{2.156}$$

- This hypothesis is more easily assessed using log-linear models (later...!).

2.6 Ordinal Variables

An *ordinal* (or ordered) categorical variable is a variable for which the categories exhibit a natural ordering.

2.6.0.1 Example

Consider the hypothetical data considered in Table 2.15 which cross-classifies the level of success (with *partial* standing for partial success) of a treatment given as either a high, medium or low dose. There is clearly a natural ordering to both the result and the dose level.

2.6.1 Scores and The Linear Trend Test of Independence

2.6.1.1 Scores

- When both classification variables of a contingency table are ordinal, we are interested in the direction of the underlying association (positive or negative).
- The ordering information of a classification variable is captured in scores, assigned to its categories.
- Thus, for an $I \times J$ table, let $x_1 \leq x_2 \leq ... \leq x_I$ and $y_1 \leq y_2 \leq ... \leq y_J$ be the scores assigned to the categories of the row and column classification variables X and Y respectively, with $x_1 < x_I$ and $y_1 < y_J$.

- The structure of the underlying association is then expressed through relations among the scores.

2.6.1.2 Pearson's Correlation for Linear Trend

- A first sensible assumption is that association exhibits a linear trend.
- The linear trend is measured by Pearson's correlation coefficient ρ_{XY} between X and Y, defined through their categories' scores as

$$\rho_{XY} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - (\mathrm{E}[X])^2}\sqrt{\mathrm{E}[Y^2] - (\mathrm{E}[Y])^2}} \tag{2.157}$$

2.6.1.3 Sample Correlation Coefficient

The sample correlation coefficient is given as

$$r_{XY} = \frac{n_{++}\sum_{k=1}^{n_{++}} x_k y_k - \sum_{k=1}^{n_{++}} x_k \sum_{k=1}^{n_{++}} y_k}{\sqrt{n_{++}\sum_{k=1}^{n_{++}} x_k^2 - (\sum_{k=1}^{n_{++}} x_k)^2}\sqrt{n_{++}\sum_{k=1}^{n_{++}} y_k^2 - (\sum_{k=1}^{n_{++}} y_k)^2}} \tag{2.158}$$

In the context of a contingency table, we have that:

$$\sum_{k=1}^{n_{++}} x_k y_k \quad = \quad \sum_{i,j} x_i y_j n_{ij} \tag{2.159}$$

$$\sum_{k=1}^{n_{++}} x_k = \sum_i x_i n_{i+} \qquad \sum_{k=1}^{n_{++}} y_k = \sum_j y_j n_{+j} \tag{2.160}$$

$$\sum_{k=1}^{n_{++}} x_k^2 = \sum_i x_i^2 n_{i+} \qquad \sum_{k=1}^{n_{++}} y_k^2 = \sum_j y_j^2 n_{+j} \tag{2.161}$$

so that

$$r_{XY} = \frac{n_{++}\sum_{i,j} x_i y_j n_{ij} - \sum_i x_i n_{i+}\sum_j y_j n_{+j}}{\sqrt{n_{++}\sum_i x_i^2 n_{i+} - (\sum_i x_i n_{i+})^2}\sqrt{n_{++}\sum_j y_j^2 n_{+j} - (\sum_j y_j^2 n_{+j})^2}} \tag{2.162}$$

2.6.1.4 Linear Trend Test

- The linear trend test restricts interest to linearly associated classification variables and tests the significance of ρ_{XY}.
- For testing independence, the hypotheses in question are

$$\mathcal{H}_0 : \rho_{XY} = 0 \tag{2.163}$$

$$\mathcal{H}_1 : \rho_{XY} \neq 0 \tag{2.164}$$

- The corresponding test statistic is

$$M^2 = (n-1) r_{XY}^2 \tag{2.165}$$

which was shown by Nathan Mantel (Mantel [1963]) to asymptotically satisfy

$$M^2 \sim \chi_1^2 \tag{2.166}$$

under $\mathcal{H}_0$.

Table 2.16: Cross multiplied XY scores for the Dose-Result example.

	Result: Success	**Partial**	**Failure**
Dose: High	1	2	3
Dose: Medium	2	4	6
Dose: Low	3	6	9

- From Equation (2.166), we also have that

$$R = \text{sign}(r_{XY)})\sqrt{M^2} \sim \mathcal{N}(0,1) \tag{2.167}$$

under $\mathcal{H}_0$, and can be used for one-sided alternative hypotheses.

2.6.1.5 Choice of Scores

- Scores are a powerful tool in the analysis of ordinal contingency tables and the development of special, very informative models.
- Different scoring systems can lead to different results. There is no direct way to measure the sensitivity of an analysis to the choice of scores used. However, test results may be particularly sensitive to the choice of scoring system when the margins of the table are highly unbalanced, or even just if some cells have considerably larger frequencies than the others.
- The most common scores used are

(a) the equally spaced scores, appropriate for ordinal classification variables, usually set equal to the category order $(1, 2, ...)$,
(b) the category midpoints for interval classification variables, and
(c) the midranks. Midranks assign to each category the mean of the ranks of its cases, where all items of the sample are ranked from 1 to n_{++}.

2.6.1.6 Example

We consider use of equally spaced scores in continuation of the example in Section 2.6.0.1 as follows:

$$D = \{high, medium, low\} = \{1, 2, 3\} \tag{2.168}$$
$$R = \{success, partial, failure\} = \{1, 2, 3\} \tag{2.169}$$

We therefore have a table of XY cross-multiplied scores given as shown in Table 2.16.

We calculate the following:

$$\sum_{k=1}^{n_{++}} x_k y_k = \sum_{i,j} x_i y_j n_{ij} = 1 \times 47 + 2 \times 25 + 3 \times 12 + 2 \times 36 + 4 \times 22 + \quad (2.170)$$

$$6 \times 18 + 3 \times 41 + 6 \times 60 + 9 \times 55 = 1379 \quad (2.171)$$

$$\sum_{k=1}^{n_{++}} x_k = \sum_{i=1}^{I} x_i n_{i+} = 1 \times 84 + 2 \times 76 + 3 \times 156 = 704 \quad (2.172)$$

$$\sum_{k=1}^{n_{++}} y_k = 1 \times 124 + 2 \times 107 + 3 \times 85 = 593 \quad (2.173)$$

$$\sum_{k=1}^{n_{++}} x_k^2 = \sum_{i=1}^{I} x_i^2 n_{i+} = 1^2 \times 84 + 2^2 \times 76 + 3^2 \times 156 = 1792 \quad (2.174)$$

$$\sum_{k=1}^{n_{++}} y_k^2 = 1^2 \times 124 + 2^2 \times 107 + 3^2 \times 85 = 1317 \quad (2.175)$$

so that

$$r_{XY} = \frac{316 \times 1379 - 704 \times 593}{\sqrt{316 \times 1792 - 704^2}\sqrt{316 \times 1317 - 593^2}} = 0.2709... \quad (2.176)$$

and

$$M^2 = (n-1) r_{XY}^2 = 315 \times 0.2709...^2 = 23.119... \quad (2.177)$$

for which we have that

$$P(\chi_1^2 \geq M^2) = 1.52 \times 10^{-6} \quad (2.178)$$

and so the test rejects the null hypothesis of independence at the 5% (or 1%, or less...) level of significance, and provides evidence of an association between dose level and result.

2.6.2 Odds Ratios

- There are various types of odds ratios available for ordinal variables.
- A fixed reference cell is no longer meaningful...

2.6.2.1 Local Odds Ratios

- *Local* odds ratios refer to locally involving two successive categories.
- We form $(I-1)(J-1)$ local 2×2 tables by taking;
  - two successive rows i and $i+1$, and
  - two successive columns j and $j+1$ of the original table, that is

$$\begin{pmatrix} \pi_{ij} & \pi_{i,j+1} \\ \pi_{i+1,j} & \pi_{i+1,j+1} \end{pmatrix} \quad (2.179)$$

- Such submatrices lead to a minimal set of $(I-1)(J-1)$ local odds ratios, given by

$$r_{ij}^L = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \quad (2.180)$$

- This minimal set of odds ratios are sufficient to describe association and derive odds ratios for any other 2×2 table formed by non-successive rows or columns, because

$$r_{ij}^{i+k,j+l} = \frac{\pi_{ij}\pi_{i+k,j+l}}{\pi_{i+k,j}\pi_{i,j+l}} = \prod_{a=0}^{k-1}\prod_{b=0}^{l-1} r_{i+a,j+b}^{L} \tag{2.181}$$

2.6.2.2 Global Odds Ratios

- *Global* odds ratios treat the variables cumulatively, and are defined by

$$r_{ij}^{G} = \frac{\left(\sum_{k\le i}\sum_{l\le j}\pi_{kl}\right)\left(\sum_{k>i}\sum_{l>j}\pi_{kl}\right)}{\left(\sum_{k\le i}\sum_{l>j}\pi_{kl}\right)\left(\sum_{k>i}\sum_{l\le j}\pi_{kl}\right)} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{2.182}$$

2.6.2.2.1 Illustration

- As an illustration for a 5×5 table, look at Figure 2.6.

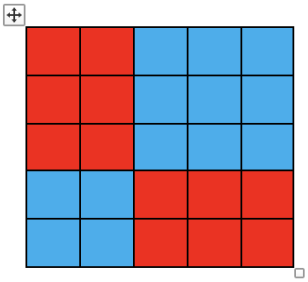


Figure 2.6: Illustration of global odds ratio.

- For cell $(i = 3, j = 2)$, the global odds ratio is defined as follows:
  - in the numerator is the product of the sums of the two groups of cells in red, while
  - in the denominator is the product of the sums of the two groups of cells in blue.

2.6.2.3 Local or Global Odds Ratios?

- Odds ratios r_{ij}^{L} and r_{ij}^{G} treat both classification variables in a symmetric way (r_{ij}^{L} locally and r_{ij}^{G} cumulatively).
- We can also just treat one classification variable cumulatively and the other locally.
- For example, we can treat the columns' variable Y cumulatively and the row variable X locally. Then the odds ratio is

$$r_{ij}^{C_Y} = \frac{(\sum_{l\le j}\pi_{il})(\sum_{l>j}\pi_{i+1,l})}{(\sum_{l>j}\pi_{il})(\sum_{l\le j}\pi_{i+1,l})} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{2.183}$$

- The cumulative odds ratio $r_{ij}^{C_X}$ is defined analogously, and is cumulative with respect to the rows, being applied on successive columns j and $j+1$.

Table 2.17: Local Odds Ratios between successive row categories i and $i+1$ and column categories j and $j+1$, these being indicated in each case by the row and column names.

	Result: Success-Partial	**Partial-Failure**
Dose: High-Medium	1.149	1.705
Dose: Medium-Low	2.395	1.120

- Cumulative and global odds ratios make sense for ordinal classification variables. They are also meaningful for tables with one ordinal classification variable and one binary.
- The ordinality of the classification variables is required only whenever a classification variable is treated cumulatively. Thus, the local odds ratios are also appropriate for nominal variables.

2.6.2.4 Sample Odds Ratios

- Corresponding sample odds ratios are calculated analogously to previous odds ratios, with, for example, the sample local odds ratio being defined as

$$\hat{r}_{ij}^{L} = \frac{n_{ij}n_{i+1,j+1}}{n_{i+1,j}n_{i,j+1}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{2.184}$$

  and the sample global odds ratio being defined as

$$\hat{r}_{ij}^{G} = \frac{\left(\sum_{k\leq i}\sum_{l\leq j} n_{kl}\right)\left(\sum_{k>i}\sum_{l>j} n_{kl}\right)}{\left(\sum_{k\leq i}\sum_{l>j} n_{kl}\right)\left(\sum_{k>i}\sum_{l\leq j} n_{kl}\right)} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{2.185}$$

- If the linear trend test provides evidence of association, sample odds ratios $\hat{r}$ can be used to investigate the association further.

2.6.2.5 Example

We calculate the different types of sample odds ratios for cell $(2, 2)$ of Table 2.15, with the odds ratios corresponding to the remaining cells presented in Tables 2.17 and 2.18 respectively.

$$\hat{r}_{22}^{L} = \frac{n_{22}n_{33}}{n_{32}n_{23}} = \frac{22 \times 55}{60 \times 18} = 1.12037 \tag{2.186}$$

$$\hat{r}_{22}^{G} = \frac{(47 + 25 + 36 + 22) \times 55}{(41 + 60) \times (12 + 18)} = 2.359... \tag{2.187}$$

2.6.2.6 Expected (under hypothesis) Odds Ratios

- The value of r_{ij} under a specific hypothesis $\mathcal{H}_0$ can also be calculated, and given as $r_{0,ij}$. In this case, we would utilise the corresponding expected frequencies E_{ij} from each cell (i, j). For example, $r_{0,ij}^{L}$ is given by:

$$r_{0,ij}^{L} = \frac{E_{ij}E_{i+1,j+1}}{E_{i+1,j}E_{i,j+1}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1$$

Table 2.18: Global Odds Ratios as defined in the text, with (i, j) being denoted by the row and column names.

	Result: Success	**Partial**
Dose: High	2.557	2.755
Dose: Medium	3.023	2.360

- As we know, E_{ij} are often not known under H_0 and thus need to be estimated by $\hat{e}_{ij}$. In this case, the maximum likelihood estimate of $r_{0,ij}$ under a specific hypothesis $\mathcal{H}_0$ can be utilised: For example, the ML estimate of $r^L_{0,ij}$ is given by:

$$\hat{r}^L_{0,ij} = \frac{\hat{E}_{ij}\hat{E}_{i+1,j+1}}{\hat{E}_{i+1,j}\hat{E}_{i,j+1}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1$$

Chapter 3

Multi-Way Contingency Tables

Multi-way contingency tables are very common in practice, derived by the presence of more than two cross-classification variables.

3.1 Description

3.1.1 Three-way Tables

- Consider an $I \times J \times K$ contingency table (n_{ijk}) for $i = 1, ..., I$, $j = 1, ..., J$ and $k = 1, ..., K$, with classification variables X (the rows), Y (the columns) and Z (the layers) respectively.
- A schematic of a generic $X \times Y \times Z$ contingency table of counts is shown in Figure 3.1.
- We can define the joint probability distribution of (X, Y, Z) as

$$\pi_{ijk} = P(X = i, Y = j, Z = k) \tag{3.1}$$

- Proportions, observed and random counts are defined similarly to the $I \times J$ contingency table cases....

3.1.1.1 Example

The table in Figure 3.2 shows an example of a 3-way contingency table. This hypothetical data cross-classifies the response (Y) to a treatment drug (X) at one of two different clinics (Z).

3.1.1.2 Partial/Conditional Tables

- *Partial*, or *conditional*, tables involve fixing the category of one of the variables.
- We denote the fixed variable in parentheses.
- For example, the set of XY-partial tables consist of the K corresponding two-way layers, denoted as $(n_{ij(k)})$ for $k = 1, ..., K$.

		$Y=1$	$Y=2$	$\cdots$	$Y=j$	$\cdots$	$Y=J$	$Y=.$
	$X=1$	n_{111}	n_{121}	$\cdots$	n_{1j1}	$\cdots$	n_{1J1}	n_{1+1}
	$X=2$	n_{211}	n_{221}	$\cdots$	n_{2j1}	$\cdots$	n_{2J1}	n_{2+1}
$Z=1$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$X=i$	n_{i11}	n_{i21}	$\cdots$	n_{ij1}	$\cdots$	n_{iJ1}	n_{i+1}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$X=I$	n_{I11}	n_{I21}	$\cdots$	n_{Ij1}	$\cdots$	n_{IJ1}	n_{I+1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$X=1$	n_{11k}	n_{12k}	$\cdots$	n_{1jk}	$\cdots$	n_{1Jk}	n_{1+k}
	$X=2$	n_{21k}	n_{22k}	$\cdots$	n_{2jk}	$\cdots$	n_{2Jk}	n_{2+k}
$Z=k$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$X=i$	n_{i1k}	n_{i2k}	$\cdots$	n_{ijk}	$\cdots$	n_{iJk}	n_{i+k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$X=I$	n_{I1k}	n_{I2k}	$\cdots$	n_{Ijk}	$\cdots$	n_{IJk}	n_{I+k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$X=1$	n_{11K}	n_{12K}	$\cdots$	n_{1jK}	$\cdots$	n_{1JK}	n_{1+K}
	$X=2$	n_{21K}	n_{22K}	$\cdots$	n_{2jK}	$\cdots$	n_{2JK}	n_{2+K}
$Z=K$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$X=i$	n_{i1K}	n_{i2K}	$\cdots$	n_{ijK}	$\cdots$	n_{iJK}	n_{i+K}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$X=I$	n_{I1K}	n_{I2K}	$\cdots$	n_{IjK}	$\cdots$	n_{IJK}	n_{I+K}
$Z=.$	$X=.$	n_{+1+}	n_{+2+}	$\cdots$	n_{+j+}	$\cdots$	n_{+J+}	n_{+++}

Figure 3.1: Generic I x J x K contingency table of counts.

		Response (Y)	
Clinic (Z)	Drug Treatment (X)	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32

Figure 3.2: Table cross-classifying hypothetical treatment drug, response and clinic.

- XZ and YZ- partial tables are denoted as $(n_{i(j)k})$ and $(n_{(i)jk})$ respectively.
- Partial/conditional probabilities:

$$\pi_{ij(k)} = \pi_{ij|k} = P(X = i, Y = j|Z = k) = \frac{\pi_{ijk}}{\pi_{++k}} \qquad k = 1, ..., K \tag{3.2}$$

- Partial/conditional proportions:

$$p_{ij(k)} = p_{ij|k} = \frac{n_{ijk}}{n_{++k}} \qquad k = 1, ..., K \tag{3.3}$$

3.1.1.3 Marginal Tables

- *Marginal* tables involve summing over all possible categories of a particular variable.
- We denote such summation using a + (as before).
- For example, the XY- marginal table is $(n_{ij+}) = (\sum_k n_{ijk})$.
- XZ and YZ- marginal tables are denoted as (n_{i+k}) and (n_{+jk}) respectively.
- Marginal probabilities:

$$\pi_{ij} = P(X = i, Y = j) = \pi_{ij+} = \sum_{k=1}^{K} \pi_{ijk} \tag{3.4}$$

- Marginal proportions:

$$p_{ij} = p_{ij+} = \sum_{k=1}^{K} p_{ijk} \tag{3.5}$$

3.1.1.4 Marginal Vectors

- Information on the single classification variables is summarised in the marginal vectors $(n_{1++}, ..., n_{I++})$, $(n_{+1+}, ..., n_{+J+})$ and $(n_{++1}, ..., n_{++K})$ respectively.

3.1.2 Generic Multiway Tables

- A multiway $I_1 \times I_2 \times ... \times I_q$ contingency table for variables $X_1, X_2, ..., X_q$ will analogously be denoted as $(n_{i_1 i_2 ... i_q})$, $i_l = 1, ..., I_l$, $l = 1, ..., q$.
- The definition of partial and marginal tables also follow analogously.
- For example, $(n_{i_1+(i_3)i_4(i_5)})$ denotes the two-way partial marginal table obtained by summing over all levels/categories i_2 of X_2 for a fixed level/category of (or conditioning on) variables $X_3 = i_3$ and $X_5 = i_5$.

3.2 Odds Ratios

- Conditional and marginal odds ratios can be defined for any two-way conditional or marginal probabilities table of a multi-way $I_1 \times I_2 \times ... \times I_q$ table with $I_l \geq 2$, $l = 1, ..., q$.

- In this case, the conditional and marginal odds ratios are defined as odds ratios for two-way tables of size $I \times J$.
- Thus, as defined for general two-way tables in Sections 2.5 and 2.6.2, there will be a (not unique) minimal set of odds ratios of nominal, local, cumulative, or global type.
- For example, for an $I \times J \times K$ table, the XY local odds ratios conditional on Z are defined by

$$r_{ij(k)}^{L,XY} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i+1,j,k}\pi_{i,j+1,k}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \quad k = 1, ..., K \tag{3.6}$$

  and the XY-marginal local odds ratios are defined by

$$r_{ij}^{L,XY} = \frac{\pi_{ij+}\pi_{i+1,j+1,+}}{\pi_{i+1,j,+}\pi_{i,j+1,+}} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{3.7}$$

- The conditional and marginal odds ratios of other types, like nominal, cumulative and global, are defined analogously.

3.3 Types of Independence

- Let (n_{ijk}) be an $I \times J \times K$ contingency table of observed frequencies with row, column and layer classification variables X, Y and Z respectively.
- We consider various types of independence that could exist among these three variables.

3.3.1 Mutual Independence

- X, Y and Z are *mutually* independent if and only if

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \qquad i = 1, ..., I \quad j = 1, ..., J \quad k = 1, ..., K \tag{3.8}$$

- Such mutual independence can be symbolised as $[X, Y, Z]$.

3.3.1.1 Example

- Following the example of Section 3.1.1.1, mutual independence would mean that clinic, drug and response were independent of each other.
- In other words, knowledge of the values of one variable doesn't affect the probabilities of the levels of the others.

3.3.2 Joint Independence

- If Y is *jointly* independent from X and Z (without these two being necessarily independent), then

$$\pi_{ijk} = \pi_{+j+}\pi_{i+k} \qquad i = 1, ..., I \quad j = 1, ..., J \quad k = 1, ..., K \tag{3.9}$$

- Such joint independence can be symbolised as $[Y, XZ]$.
- By symmetry, there are two more hypotheses of this type, which can be expressed in a symmetric way to Equation (3.9) for X or Z being jointly independent from the remaining two variables. These could be symbolised as $[X, YZ]$ and $[Z, XY]$ respectively.

3.3.2.1 Example

If $[Z, XY]$, then the clinic is independent of the drug and the response. In other words, the response of a subject to treatment may depend on the drug they received, but neither of these are associated with the clinic that they went to.

3.3.3 Marginal Independence

- X and Y are *marginally* independent (ignoring Z) if and only if

$$\pi_{ij+} = \pi_{i++}\pi_{+j+} \qquad i = 1, ..., I \quad j = 1, ..., J \quad k = 1, ..., K \tag{3.10}$$

- Here, we actually ignore Z.
- Such marginal independence is symbolised $[X, Y]$.

3.3.3.1 Example

- If Y and Z are marginally independent[1] (that is $[Y, Z]$), then this would imply that response to treatment is not associated with the clinic attended if we ignore which drug was received.

3.3.4 Conditional Independence

- Under a multinomial sampling scheme, the joint probabilities of the three-way table cells π_{ijk} can be expressed in terms of conditional probabilities as

$$\begin{aligned} \pi_{ijk} &= P(X = i, Y = j, Z = k) && (3.11)\\ &= P(Y = j|X = i, Z = k)\, P(X = i, Z = k) && (3.12)\\ &= \pi_{j|ik}\pi_{i+k} && (3.13) \end{aligned}$$

- X and Y are *conditionally* independent given Z if

$$\pi_{ij|k} = \pi_{i|k}\pi_{j|k} \qquad k = 1, ..., K \tag{3.14}$$

- We can consequently show that

$$\pi_{j|ik} = \pi_{j|k} \tag{3.15}$$

[1] ignoring X

and therefore that

$$\begin{aligned}\pi_{ijk} = \pi_{j|k}\pi_{i+k} &= P(Y=j|Z=k)P(X=i,Z=k) \\ &= P(X=i,Z=k)\frac{P(Y=j,Z=k)}{P(Z=k)} \\ &= \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \qquad (3.16) \\ &\qquad i=1,...,I \quad j=1,...,J \quad k=1,...,K\end{aligned}$$

- Note that we here assumed that Y was the response variable. The conditioning approach with $X = i$ as response variable would also lead to Equation (3.16), which is symmetric in terms of X and Y.
- This conditional independence of X and Y given Z can be symbolised as $[XZ, YZ]$.
- The hypotheses of conditional independence $[XY, YZ]$ and $[XY, XZ]$ are formed analogously to Equation (3.16).

3.3.4.1 Example

If Y and Z are conditionally independent given X (that is, $[XY, XZ]$), this implies that response to treatment is independent of clinic attended given knowledge of which drug was received.

3.3.4.2 Odds Ratios

- Under conditional independence of X and Y given Z ([XZ,YZ]), the XZ odds ratios conditional on Y are equal to the XZ marginal odds ratios, that is[2]

$$r^{XZ}_{i(j)k} = r^{XZ}_{ik} \qquad i=1,...,I-1 \quad j=1,...,J \quad k=1,...,K-1 \qquad (3.17)$$

  In other words, the marginal and conditional XZ associations coincide.

- By symmetry, we also have that

$$r^{YZ}_{(i)jk} = r^{YZ}_{jk} \qquad i=1,...,I \quad j=1,...,J-1 \quad k=1,...,K-1 \qquad (3.18)$$

  that is, the marginal and conditional YZ associations coincide.

- However, the XY marginal and conditional associations do not coincide, that is:

$$r^{XY}_{ij(k)} \neq r^{XY}_{ij} \qquad (3.19)$$

  in general.

- Such arguments for $[XY, YZ]$ and $[XY, XZ]$ are analogous.

3.3.5 Conditional and Marginal Independence

Important: Conditional independence does not imply marginal independence, and marginal independence does not imply conditional independence.

[2]Note that we don't superscript L or G here, as the result holds for both. Q3-2 involves showing that Equation (3.17) holds for local odds ratios.

3.3.5.1 Example

3.3.5.1.1 Marginal but not Conditional Independence

- Suppose response Y and clinic Z are marginally independent (ignoring treatment drug X). However, there may be a conditional association between response to treatment Y and clinic attended Z on the drug received X.
- Example potential explanation[3]: some clinics may be better prepared to care for subjects on some treatment drugs than others, but without knowledge of the treatment drug received, neither clinic is more associated with a successful response.

3.3.5.1.2 Conditional but not Marginal Independence

- Suppose Y and Z are conditionally independent given X (that is, $[XY, XZ]$), then this implies that response to treatment is independent of clinic attended given knowledge of which drug was received. However, there may be a marginal association between response to treatment Y and clinic attended Z if we ignore which treatment drug X was received.
- Example potential explanation: Given knowledge of the treatment drug, it does not matter which clinic the subject attends. However, without knowledge of the treatment drug, one clinic may be more associated with a successful response (perhaps because their stock of the more successful drug is greater...).

3.3.6 Homogeneous Associations

- *Homogeneous associations* (also known as *no three-factor interactions*) mean that the conditional relationship between any pair of variables given the third one is the same at each level of the third variable; but not necessarily independent.
- This relation implies that if we know all two-way tables between the three variables, we have sufficient information to compute (π_{ijk}).
- However, there are no separable closed-form estimates for the expected joint probabilities $(\hat{\pi}_{ijk})$, hence maximum likelihood estimates must be computed by an iterative procedure such as Iterative Proportional Fitting or Newton-Raphson.
- Such homogeneous associations are symbolised $[XY, XZ, YZ]$.

3.3.6.1 Odds Ratios

- Homogeneous associations can be thought of in terms of conditional odds ratios as follows:
  - the XY partial odds ratios at each level of Z are identical: $r^{XY}_{ij(k)} = r^{XY,\star}_{ij}$
  - the XZ partial odds ratios at each level of Y are identical: $r^{XZ}_{i(j)k} = r^{XZ,\star}_{ik}$

[3]Note that this is precisely what this is - a potential explanation - it would be incorrect to conclude that this is definitely the reason for the hypothesised independence scenarios. We all know (I hope...) that *correlation* (or in this case, *association*) *does not imply causation.*

  - the YZ partial odds ratios at each level of X are identical: $r^{YZ}_{(i)jk} = r^{YZ,\star}_{jk}$

- Note that $r^{XY,\star}_{ij}, r^{XZ,\star}_{ik}, r^{YZ,\star}_{jk}$ are not necessarily the same as the corresponding marginal odds ratios $r^{XY}_{ij}, r^{XZ}_{ik}, r^{YZ}_{jk}$.

3.3.6.2 Example

The treatment response and treatment drug have the same association for each clinic.

More precisely, we have

$$r^{XY}_{A,S,(k)} = r^{XY,\star}_{A,S} \iff \frac{\pi_{A,S,(k)}}{\pi_{A,F,(k)}} = r^{XY,\star}_{A,S}\frac{\pi_{B,S,(k)}}{\pi_{B,F,(k)}} \qquad k = 1, 2 \tag{3.20}$$

which means that each drug has a different odds of success depending on the clinic, however, the odds of treatment success of drug A are a fixed constant $r^{XY,\star}_{A,S}$ greater than the odds of treatment success of drug B, regardless of the clinic.

3.3.7 Tests for Independence

- Marginal independence (Equation (3.10)) can be tested using the test for independence presented in Section 2.4.3.1 applied on the corresponding two-way marginal table.
- Hypotheses of the independence statements defined by Equations (3.8), (3.9) and (3.16) could be tested analogously using the relevant marginal counts.
- We do not consider these tests, but defer to log-linear models (soon!).
- A specific test of independence of XY at each level of Z for $2 \times 2 \times K$ tables is presented in Section 3.3.10.

3.3.8 Summary of Relationships

We present a summary of which independence relationships can be implied from which others, and which can't, in Figure 3.3.

3.3.9 Multi-way Tables

- Analogous definitions of the various types of independence exist for general multi-way tables.

3.3.9.1 Example

We might analyse $(n_{i_1+(i_3)i_4(i_5)})$ across the different levels of (I_3, I_5) to see if (I_1, I_4) and (I_3, I_5) are marginally independent (ignoring I_2). We will explore this a bit, but in general, we look to log-linear models.

3.3.10 Mantel-Haenszel Test for $2 \times 2 \times K$ Tables

- We will discuss the particular case of X and Y being binary variables that are cross-classified across the K layers of a variable Z, forming K 2×2 partial tables $n_{ij(k)}$, $k = 1, ..., K$.

$$
\begin{aligned}
\text{Mutual} &\implies \text{Joint} \\
\\
\text{Joint} &\implies \text{Marginal} \\
[XY, Z] &\implies [X, Z], [Y, Z] \\
\\
\text{Joint} &\implies \text{Conditional} \\
[XY, Z] &\implies [XY, YZ], [XY, XZ] \\
\\
\text{Marginal} &\not\implies \text{Joint} \\
\text{Conditional} &\not\implies \text{Joint} \\
\text{Marginal} &\not\implies \text{Conditional} \\
\text{Conditional} &\not\implies \text{Marginal}
\end{aligned}
$$

Figure 3.3: Summary of relationships between independencies.

- The Mantel-Haenszel Test is for testing the conditional independence of X and Y given Z for these $2 \times 2 \times K$ tables, that is, it considers the hypotheses

$$\mathcal{H}_0 : \quad X, Y \text{are independent conditional on the level of } Z. \tag{3.21}$$
$$\mathcal{H}_1 : \quad X, Y \text{are not independent conditional on the level of } Z. \tag{3.22}$$

  or in other words[4]

$$\mathcal{H}_0 : \quad r_{12(k)} = 1, \text{ for all } k = 1, ..., K \tag{3.23}$$
$$\mathcal{H}_1 : \quad r_{12(k)} \neq 1, \text{ for some } k = 1, ..., K \tag{3.24}$$
$$\tag{3.25}$$

- The Mantel-Haenszel Test conditions on the row and column marginals of each of the K partial tables.

- Under $\mathcal{H}_0$, every partial table has that n_{11k} follows a hypergeometric distribution[5]

[4]Note that we revert back to the r_{12} notation here since each of the K layers is a 2×2 table.
[5]See Section 1.4.4.

$\mathcal{H}g(N = n_{++k}, M = n_{1,+,k}, q = n_{+,1,k})$[6], and thus has mean and variance

$$\hat{E}_{11k} = \frac{n_{1+k}n_{+1k}}{n_{++k}} \qquad \hat{\sigma}^2_{11k} = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n^2_{++k}(n_{++k} - 1)}$$

- $\sum_k n_{11k}$ therefore has mean $\sum_k \hat{E}_{11k}$ and variance $\sum_k \hat{\sigma}^2_{11k}$, since the values of n_{11k} are independent of each other (having conditioned on $Z = k$).
- The Mantel–Haenszel test statistic is defined as[7]

$$T_{MH} = \frac{[\sum_k (n_{11k} - \hat{E}_{11k})]^2}{\sum_k \hat{\sigma}^2_{11k}} \tag{3.26}$$

- T_{MH} is asymptotically χ^2_1 under $\mathcal{H}_0$.
- If $T_{MH(obs)}$ is the observed value of the test statistic for a particular case, then the p-value is $P(\chi^2_1 > T_{MH(obs)})$.
- When the XY association is similar across the partial tables, then the test is more powerful.
- It loses in power when the underlying associations vary across the layers, especially when they are of different direction, since the differences $n_{11k} - \hat{E}_{11k}$ will then cancel out in the sum of the statistic given by Equation (3.26).

[6]Why hypergeometric? Well, for any 2×2 table we have an assumed total of $N = n_{++k}$ items. We condition on row and column margins, so we assume knowledge of $n_{i,+,k}$ and $n_{+,j,k}$. In that population, we know that $M = n_{1,+,k}$ of these items are such that $i = 1$. If the two variables X and Y are conditionally independent given Z, then we could view $N_{1,1,k}$ to be the result of picking $q = n_{+,1,k}$ items (those going into column 1) randomly from $N = n_{++k}$, and calculating how many of those are from row 1 (given that we know that there are $M = n_{1,+,k}$ items out of the N that will go into row 1 in total). Therefore $N_{1,1,k} \sim \mathcal{H}g(N = n_{++k}, M = n_{1,+,k}, q = n_{+,1,k})$

[7]Note that the square is outside of the summation.

Chapter 4

Log-Linear Models

- We have dealt with inter-relationships among and between several categorical variables in Sections 2 and 3.
- *Log-Linear Models (LLMs)* describe the way the involved categorical variables and their association (if appropriate/significant) influence the count in each of the cells of the cross-classification table of these variables.
- They are appropriate when there is no clear distinction between response and explanatory variables (this is in contrast to GLMs (later...)).
- So, how precisely are they formed...?

4.1 LLMs for Two-Way Tables

4.1.1 Model of Independence

- Independence between classification variables X and Y can equivalently be expressed in terms of the expected-under-independence cell frequencies E_{ij} in a log-linear model form as

$$\log E_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \qquad \forall i, j \tag{4.1}$$

  where
  - λ_i^X corresponds to the effect of the i^{th} level of X, and
  - λ_j^Y corresponds to the effect of the j^{th} level of Y
  - What about λ? Well that depends on the choice of constraints used (Section 4.1.1.4). The precise interpretation of λ_i^X and λ_j^Y also depends on the choice of constraints used.

4.1.1.1 Interpretation via Odds

- Interpretation is carried out in terms of odds. For given column category j, under the model given by Equation (4.1), the odds of being in row i_1 instead of row i_2, for

$i_1, i_2 \in \{1, ..., I\}, i_1 \neq i_2$, are

$$\frac{E_{i_1 j}}{E_{i_2 j}} = \frac{\exp(\lambda + \lambda^X_{i_1} + \lambda^Y_j)}{\exp(\lambda + \lambda^X_{i_2} + \lambda^Y_j)} = \exp(\lambda^X_{i_1} - \lambda^X_{i_2}) \qquad \forall j \tag{4.2}$$

which shows that being in row i_1 instead of i_2 is determined only by the distance of the corresponding row main effect values, and hence independent of j. Note that Equation (4.2) also implies that

$$\frac{P(X = i_1 | Y = j)}{P(X = i_2 | Y = j)} = \exp(\lambda^X_{i_1} - \lambda^X_{i_2}) \qquad \forall j \tag{4.3}$$

- Similarly for columns j_1 and j_2 such that $j_1, j_2 \in \{1, ..., J\}, j_1 \neq j_2$, we have

$$\frac{E_{ij_1}}{E_{ij_2}} = \exp(\lambda^Y_{j_1} - \lambda^Y_{j_2}) \qquad \forall i \tag{4.4}$$

hence independent of i.

4.1.1.2 Odds Ratios

- Local odds ratios in this case are given by

$$r^L_{ij} = \frac{E_{ij}/E_{i+1,j}}{E_{i,j+1}/E_{i+1,j+1}} = \frac{\exp(\lambda^X_i - \lambda^X_{i+1})}{\exp(\lambda^X_i - \lambda^X_{i+1})} = 1 \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \tag{4.5}$$

as expected.

4.1.1.3 Non-Identifiability Issues

- We observe that the model given by Equation (4.1) has $1 + I + J$ parameters to be learned.
- However, under a Poisson sampling scheme, there should be $1 + (I-1) + (J-1)$ free parameters, because
  - n_{++} is not fixed, hence we have to learn it (1 parameter),
  - the set of marginal sums n_{i+} must themselves sum to n_{++}, hence $(I-1)$ parameters,
  - the set of marginal sums n_{+j} must themselves sum to n_{++}, hence $(J-1)$ parameters.

4.1.1.4 Identifiability Constraints

- To fix the non-identifiability issues, we impose a number of constraints equal to the difference between the number of free parameters of the unconstrained model and the number it should have.
- For the two-way independence model, this is:

$$(1 + I + J) - (1 + (I-1) + (J-1)) = 2$$

- We discuss two popular choices.

4.1.1.4.1 Zero-Sum Constraints

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0 \tag{4.6}$$

Then λ_i^X and λ_j^Y account for deviations that X and Y have from the overall mean (λ).

4.1.1.4.2 Corner Point Constraints

$$\lambda_I^X = \lambda_J^Y = 0 \tag{4.7}$$

where I and J are called *reference* levels for variables X, Y.

- Then λ_i^X, $i = 1, ..., I-1$ and λ_j^Y, $j = 1, ..., J-1$ account for deviations from the reference levels I and J.
- Of course, any other level could serve as reference levels instead. For example,

$$\lambda_1^X = \lambda_1^Y = 0 \tag{4.8}$$

  is a common choice too.

4.1.1.4.3 Interpretations of λ

- Different identifiability constraints lead to different interpretation for the λ's, but to the same inference.

4.1.1.4.3.1 Example For a 2×2 contingency table, we have

- with zero-sum constraints ($\lambda_1^X + \lambda_2^X = \lambda_1^Y + \lambda_2^Y = 0$), that

$$\frac{\pi_{2|i}}{\pi_{1|i}} = \exp(-2\lambda_1^Y) \tag{4.9}$$

- or with corner point constraints ($\lambda_2^X = \lambda_2^Y = 0$), that

$$\frac{\pi_{2|i}}{\pi_{1|i}} = \exp(-\lambda_1^Y) \tag{4.10}$$

4.1.2 The Saturated Model

If X, Y are not independent, then we add an XY-interaction term to the LLM.

$$\log E_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \qquad \forall i, j \tag{4.11}$$

4.1.2.1 Log Odds Ratio

The local odds ratios are directly derived from the interaction parameters, since

$$\log r_{ij}^L = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i+1,j}^{XY} - \lambda_{i,j+1}^{XY} \qquad i = 1, ..., I-1,\ j = 1, ..., J-1 \tag{4.12}$$

4.1.2.2 Constraints

4.1.2.2.1 Zero-Sum Constraints

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = \sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0 \tag{4.13}$$

4.1.2.2.2 Corner Point Constraints With reference categories I, J, these are

$$\lambda_I^X = \lambda_J^Y = \lambda_{Ij}^X = \lambda_{iJ}^Y = 0 \qquad \forall i, j \tag{4.14}$$

4.1.2.3 Free Parameters

The number of free parameters (under Poisson sampling) is thus

$$d = 1 + (I-1) + (J-1) + (I-1)(J-1) = IJ \tag{4.15}$$

as it should, corresponding to the number of cells in the table.

Remark: Under multinomial sampling, since n_{++} is fixed, there should be $IJ - 1$ parameters. This is fixed by removing λ as a parameter.

4.1.3 Interpretation of λ

From Equation (4.11), assuming the zero-sum constraints as given by Equation (4.13), we have that[1]

$$\lambda = \frac{1}{IJ} \sum_{i,j} \log E_{ij} \tag{4.16}$$

$$\lambda_i^X = \frac{1}{J} \sum_{j} \log E_{ij} - \lambda \qquad i = 1, ..., I \tag{4.17}$$

$$\lambda_j^Y = \frac{1}{I} \sum_{i} \log E_{ij} - \lambda \qquad j = 1, ..., J \tag{4.18}$$

$$\lambda_{ij}^{XY} = \log E_{ij} - \lambda - \lambda_i^X - \lambda_j^Y \qquad i = 1, ..., I, j = 1, ..., J \tag{4.19}$$

Hence,

- λ represents the mean in terms of being the mean log expected count value across all cells in the table.
- λ_i^X represents the effect of the i^{th} level of X as the difference between the average log expected count value across the levels of Y conditional on $X = i$, and the overall average log expected count value across all cells in the table[2].
- λ_{ij}^{XY} represents the joint effect of the i^{th} level of X and the j^{th} level of Y as the difference between the log expected cell count for cell (i, j), and the sum of the marginal contributions attributed to λ, λ_i^X and λ_j^Y discussed above.

[1] Q4-1a involves showing that these expressions hold.
[2] λ_j^Y can be interpreted analogously.

4.1.4 Inference and Fit

- The likelihood function ($Poi(E_{ij})$-distributed for each cell (i, j)):

$$L = \prod_{i,j} \frac{e^{-E_{ij}} E_{ij}^{n_{ij}}}{n_{ij}!} \tag{4.20}$$

so that

$$\begin{aligned} l &= \sum_{i,j} (-E_{ij} + n_{ij} \log E_{ij} - \log n_{ij}!) && (4.21) \\ &\propto \sum_{i,j} (n_{ij} \log E_{ij} - e^{\log E_{ij}}) && (4.22) \end{aligned}$$

as an expression of $\log E_{ij}$.

- Substituting $\log E_{ij}$ using the relevant expression (e.g. from Equation (4.1) for the independence model or Equation (4.11) for the saturated model) will make this an expression of the LLM λ parameters.
- We now proceed for the independence model specifically.

4.1.4.1 Inference and Fit for Independence Model under Zero-Sum Constraints

- Given model $[X, Y]$, we have

$$\begin{aligned} \pi_{ij} &= \pi_{i+}\pi_{+j} \\ \implies E_{ij} &= \frac{E_{i+}E_{+j}}{n_{++}} \end{aligned} \tag{4.23}$$

since $E_{ij} = n_{++}\pi_{ij}$, $E_{i+} = n_{++}\pi_{i+}$, etc.

- Hence we have that

$$\hat{E}_{ij} = \frac{\hat{E}_{i+}\hat{E}_{+j}}{n_{++}} \tag{4.24}$$

where $\hat{E}_{ij} = n_{++}\hat{\pi}_{ij}$, $\hat{E}_{i+} = n_{++}\hat{\pi}_{i+}$ etc.

- In Section 2.4.3.2, we used the method of Lagrange multipliers to find MLEs $\hat{\pi}_{i+}$ and $\hat{\pi}_{+j}$ using appropriate independence constraints. These can then be used to yield $\hat{E}_{i+}$ and $\hat{E}_{+j}$.
  Alternatively, however, we could also use the method of Lagrange multipliers for λ on the log-likelihood of the LLM expression (such as given by Equation (4.22)) to elicit MLEs $\hat{E}_{i+} \equiv E_{i+}(\hat{\boldsymbol{\lambda}})$ and $\hat{E}_{+j} \equiv E_{+j}(\hat{\boldsymbol{\lambda}})$[3].
  Either way, we find that

$$\hat{E}_{i+} = n_{i+} \qquad \hat{E}_{+j} = n_{+j} \tag{4.25}$$

so that

$$\hat{E}_{ij} = \frac{\hat{E}_{i+}\hat{E}_{+j}}{n_{++}} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{4.26}$$

[3]Using Lagrange multipliers in this way is addressed as part of Q4-2.

- We can then substitute these estimates $\hat{E}_{ij}$ into Equations (4.16) - (4.18)[4] to obtain MLEs for the LLM parameters as given by:[5] [6]

$$\hat{\lambda} = \frac{1}{I}\sum_s \log n_{s+} + \frac{1}{J}\sum_t \log n_{+t} - \log n_{++} \tag{4.27}$$

$$\hat{\lambda}_{i+} = \log n_{i+} - \frac{1}{I}\sum_s \log n_{s+} \qquad i = 1, ..., I \tag{4.28}$$

$$\hat{\lambda}_{+j} = \log n_{+j} - \frac{1}{J}\sum_t \log n_{+t} \qquad j = 1, ..., J \tag{4.29}$$

4.2 LLMs for Three-Way Tables

- Consider a three-way contingency table, cross-classifying the variables X, Y and Z.

4.2.1 Mutual Independence

- Denoted $[X, Y, Z]$, this is given as

$$\log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \qquad \forall i, j, k \tag{4.30}$$

with either zero-sum constraints:

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = \sum_{k=1}^{K} \lambda_k^Z = 0 \tag{4.31}$$

or corner point constraints:

$$\lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0 \tag{4.32}$$

so that the number of free parameters is

$$d = 1 + (I-1) + (J-1) + (K-1) = I + J + K - 2 \tag{4.33}$$

4.2.2 Joint Independence

- Joint independence of Y from X and Z (denoted $[Y, XZ]$) corresponds to the following LLM:

$$\log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} \qquad \forall i, j, k \tag{4.34}$$

either satisfying the zero-sum constraints given by Equation (4.31) in addition to

$$\sum_{i=1}^{I} \lambda_{ik}^{XZ} = \sum_{k=1}^{K} \lambda_{ik}^{XZ} = 0 \tag{4.35}$$

for all possible values of the non-marginal subscript, or satisfying the cornerpoint constraints given by Equation (4.32) in addition to

$$\lambda_{Ik}^{XZ} = \lambda_{iK}^{XZ} = 0 \tag{4.36}$$

for all possible values of the non-conditional subscript.

[4] Note that these three hold even for the independence model, it's just we don't have Equation (4.19) in this case.

[5] Note that we sum over s, t to distinguish from the λ LLM parameters corresponding to each i, j.

[6] Q4-1b involves showing that these equations hold given Equation (4.26) .

- As a result, the number of free parameters for this model is given by

$$d = 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) = IK + J - 1 \tag{4.37}$$

4.2.3 Further LLMs

Hopefully you can see that LLMs corresponding to the types of independence discussed in Sections 3.3.1 to 3.3.6 are also possible, as well as a three-way saturated model for the situation where all three parameters are all jointly associated with each other.

4.2.3.1 Homogeneous Associations Model

The homogeneous associations model is

$$\log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \qquad \forall i, j, k \tag{4.38}$$

with zero-sum constraints[7] given by

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = \sum_{k=1}^{K} \lambda_k^Z = \sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = \sum_{i=1}^{I} \lambda_{ik}^{XZ} = \sum_{k=1}^{K} \lambda_{ik}^{XZ} = \sum_{j=1}^{J} \lambda_{jk}^{YZ} = \sum_{k=1}^{K} \lambda_{jk}^{YZ} = 0 \tag{4.39}$$

with the two-factor constraints holding for all values of the non-marginalised subscript.

4.2.3.2 The Saturated Model

The saturated model is

$$\log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \qquad \forall i, j, k \tag{4.40}$$

with zero-sum constraints[8] as given by Equation (4.39) in addition to

$$\sum_{i=1}^{I} \lambda_{ijk}^{XYZ} = \sum_{j=1}^{J} \lambda_{ijk}^{XYZ} = \sum_{k=1}^{K} \lambda_{ijk}^{XYZ} = 0 \tag{4.41}$$

each holding for all pairs of values of the non-marginalised subscripts, resulting in

$$\begin{aligned} d &= 1 + (I-1) + (J-1) + (K-1) && (4.42)\\ &\quad + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) && (4.43)\\ &\quad + (I-1)(J-1)(K-1) = IJK && (4.44) \end{aligned}$$

free parameters.

4.3 Hierarchical LLMs for Multiway Tables

- LLMs can be defined analogously for multiway tables of $q > 3$ categorical variables.

[7] Corner point constraints could also be used.
[8] Corner point constraints could also be used.

- LLMs for multiway tables include higher-order interactions, up to order q.
- The number of possible models increases with the dimension of the table, involving the procedure of deciding which one is appropriate for describing the underlying structure of association.
- In order to impose a structure on model building, especially helpful in model selection, log-linear modelling is usually restricted to the family of Hierarchical Log-Linear Models (HLLMs).
- *Hierarchical* here means that if an interaction term of b variables is included in the model, then all the marginal terms, corresponding to lower-level interactions of any subset of these b variables, will also be included.

4.3.1 Example: $[XYW, ZW]$

- In a four-way table cross-classifying variables X, Y, Z and W, how could we understand and explain that variable X does not interact with Y (absence of the λ_{XY} term from the model) but it interacts simultaneously with Y and W (model includes the λ_{ijl}^{XYW} term)?
- Therefore, if λ_{ijl}^{XYW} is included, so too must λ_{ij}^{XY}, λ_{il}^{XW} and λ_{jl}^{YW}, and by extension or directly, λ_i^X, λ_j^Y and λ_l^W must be included as well.
- $[XYW, ZW]$ therefore implies the following model

$$\log E_{ijkl} = \lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_l^W+\lambda_{ij}^{XY}+\lambda_{il}^{XW}+\lambda_{jl}^{YW}+\lambda_{kl}^{ZW}+\lambda_{ijl}^{XYW} \qquad \forall i,j,k,l \quad (4.45)$$

- This model implies that XY are jointly independent of Z conditional on W, since the two-factor interaction terms XZ and YZ are missing. Another way to see this is that no term involves both Z and 'X or Y', so if W is fixed at level l, then Z and XY affect $\log E_{ijkl}$ independently. Note, however, that the model structure does not imply that Z and XY are jointly independent if we marginalise over W.

4.3.2 Example: $[XYW, Z]$

- If the model of interest is notated $[XYW, Z]$ (that is, XYW are jointly independent of Z), then the model is given by

$$\log E_{ijkl} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{il}^{XW} + \lambda_{jl}^{YW} + \lambda_{ijl}^{XYW} \qquad \forall i,j,k,l \quad (4.46)$$

4.3.3 Example: $[X_1X_2, X_1X_5, X_3X_4X_5, X_5X_6X_7]$

- $[X_1X_2, X_1X_5, X_3X_4X_5, X_5X_6X_7]$ implies the following model

$$\begin{aligned} \log E_{i_1\ldots i_7} \;=\; & \lambda + \sum_{q=1}^{7} \lambda_{i_q}^{X_q} \\ & +\lambda_{i_1i_2}^{X_1X_2} + \lambda_{i_1i_5}^{X_1X_5} + \lambda_{i_3i_4}^{X_3X_4} + \lambda_{i_3i_5}^{X_3X_5} + \lambda_{i_4i_5}^{X_4X_5} + \lambda_{i_5i_6}^{X_5X_6} + \lambda_{i_5i_7}^{X_5X_7} + \lambda_{i_6i_7}^{X_6X_7} \\ & +\lambda_{i_3i_4i_5}^{X_3X_4X_5} + \lambda_{i_5i_6i_7}^{X_5X_6X_7} \qquad \forall i_1,i_2,i_3,i_4,i_5,i_6,i_7 \end{aligned} \quad (4.47)$$

- We can see that variable pairs (X_1, X_2), (X_3, X_4) and (X_6, X_7) are mutually independent conditional on X_5.

4.3.4 Notation

- Notation can get ridiculous here...
- We can also write the model (4.47) as

$$\log E_{i_1 \ldots i_7} = \lambda + \sum_{\mathcal{I} \in \mathcal{J}} \lambda_{\mathcal{I}}^{\mathcal{X}} \tag{4.48}$$

where $\mathcal{I}$ corresponds to the levels of a combination of variables $\mathcal{X}$ included in a particular model term, and $\mathcal{J}$ corresponds to the set of those combinations which make up the terms of the whole model.

- Using this notation, for the model given by Equation (4.47), we have

$$\begin{aligned} \mathcal{J} = \{\mathcal{I}\} \;=\; & \{i_1, i_2, i_3, i_4, i_5, i_6, i_7, \\ & (i_1, i_2), (i_1, i_5), (i_3, i_4), (i_3, i_5), (i_4, i_5), (i_5, i_6), (i_5, i_7), (i_6, i_7), \\ & (i_3, i_4, i_5), (i_5, i_6, i_7)\} \end{aligned} \tag{4.49}$$

corresponding to the sets of variables

$$\begin{aligned} \{\mathcal{X}\} \;=\; & \{X_1, X_2, X_3, X_4, X_5, X_6, X_7, \\ & (X_1, X_2), (X_1, X_5), (X_3, X_4), (X_3, X_5), (X_4, X_5), (X_5, X_6), (X_5, X_7), (X_6, X_7), \\ & (X_3, X_4, X_5), (X_5, X_6, X_7)\} \end{aligned} \tag{4.50}$$

4.4 MLE for LLMs

This is the general five-step process of obtaining MLEs for the LLM λ parameters. Some of these involve quite some work in themselves.[9]

1. Write down log-likelihood expression.

- Similar to Section 4.1.4 for the two-way LLM, assuming a $Poi(E_{i_1 \ldots i_q})$ distribution for each cell $(i_1, ..., i_q)$, the general log-likelihood is

$$l(\boldsymbol{\lambda}) \propto \sum_{i_1, \ldots, i_q} n_{i_1 \ldots i_q} \log(E_{i_1 \ldots i_q}) - E_{i_1 \ldots i_q} = \sum_{i_1, \ldots, i_q} n_{i_1 \ldots i_q} \log(E_{i_1 \ldots i_q}) - e^{\log(E_{i_1 \ldots i_q})} \tag{4.51}$$

where $E_{i_1 \ldots i_q} \equiv E_{i_1 \ldots i_q}(\boldsymbol{\lambda})$ can be viewed as a function of $\boldsymbol{\lambda}$ which is defined by the LLM expression of interest.

2. $E_{i_1 \ldots i_q}$ only depend on $E_{\mathcal{I}}$.

[9] Note that some of these steps can be done in a different order - the aim here is to create an outline for the whole process in case it is helpful for you. Q4-2 involves going through these steps from start to finish for the two-way independence model assuming corner-point constraints. This will also be a problems class question at some point.

- We can use knowledge of the model format/dependency structure to express $E_{i_1 \ldots i_q}$ in terms of $E_{\mathcal{I}}$. In other words, we only require $E_{\mathcal{I}}$ corresponding to $\mathcal{I} \in \mathcal{J}$, from which $E_{i_1 \ldots i_q}$, $\forall i_1, ..., i_q$, can be obtained given the dependency structure corresponding to $\mathcal{J}$. It may be possible to express this dependency in an analytic form (such as Equation (4.23) for model $[X, Y]$, or Equation (4.54) below for model $[Y, XZ]$), or it may be that a tractable expression is not available, in which case numerical methods such as Newton-Raphson or Iterative Proportional Fitting must be used, as discussed in Section 3.3.6[10]). See Section 4.4.1 for examples.

3. Use Lagrange multipliers to elicit required $\hat{E}_{\mathcal{I}} \equiv E_{\mathcal{I}}(\hat{\lambda})$-terms.

- We do this either

  a. through first eliciting required $\hat{\pi}$-terms, or

  b. for $\hat{\lambda}$ on the log-likelihood of LLM expression, with corresponding constraints.

Either way, we find that

$$\hat{E}_{\mathcal{I}} = n_{\mathcal{I}} \tag{4.52}$$

for all $\mathcal{I} \in \mathcal{J}$, where $n_{\mathcal{I}}$ corresponds to the marginal sums over all levels of the variables not represented in the set $\mathcal{I}$.

4. Obtain MLEs for $\hat{E}_{i_1 \ldots i_q} \equiv E_{i_1 \ldots i_q}(\hat{\boldsymbol{\lambda}})$ terms.

- We do this by either using the expression from (2.) if obtained, or by appropriate numerical approximation method[11] if no analytical expression exists.
- Either way, MLEs for the remaining $\hat{E}_{i_1,\ldots,i_q} \equiv E_{i_1,\ldots,i_q}(\hat{\boldsymbol{\lambda}})$ terms, $\forall i_1, ..., i_q$, only depend on (that is, can be obtained from) the $\hat{E}_{\mathcal{I}}$ terms corresponding to those $\mathcal{I} \in \mathcal{J}$ included in the model (see Section 4.4.1 for examples).

5. Substitute these estimates $\hat{E}_{i_1 \ldots i_q}$ into expressions for $\boldsymbol{\lambda}$

- The $\hat{\lambda}$ LLM parameters can then be found by:
  - first rearranging the equations corresponding to the LLM expressions in terms of the λ parameters (such as for the two-way saturated model in Section 4.1.3); and
  - then substituting the relevant estimates $\hat{E}_{i_1 \ldots i_q}$ into these expressions for $\boldsymbol{\lambda}$ (such as for the two-way independence model in Section 4.1.4.1).

4.4.1 Examples of Eliciting $\hat{E}_{i_1,\ldots,i_q}$

4.4.1.1 $[Y, XZ]$

- E_{ijk} only depend on E_{+j+} and E_{i+k} as follows

$$E_{ijk} = \frac{E_{+j+} E_{i+k}}{n_{+++}} \tag{4.53}$$

[10] We do not go into the details of these in this course.

[11] e.g. Newton Raphson, Iterative Proportional Fitting, etc.

- Using Lagrange multipliers, we can show that $\hat{E}_{+j+} = n_{+j+}$ and $\hat{E}_{i+k} = n_{i+k}$, so that

$$\hat{E}_{ijk} = \frac{\hat{E}_{+j+}\hat{E}_{i+k}}{n_{+++}} = \frac{n_{+j+}n_{i+k}}{n_{+++}} \tag{4.54}$$

4.4.1.2 $[XY, XZ, YZ]$

$\hat{E}_{ijk}$ depends only on $\hat{E}_{ij} = n_{ij}$, $\hat{E}_{ik} = n_{ik}$ and $\hat{E}_{jk} = n_{jk}$. However, this relation cannot be expressed by a tractable expression (hence numerical methods such as Newton-Raphson and Iterative Proportional Fitting must be used, as discussed in Section 3.3.6[12]).

4.5 Model Fit and Selection

4.5.1 Nested Models

Model $\mathcal{M}_0$ is nested in model $\mathcal{M}_1$ if and only if $\mathcal{M}_1$ can be reduced to the simpler $\mathcal{M}_0$ by assigning specific fixed values (usually 0) to some of the parameters of $\mathcal{M}_1$.

4.5.1.1 Example

Model $[Y, XZ]$ is nested in model $[XY, XZ]$ since

$$\log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} \qquad \forall i, j, k \tag{4.55}$$

can be obtained from

$$\log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \qquad \forall i, j, k \tag{4.56}$$

by setting $\lambda_{ij}^{XY} = 0 \qquad \forall i, j.$

4.5.1.2 For Three-Way Tables

Sequences of nested models for three-way tables, from the saturated $[XYZ]$ model to the model of mutual independence $[X, Y, Z]$, are displayed in Figure 4.1.

4.5.2 Generalised Goodness-of-Fit Test

- We use a Goodness-of-Fit test to test between the two hypotheses:

$$\mathcal{H}_0: \quad \text{model (that is, dependency type) } \mathcal{M}_0 \tag{4.57}$$
$$\mathcal{H}_1: \quad \text{model (that is, dependency type) } \mathcal{M}_1 \tag{4.58}$$

  where $\mathcal{M}_0$ is nested in $\mathcal{M}_1$.

- Using Wilks' Theorem (Section 1.2.3.4), under $\mathcal{H}_0$ we have that

$$G^2(\mathcal{M}_0|\mathcal{M}_1) = 2\sum_{i_1 \ldots i_q} n_{i_1 \ldots i_q} \log\left(\frac{\hat{E}^{(1)}_{i_1 \ldots i_q}}{\hat{E}^{(0)}_{i_1 \ldots i_q}}\right) \sim \chi^2_{df} \tag{4.59}$$

[12] We do not go into the details of these in this course.

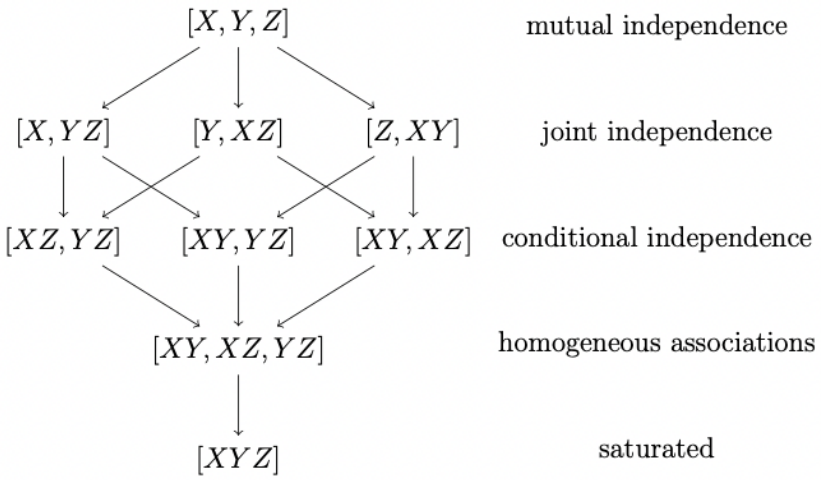


Figure 4.1: Sequences of nested models for three-way tables.

where $\hat{E}^{(0)}_{i_1 \ldots i_q}$ is the MLE of $E^{(0)}_{i_1 \ldots i_q}$ under the model $\mathcal{M}_0$, $\hat{E}^{(1)}_{i_1 \ldots i_q}$ is the MLE of $E^{(1)}_{i_1 \ldots i_q}$ under the model $\mathcal{M}_1$, and the degrees of freedom df are given by

$$df = d_1 - d_0 \tag{4.60}$$

where d_0 is the number of free parameters in model $\mathcal{M}_0$, and d_1 is the number of free parameters in model $\mathcal{M}_1$.

- Notice that we can show:

$$G^2(\mathcal{M}_0|\mathcal{M}_1) = G^2(\mathcal{M}_0|\mathcal{M}_{sat}) - G^2(\mathcal{M}_1|\mathcal{M}_{sat}) \tag{4.61}$$

where $\mathcal{M}_{sat}$ is the saturated model of dimension q, and q is the number of variables.

- The corresponding rejection area of the GLR test of null hypothesis $\mathcal{H}_0$ against alternative hypothesis $\mathcal{H}_1$, at significance level α, is

$$R = \{\boldsymbol{n} : G^2(\mathcal{M}_0|\mathcal{M}_1) \geq \chi^2_{df,\alpha}\} \tag{4.62}$$

4.5.3 Model Selection

LLM selection consists of a sequential search between hierarchical nested models:

- *Backwards model selection:* start from the saturated model, and at each step remove the least significant term. This term can be found by conditionally testing the significance of each potential term to be removed (ensuring the model is still hierarchical). The process stops and decides for the model for which the next term to be removed leads to a large value of the test statistic given by Equation (4.59).
- *Forwards model selection:* Start from the model of complete independence and at each step add the most significant potential term. This term can be found by conditionally testing the significance of each potential term to be added (ensuring the model is still hierarchical). The process stops and decides for the model for which the next term to be added leads to a small value of the test statistic given by Equation (4.59).

4.5.3.1 Selection Criteria

We need to use some form of criteria to decide between non-nested models in the above stepwise algorithms. We here consider AIC and BIC.

4.5.3.1.1 AIC

- One option is to minimise the Akaike Information Criterion (AIC).
- For model $\mathcal{M}$, the AIC is given by

$$-2l_{\mathcal{M}}(\hat{\boldsymbol{\lambda}}) + 2d_{\mathcal{M}} \tag{4.63}$$

where $l_{\mathcal{M}}(\hat{\boldsymbol{\lambda}})$ is the value of the log-likelihood function at the MLE for $\boldsymbol{\lambda}$ for the appropriate set of λ parameters corresponding to model $\mathcal{M}$, and $d_{\mathcal{M}}$ is the number of free parameters in model $\mathcal{M}$.

4.5.3.1.2 BIC

- Another option is to minimise the Bayesian Information Criterion (BIC).
- For model $\mathcal{M}$, the BIC is given by

$$-2l_{\mathcal{M}}(\hat{\boldsymbol{\lambda}}) + \log(n_{+\ldots+})d_{\mathcal{M}} \tag{4.64}$$

where $l_{\mathcal{M}}(\hat{\boldsymbol{\lambda}})$ is the value of the log-likelihood function at the MLE for $\boldsymbol{\lambda}$ for the appropriate set of λ parameters corresponding to model $\mathcal{M}$, $d_{\mathcal{M}}$ is the number of free parameters in model $\mathcal{M}$, and $n_{+\ldots+}$ is the total number of observations.

4.5.3.1.3 Example: $[Y, XZ]$

- Following Section 4.2.2, assuming Poisson sampling and a model $\mathcal{M}$ given by $[Y, XZ]$, we have that

$$\hat{E}_{ijk} = \frac{\hat{E}_{i+k}\hat{E}_{+j+}}{n_{+++}} = \frac{n_{i+k}n_{+j+}}{n_{+++}} \tag{4.65}$$

so that

$$\begin{aligned} l(\hat{\boldsymbol{\lambda}}) &= \sum_{ijk}(-\hat{E}_{ijk} + n_{ijk}\log\hat{E}_{ijk} - \log n_{ijk}!) && (4.66)\\ &= \sum_{ijk}\left(n_{ijk}\log\left(\frac{n_{i+k}n_{+j+}}{n_{+++}}\right) - \frac{n_{i+k}n_{+j+}}{n_{+++}} - \log n_{ijk}!\right) && (4.67) \end{aligned}$$

- Then we have that

$$\begin{aligned} AIC(\mathcal{M}) &= -2\sum_{ijk}\left(n_{ijk}\log\left(\frac{n_{i+k}n_{+j+}}{n_{+++}}\right) - \frac{n_{i+k}n_{+j+}}{n_{+++}} - \log n_{ijk}!\right) + 2(IK + J - 1)\\ BIC(\mathcal{M}) &= -2\sum_{ijk}\left(n_{ijk}\log\left(\frac{n_{i+k}n_{+j+}}{n_{+++}}\right) - \frac{n_{i+k}n_{+j+}}{n_{+++}} - \log n_{ijk}!\right) + \log(n_{+++})\,(IK + J - 1) \end{aligned}$$

4.5.3.2 Word of Caution

- We should not let an algorithm blindly decide the model. Sometimes, the nature of the problem or experimental conditions dictate the presence of insignificant terms in the model.

4.5.3.2.1 Example

- Suppose in a survey that responders are cross-classified according to their educational level (X), marital status (Y), gender (Z), and age in categories (W).
- Suppose that according to the experimental design it is controlled over gender and age, in the sense that the number of people of each gender participating in the survey is prespecified for each of the age categories.
- This means that the table marginals n_{++kl}, $\forall k, l$ are fixed by design.
- If the ZW interaction term is not found to be significant by the selection algorithm, and is thus not included in the model, then the corresponding likelihood equations, given by

$$E_{++kl} = n_{++kl} \qquad \forall k, l \tag{4.68}$$

  are missing.
- Consequently, the number of subjects of each gender assigned by the adopted model to each age group will not agree with the known prespecified numbers.
- Thus, the ZW interaction term should be included in the model, even if it is insignificant.
- In this case, the λ_{kl}^{ZW} terms signal the underlying product multinomial sampling design and not the physical significance of this interaction.

Part II

Generalised Linear Models

Chapter 5

Motivation and Binary Regression

5.1 Motivation

5.1.1 Modelling Functional Relationships

In many cases in statistics, we wish to describe a *functional relationship* between two quantities. These types of relationships are very common; indeed they are almost what we mean when we talk of cause and effect. Mathematically speaking, they mean that the value of one variable, Y, taking values in a space $\mathcal{Y}$, depends on the value of another variable, X, taking values in a space $\mathcal{X}$, only via the value of a function f of X.[1]

The extreme case of this situation is where the value $f(x)$ determines the value of Y exactly, in which case we might as well say that $y = f(x)$. Since we are doing statistics, we will write this in terms of a probability distribution. For discrete $\mathcal{Y}$,[2]

$$P(y|x, f) = \begin{cases} 1 & \text{if } y = f(x) \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

If $\mathcal{Y}$ is continuous, the idea is the same but a little more complicated.

In general, however, knowing $f(x)$ will not be sufficient to determine the value of Y with certainty; there will be uncertainty due to other quantities, known or unknown, that affect the value. The uncertainty in our knowledge of the value of Y caused by these other effects is described by a probability distribution

$$P(y|x, f, K) = P(y|f(x), K) \tag{5.2}$$

that nevertheless only depends on $f(x)$[3].

The variable $f(X)$ thus acts as a parameter for the distribution for Y: different values x of X lead to different distributions. However, this does not seem sufficient to capture the

[1]The notion of variables (and random variables) was discussed in Q1-1.

[2]The notation $P(y|\ldots)$ will be used to mean the probability of $Y = y$ in the case that $\mathcal{Y}$ is countable, and to represent the density $P(Y \in [y, y + dy])/dy$ in the case that $\mathcal{Y}$ is continuous.

[3]Here K represents all the other knowledge, e.g. values of other parameters, that may be relevant. We will, however, often drop the K for notational convenience.

idea of a functional relationship. For example, we could have a Gaussian distribution for Y whose variance was given by a function of X, but we would not regard this as expressing a functional relationship between X and Y. We therefore need some further constraint on the relationship between the value of the function and the probability distribution.

One way to think of such a constraint is that the function output should be a prediction of the value of Y, based on the probability distribution. The natural concept of a prediction for the value of Y is simply the mean of the distribution, $\mathrm{E}[Y|f,x]$. This then leads to a very important equation, one that defines the models to be discussed in this part of the course: we will say that a probability distribution for Y, $P(y|f,x)$, describes a functional relationship between X and Y when

$$\mathrm{E}[Y|f,x] = f(x) \tag{5.3}$$

In other words, instead of the *value* of Y being given by $f(x)$ (as in the deterministic case of Equation (5.1)), rather the *expectation* of the value of Y under the probability distribution is given by $f(x)$.

5.1.1.1 Choice of Function f

The function f will come from some given space of functions $\mathcal{F}$, whose elements map $\mathcal{X}$ to the space containing possible values of $\mathrm{E}[Y|f,x]$, which in our case will always be $\mathbb{R}$: that is, the elements of $\mathcal{F}$ are real-valued functions of X. However, in most cases, we do not know *a priori* which function in $\mathcal{F}$ is involved; indeed the main task is usually to *estimate* this function given some examples of input/output pairs $\{(x_i, y_i)\}_{i\in[1..n]}$.[4] Thus, the function too is a variable, F, taking values in $\mathcal{F}$.

There are thus several ingredients to this type of modelling:

- the input *predictor* space $\mathcal{X}$ and output *response* space $\mathcal{Y}$;
- the space of possible functions $\mathcal{F}$;
- the probability distribution, $P(y|f(x),K)$, satisfying $\mathrm{E}[Y|f,x] = f(x)$.

5.1.1.2 Example: Linear Models

Consider the standard linear model (reviewed in detail in Section 1.3):

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Here, y is the *response* variable, the x_j's are the *predictor* variables and ϵ is an *error* term. Note that we typically have $x_1 \equiv 1$, so that β_1 is the intercept term.

Linear models correspond to the following ingredients.

- The space of responses $\mathcal{Y} = \mathbb{R}$; the space of predictors $\mathcal{X} = \mathbb{R}^p$.[5]
- The space of possible functions $\mathcal{F}$ used in linear models is generated by the linear maps $f : \mathcal{X} \to \mathbb{R}$, i.e. from $\mathbb{R}^p$ to $\mathbb{R}$:

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R} : \; f(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x} \right\} \tag{5.4}$$

[4] The notation $[m..n]$ will indicate all natural numbers i such that $m \leq i \leq n$.

[5] Recall that the space of predictors is not the same as the space of covariates. For example, there might only be one covariate, $z \in \mathbb{R}$, but many predictors, corresponding, for example, to $x_a = z^{a-1}$, $a \in [1..p]$.

where $\boldsymbol{x} = (x_1 = 1, x_2, x_3, ..., x_p)$ and necessarily $\boldsymbol{\beta} \in \mathbb{R}^p$.

- The probability distribution is Gaussian:

$$P(y|\boldsymbol{x}, f, K) = \mathcal{N}(y; f(\boldsymbol{x}), \sigma^2) \tag{5.5}$$

  Note that this guarantees that Equation (5.3) is satisfied.

While linear models offer significant inferential tools, note, however, that the use of a Gaussian distribution is restrictive; we necessarily must have that $\mathcal{Y} = \mathbb{R}$, and the linear functions have a range, which in principle, is also all of $\mathbb{R}$. There are many quantities that must be represented by variables with more limited ranges, even a discrete set of values, and linear models do not seem suited to these cases.

5.1.2 Functional Relationships: A Summary

- A functional relationship attempts to explain the dependency between a response variable Y on some other (predictor) variables X only via some function f of X.
- More precisely, a probability distribution for Y, notated $P(y|f, \boldsymbol{x})$, describes a functional relationship between X and Y when

$$\mathrm{E}[Y|f, \boldsymbol{x}] = f(\boldsymbol{x}) \tag{5.6}$$

- Such modelling removes the requirement that $\mathcal{Y} = \mathbb{R}$, as is the case for linear models. For example, $\mathcal{Y}$ may be binary, discrete or constrained continuous, such as is the case for the datasets presented in Section 5.2.
- There are several ingredients to this type of modelling:
  - the input *predictor* space $\mathcal{X}$ and output *response* space $\mathcal{Y}$;
  - the space of possible functions $\mathcal{F}$, whose elements map $\mathcal{X}$ to the space containing possible values of $\mathrm{E}[Y|f, \boldsymbol{x}]$, which in our case will always be $\mathbb{R}$;
  - a probability distribution, $P(y|f(\boldsymbol{x}), K)$, satisfying $\mathrm{E}[Y|f, \boldsymbol{x}] = f(\boldsymbol{x})$.

5.2 Example Datasets

We introduce some of the datasets used throughout this part of the course.

Considerations:

- Can we fit a standard linear model to these datasets?
- If not, why not?
- What might we do instead?

5.2.1 Dataset A: Birthweight Data

Consider the dataset `bpd` from R libary `SemiPar` (type `?bpd` for information).

```
library( "SemiPar" )
data( "bpd" )
```

This dataset is from a study of low-birthweight infants in a neonatal intensive care unit, with data collected on $n = 223$ babies by day 28 of life. We model `bpd` as response, with `1` indicating presence, and `0` indicating absence, of Bronchopulmonary dysplasia (BPD), taking `birthweight` as covariate. First, let's plot the data.

```
plot( bpd )
```

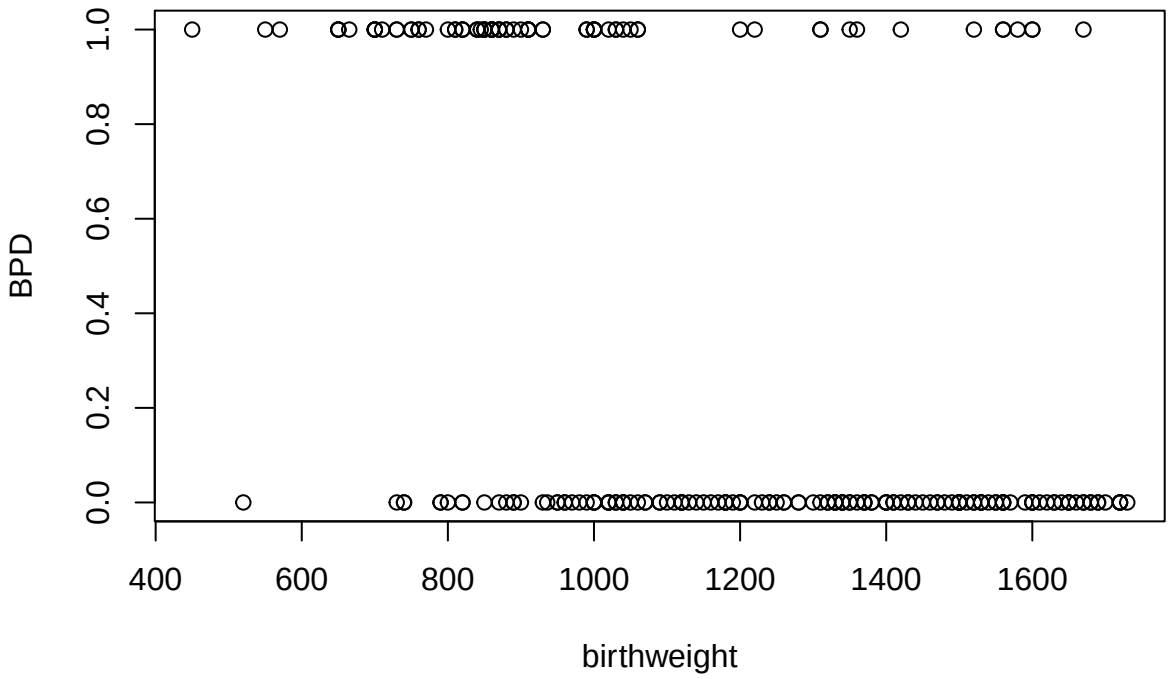


Figure 5.1: Plot of presence of BPD against birthweight for the `bpd` dataset.

Question

- *How does the probability of the development of BPD depend on birthweight?*

Response Type

- `bpd`
- $\mathcal{Y} \in \{0, 1\}$
- $\mathrm{E}[Y|f, x] \in [0, 1]$
- $P(y|f, x)$.....Bernoulli?

5.2.2 Dataset B: US Polio Data

Consider the polio data from library `gamlss.data`.

```
library( "gamlss.data" )
data( "polio" )
```

This dataset is a matrix of count data, giving the monthly number of polio cases in United States from 1970 to 1983. We wish to convert this into a usable time series. In other words, we wish to covert this into a matrix with two columns with

- covariate `time` in the first column ranging from 1 to 168, starting with January 1970.
- response `cases` in the second column, indicating the monthly number of polio cases.

We do this as follows

```
uspolio <- as.data.frame( matrix( c( 1:168, t( polio ) ), ncol = 2 ) )
colnames( uspolio ) <- c("time", "cases")
```

First, let's plot the data.

```
plot( uspolio, type = "h" )
```

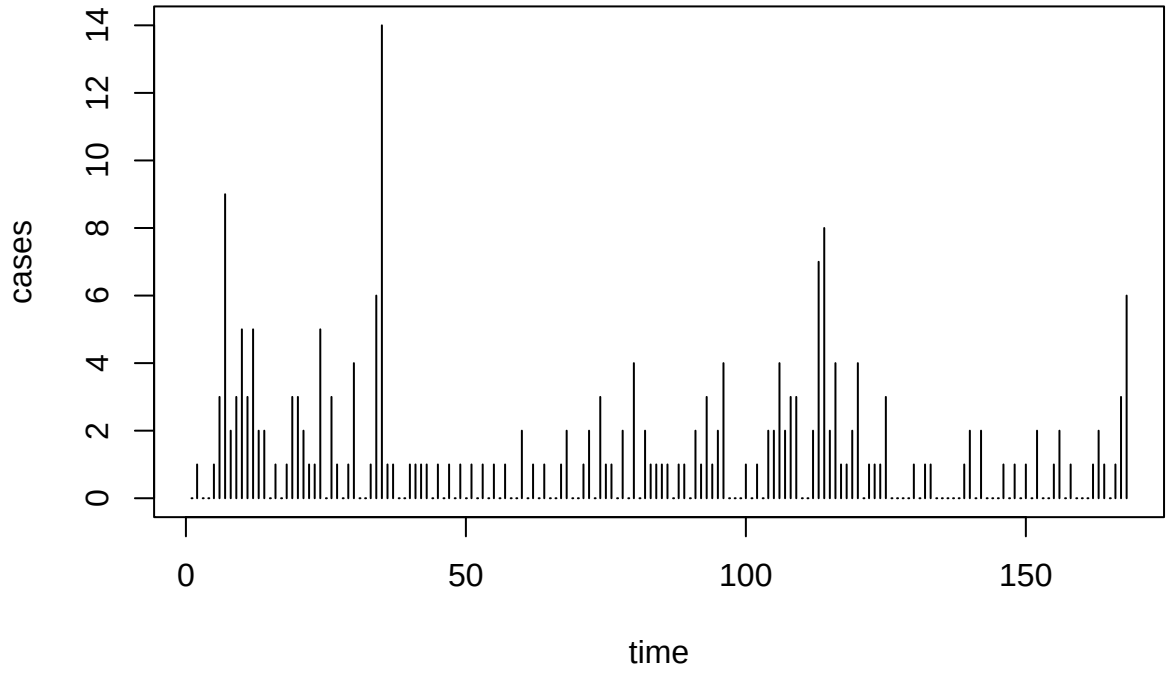


Figure 5.2: Plot of monthly reported number of Polio cases in the US between 1970 and 1983.

Question

- *How has Polio incidence changed over time?*

Response Type

- `cases`
- $\mathcal{Y} \in \mathbb{N}$
- $\mathrm{E}[Y|f,x] \in \mathbb{R}_{\geq 0}$
- $P(y|f,x)$.....Poisson?

5.2.3 Dataset C: Hospital Stay Data

The data, from R package `npmlreg`, are a subset of a larger data set collected on persons discharged from a selected Pennsylvania hospital as part of a retrospective chart review of

antibiotic use in hospitals. Let's load the data and then generate a plot.

```
library(npmlreg)
data(hosp)
plot(hosp[,c("duration","age","temp1","wbc1")])
```

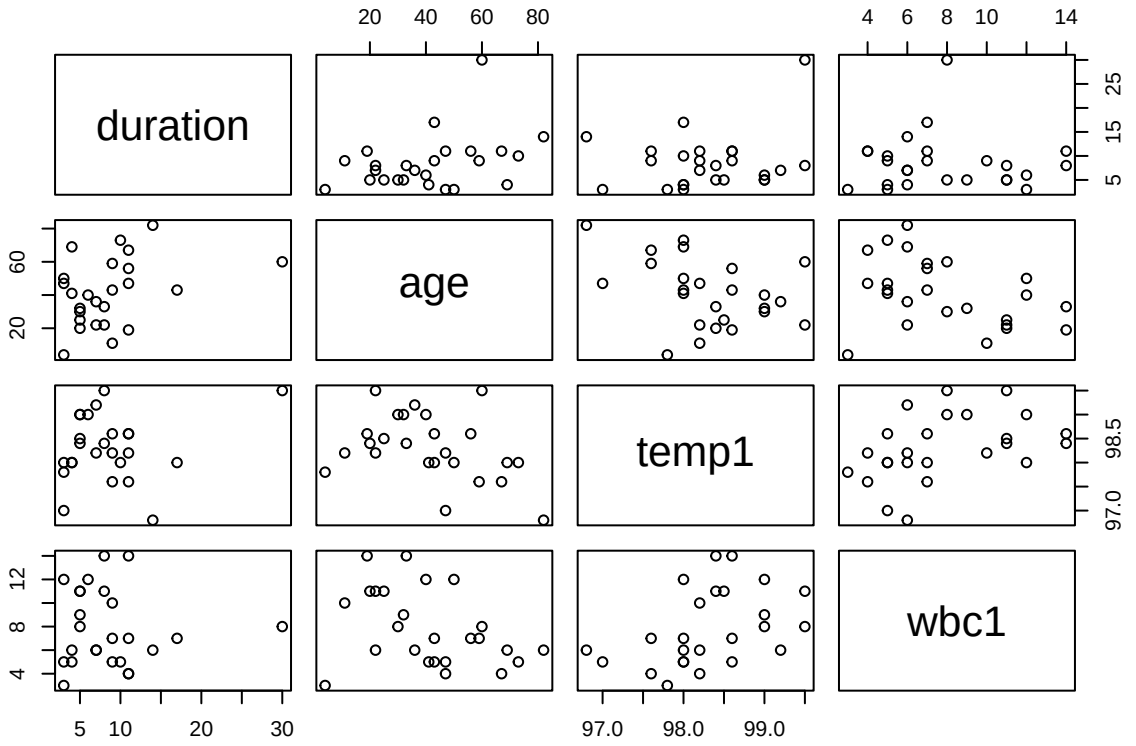


Figure 5.3: Scatter plots of length of hospital stay against three of the most significant covariates.

Question

- *How does the duration of a patient's hospital stay vary, based on data available at the time of admission?*

Response Type

- `duration`
- $\mathcal{Y} \in \mathbb{R}_{\geq 0}$
- $\mathrm{E}[Y|f,x] \in \mathbb{R}_{\geq 0}$
- $P(y|f,x)$.....Gamma?

5.3 Binary Regression

- We consider modelling for the special case in which the data takes on binary values: $Y \in \{0,1\}$.
- Binary values usually denote membership or not of some subset of a set: having a disease or not; a credit card transaction being fraudulent or genuine; and so on (can you think of some other examples?).

- In this section, we will study this special case, both as important in its own right, and as an introduction to the main ideas of Generalised Linear Models (GLMs), to be discussed in full generality later.
- Dataset A (the `bpd` dataset) considered in Section 5.2.1 is an example in which response variable Y is binary, taking values 1 or 0, representing presence or absence of BPD respectively.

5.3.1 Modelling Considerations

- If a response variable Y is binary, we know that its expectation must lie in the interval $[0, 1]$, that is
$$\mathrm{E}[Y|f, \boldsymbol{x}] = f(\boldsymbol{x}) \in [0, 1] \tag{5.7}$$
- The only probability distribution available is the Bernoulli distribution, completely characterized by the probability π that $Y = 1$.
- The same quantity, π, is also the expectation of Y under the Bernoulli distribution, that is
$$P(Y = 1|f, \boldsymbol{x}) = \pi(\boldsymbol{x}) = \mathrm{E}[Y|f, \boldsymbol{x}] = f(\boldsymbol{x}) \tag{5.8}$$
- Thus the spaces $\mathcal{Y}$ and $\mathcal{X}$, along with a probability distribution satisfying Equation (5.6), are in place. The remaining task is to think of a suitable set of possible functions, $\mathcal{F}$.
- A first guess might be to try a linear model:
$$f(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} \tag{5.9}$$
What is wrong with this choice? Well, as you might recall from Section 5.2.1, such a function is unsatisfactory, because for any given value of $\boldsymbol{\beta}$, there will always be values of $\boldsymbol{x}$ such that $\boldsymbol{\beta}^T\boldsymbol{x} \notin [0, 1]$. While it may be that there are particular cases where $\boldsymbol{x}$ only takes on a certain range of values, thereby respecting the constraint on $\mathrm{E}[Y|f, \boldsymbol{x}]$, this cannot be true in general.
- The idea that lies behind GLMs is to apply a *response* function to the linear predictor $\eta = \boldsymbol{\beta}^T\boldsymbol{x}$, in order to constrain it to have the correct range. We thus need to choose a function $h : \mathbb{R} \to [0, 1]$, and then set
$$\mathrm{E}[Y|f, \boldsymbol{x}] = f(\boldsymbol{x}) = h(\boldsymbol{\beta}^T\boldsymbol{x}) \tag{5.10}$$
Such a function is shown in Figure 5.4, with the R code used to plot this given below.

```
eta <- seq( from = -6, to = 6, length = 1000 )
response <- function( x ){ exp(x)/(1+exp(x)) }
plot( eta, response(eta), type = "l",
      xlab = expression( eta ), ylab = expression( h(eta) ) )
```

- What possibilities for h exist? Given the definition, it is apparent that any cumulative distribution function would work, but we would like to be more specific than that.

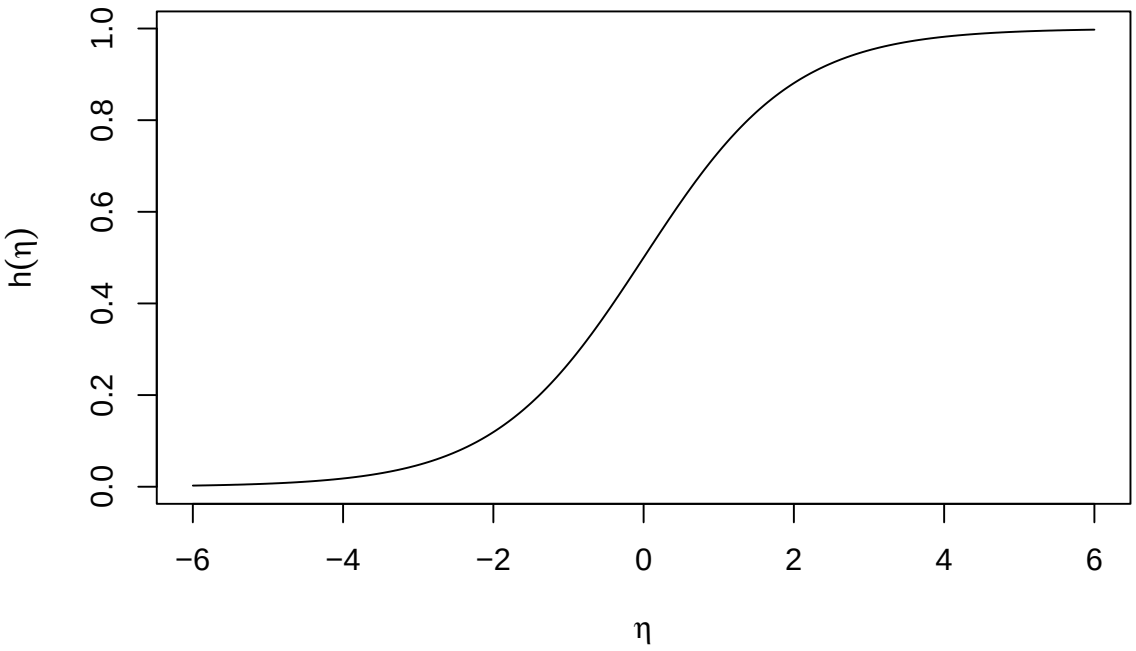


Figure 5.4: Graph of a bijective function $\mathbb{R} \to [0, 1]$

- In fact, there is a *standard*, or at least most usual, choice: the *logistic* function (it is this function of η that is plotted in Figure 5.4):

$$h(\eta) = \frac{e^\eta}{1 + e^\eta} = \frac{1}{1 + e^{-\eta}} \tag{5.11}$$

  This choice for h leads to the *logistic regression model* or just *logistic model*.

5.3.2 The Logistic Regression Model

The logistic regression model consists of three components:

- The *linear predictor*:

$$\eta = \boldsymbol{\beta}^T \boldsymbol{x} \tag{5.12}$$

- The logistic *response function*:

$$\mathrm{E}[Y|f, \boldsymbol{x}] = \pi(\boldsymbol{x}) = f(\boldsymbol{x}) = h(\eta) = \frac{e^\eta}{1 + e^\eta} \tag{5.13}$$

- The *probability distribution*:

$$Y \sim \text{Bernoulli}(\pi(\boldsymbol{x})) \tag{5.14}$$

  or, in other words,

$$P(Y = 1|f, \boldsymbol{x}) = P(Y = 1|\boldsymbol{\beta}, \boldsymbol{x}) = \pi(\boldsymbol{x}) \tag{5.15}$$

So, to summarise,

- Y is a binary variable, that is $Y \in \{0, 1\}$.

- At $\boldsymbol{x} \in \mathcal{X}$ for which the response y is unobserved, we wish to say something about the distribution of y.
- We assume a particular distibutional form for y (in this case bernoulli), and would like to connect $\boldsymbol{x}$ with some feature of this distribution, namely the expectation, by a function f.
- However, since the expectation must take on a value in $[0, 1]$, the standard linear model $\eta = \boldsymbol{\beta}^T x$, with potential range $\mathbb{R}$, cannot be used.
- We therefore transform the linear predictor using the response function h, which we ensure has potential range $[0, 1]$, as required.

5.3.2.1 Reformulations

The following are all equivalent ways of reformulating the logistic model's relation between π and the linear predictor $\eta = \boldsymbol{\beta}^T \boldsymbol{x}$, providing slightly different perspectives.

- The logistic model:

$$\pi(\boldsymbol{x}) = \frac{e^{\eta}}{1 + e^{\eta}} \tag{5.16}$$

- A model for the *odds*:

$$\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = e^{\eta} \tag{5.17}$$

- A linear model for the log odds, or *logits*:

$$\text{logit}(\pi(\boldsymbol{x})) = \log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \eta \tag{5.18}$$

5.3.2.2 Interpretation of Parameter Values

Suppose we increase the value of the a^{th} predictor x_a by 1, that is

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + (0, \ldots, 0, 1, 0, \ldots, 0) \tag{5.19}$$

where the 1 is in the a^{th} position. Then

$$\text{logit}\left(\pi(\tilde{\boldsymbol{x}})\right) = \boldsymbol{\beta}^T \tilde{\boldsymbol{x}} = \boldsymbol{\beta}^T \boldsymbol{x} + \beta_a \tag{5.20}$$

or (by applying the exponential function to both sides)

$$\frac{\pi(\tilde{\boldsymbol{x}})}{1 - \pi(\tilde{\boldsymbol{x}})} = e^{\beta_a} \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} \tag{5.21}$$

Thus if the a^{th} predictor increases in value by 1, the odds on $Y = 1$ change multiplicatively by e^{β_a}.

5.3.2.3 Practical Example: Dataset A: `bpd` Data

We will look in detail at modelling the dataset introduced in Section 5.2.1 later. For now, just consider the linear predictor is given by

$$\eta(\boldsymbol{x}) = \beta_1 + \beta_2 \texttt{birthweight} \tag{5.22}$$

The resulting parameter estimates (we will look at parameter estimation in Section 5.3.3) turn out to be

$$\hat{\beta}_1 = 4.034 \tag{5.23}$$

$$\hat{\beta}_2 = -0.004229 \tag{5.24}$$

Thus, according to the model, when birth weight increases by 1, we estimate that the odds of developing BPD are multiplied by

$$e^{\hat{\beta}_2} = e^{-0.004229} = 0.996 \tag{5.25}$$

that is, according to this model, each additional gram of birth weight reduces the odds of BPD by about 0.4%.

5.3.3 Estimation

- We have now set up the logistic model for binary regression.
- In practice, of course, we do not know the parameters $\boldsymbol{\beta}$ (or in other words, the function $f \in \mathcal{F}$), and in fact these are usually the quantities in which we are most interested.
- We therefore wish to make inferences about them: given data $\{y_i\}_{i\in[1..n]}$ corresponding to covariate values $\{\boldsymbol{x}_i\}_{i\in[1..n]}$, how can we estimate $\boldsymbol{\beta}$?
- We will use maximum likelihood for this purpose.
- The probability of getting the given data $\{y_i\}$ given knowledge of the $\{\boldsymbol{x}_i\}$ and $\boldsymbol{\beta}$ is

$$\begin{aligned} P(\{y_i\} \mid \{\boldsymbol{x}_i\}, \boldsymbol{\beta}) &= \prod_i P(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) && (5.26) \\ &= \prod_i \pi(\boldsymbol{x}_i)^{y_i} \left(1 - \pi(\boldsymbol{x}_i)\right)^{1-y_i} && (5.27) \end{aligned}$$

  where we have assumed that:

  - given the $\{\boldsymbol{x}_i\}, i = 1, ..., n$, our knowledge of any particular $y_k, k \in \{1, ..., n\}$ is not altered by knowledge of other y_j for $j \neq k$;
  - given $\boldsymbol{x}_k$, our knowledge of y_k is not altered by knowledge of other $\boldsymbol{x}_j$ for $j \neq k$.

Note that these are distinct assumptions, the first corresponding to the y_i being independent of each other.[6]

Abbreviating $\pi(\boldsymbol{x}_i)$ by π_i, the log likelihood is thus

$$\begin{aligned} l\left(\boldsymbol{\beta}; \{\boldsymbol{x}_i\}, \{y_i\}\right) &= \log P\left(\{y_i\} \mid \{\boldsymbol{x}_i\}, \boldsymbol{\beta}\right) && (5.28)\\ &= \sum_i \Big[y_i \log \pi_i + (1-y_i)\log(1-\pi_i) \Big] && (5.29)\\ &= \sum_i \Big[y_i \log \frac{\pi_i}{1-\pi_i} + \log(1-\pi_i) \Big] && (5.30)\\ &= \sum_i \Big[y_i \boldsymbol{\beta}^T \boldsymbol{x}_i - \log(1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}) \Big] && (5.31) \end{aligned}$$

where for the final line we note that

$$1 - \pi_i = 1 - \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}} = \frac{1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i} - e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}} = \frac{1}{1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}} \tag{5.32}$$

5.3.4 The Score Function

- The *score function* is given by

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta}) = \nabla l\left(\boldsymbol{\beta}\right) \tag{5.33}$$

  that is, the derivative of the log-likelihood function with respect to $\boldsymbol{\beta}^T$.

- The Maximum Likelihood Estimate (MLE) of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}$, will, for well-behaved likelihoods, satisfy the equation

$$\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0 \tag{5.34}$$

  Equation (5.34) is known as the *score equation.*

- To be a maximum, we must also have

$$\boldsymbol{H}(l(\hat{\boldsymbol{\beta}})) \leq 0 \tag{5.35}$$

[6]The word *assumption* is a bad one. It sounds as if we are *assuming* that something is *true* when in reality, it might be *false.* However, independence is not a property of the world at all. For example, when two people, A and B, walk into a building one after another, the propositions *A carried an umbrella* and *B carries an umbrella* are not independent if I am inside the building and do not know whether it is raining or not, because the first being true suggests that it is raining, which increases the probability of the second being true. However, if I know (condition on the fact) that it is raining, then they are independent, or nearly so, because knowledge that one is true does not affect the probability of the second. Nothing changed in the world, only in my knowledge of it. Thus independence says that given the conditioning knowledge, a change in our knowledge of one quantity does not affect our knowledge of another. Usually independence is used because we judge the relevant changes in knowledge to be small, so that it is a good approximation to our state of knowledge, and because it simplifies calculations greatly.

where ≤ 0 here denotes *negative semidefinite*[7], and $\boldsymbol{H}$ denotes the *Hessian* matrix[8]:

$$\boldsymbol{H}(l(\boldsymbol{\beta})) = \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\boldsymbol{\beta}) \tag{5.36}$$

that is, the second derivative of the log-likelihood function.

5.3.4.1 For Logistic Regression...

- For logistic regression, the score function is given by

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_i \left[y_i \boldsymbol{x}_i - \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}} \boldsymbol{x}_i \right] && (5.37) \\ &= \sum_i (y_i - \pi_i) \boldsymbol{x}_i && (5.38) \end{aligned}$$

where the first line is obtained by use of the chain rule with $u = 1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}$:

$$\frac{\partial}{\partial \boldsymbol{\beta}^T} \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}) = \frac{1}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}} \boldsymbol{x}_i e^{\boldsymbol{\beta}^T \boldsymbol{x}_i} \tag{5.39}$$

5.3.4.2 Solving The Score Equation

- The score equation (for binary regression, or more generally)

$$\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0 \tag{5.40}$$

is a system of nonlinear equations of $\hat{\boldsymbol{\beta}}$ (the parameters).

- In general, no analytical solution exists, so we need to use a numerical method to find a solution. We thus need an algorithm to perform the task we have set ourselves, of finding the MLE, and, in the interests of time efficiency, a computer to perform the computations. There are two broad classes of algorithm that are relevant:[9]
  - Optimization algorithms, which will attempt to find a global or local minimum of l;
  - Nonlinear equation solvers, which will attempt to solve $\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = 0$.

The most common method in our context is *iterative weighted least squares* (IWLS)[10].

[7] Recall: A symmetric matrix A is said to be positive definite if and only if $b^T A b > 0$ for all vectors $b \neq 0$, and positive semi-definite if the inequality is not strict (i.e., $\geq$). Similarly, A is said to be negative definite if and only if $b^T A b < 0$ for all vectors $b \neq 0$, and negative semi-definite if the inequality is not strict.

[8] Note that equations (5.34) and (5.35) are not sufficient for $\hat{\boldsymbol{\beta}}$ to be the MLE. The log likelihood may have many local maxima, all of which will satisfy both these equations, yet, generically, only one can be the global maximum. In fact, these equations are not necessary either: the maximum may not occur at a differentiable point, usually because it occurs at a boundary (e.g. $e^{-\boldsymbol{x}}$ on $\mathbb{R}_{\geq 0}$), but not always (e.g. $(1 - |\boldsymbol{x}|)$ on $[-1, 1]$). Note too that if the second derivative is not negative definite, there will be an infinite number of maxima in the *flat* direction.

[9] In practice, the former are quite often implemented via the latter, although this need not be the case. There are advantages to optimization algorithms that do not require derivatives.

[10] This name will be demystified later when we describe the method.

5.3.4.2.1 Exception An exception to the need for an algorithm is the case with $p = 2$, $x_1 = 1$, and $x_2 \in \{0, 1\}$ (that is, a single binary covariate). In this case, an analytical solution is available. It is left as an exercise for the reader[11].

5.3.5 Logit and Probit

Are there any alternatives to the logistic function? The answer, of course, is *yes*: any cdf will do. In practice, the second most popular response function for binary regression is the cdf of the standard normal distribution

$$\pi(\boldsymbol{x}) = \Phi(\boldsymbol{\beta}^T \boldsymbol{x}) \tag{5.41}$$
$$\Phi^{-1}(\pi(\boldsymbol{x})) = \boldsymbol{\beta}^T \boldsymbol{x} \tag{5.42}$$

The inverse, Φ^{-1}, is known as the *probit* function. A comparison of the logit and probit models is presented in the table in Figure 5.5.

	Logit	Probit
Response function h	logistic function	Gaussian cdf
Model	$\pi(x) = \frac{e^{\beta^T x}}{1+e^{\beta^T x}}$	$\pi(x) = \Phi(\beta^T x)$
Inverse	$\text{logit}(\pi(x)) = \beta^T x$	$\Phi^{-1}(\pi(x)) = \beta^T x$
GLM in R	`link = "logit"`	`link = "probit"`

Figure 5.5: A comparison of the logit and probit models.

5.3.5.1 Practical Example: Dataset A

- We fit a logit and a probit model to the the `bpd` data introduced in Section 5.2.1.
- This uses the `glm` command from R package `SemiPar`, which is the general R command for fitting a generalised linear model.
- Notice, however, similarities and differences with the `lm` command in R, and also where we inform R of whether we wish to fit a `logit` or a `probit` model.

```
# Load the data.
library( SemiPar )
data( bpd );
attach( bpd )

# Fit logistic regression model
fitl <- glm( BPD ~ birthweight, family = binomial( link = logit ) )
fitl
```

[11] Q5-1c, in fact.

```
##
## Call:  glm(formula = BPD ~ birthweight, family = binomial(link = logit))
##
## Coefficients:
## (Intercept)  birthweight
##    4.034291    -0.004229
##
## Degrees of Freedom: 222 Total (i.e. Null);  221 Residual
## Null Deviance:       286.1
## Residual Deviance: 223.7     AIC: 227.7
```

```
# Comparison with probit link
fitp <- glm( BPD ~ birthweight, family = binomial( link = probit ) )
fitp
```

```
##
## Call:  glm(formula = BPD ~ birthweight, family = binomial(link = probit))
##
## Coefficients:
## (Intercept)  birthweight
##    2.276856    -0.002383
##
## Degrees of Freedom: 222 Total (i.e. Null);  221 Residual
## Null Deviance:       286.1
## Residual Deviance: 225.4     AIC: 229.4
```

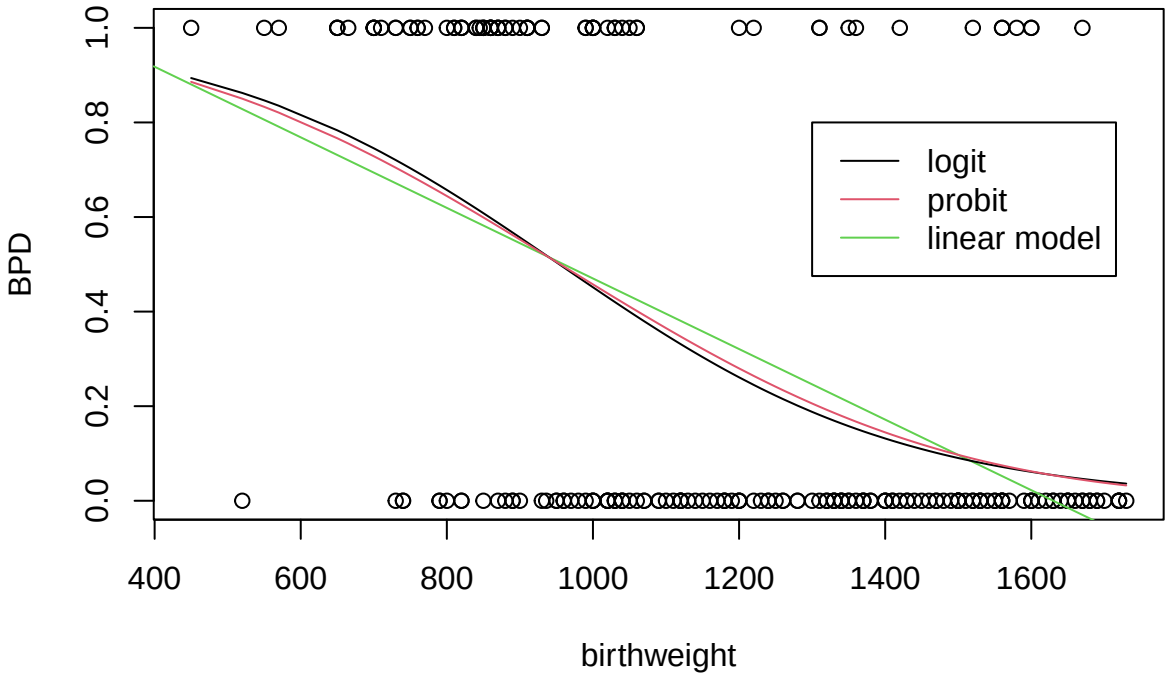
Note that although the parameter values are quite different, if we plot the expectation functions, we can see that they are very similar. We also add on the regression line from a standard linear model for comparison purposes. Notice how the logit and probit models are constrained between 0 and 1 (and would be even if you extended the x-axis beyond the visualised range), however, the linear model visually predicts outside this range on the displayed plot.

```
# Plot the bpd dataset.
plot( birthweight, BPD )

# Try a standard regression line using lm command:
abline( lm( BPD ~ birthweight ), col = 3 )

# Include fitted pi(x_i) into plot
lines( birthweight[order( birthweight )], fitl$fitted[order( birthweight )] )
lines( birthweight[order( birthweight )], fitp$fitted[order( birthweight )],
       col = 2 )

# Add a legend
legend( x = 1300, y = 0.8, col = 1:3, lty = 1,
        legend=c( "logit", "probit", "linear model" ) )
```


BPD
0.0
0.2
0.4
0.6
0.8
1.0
400
600
800
1000
1200
1400
1600
birthweight
logit
probit
linear model

Chapter 6

Exponential Dispersion Family

Our models of functional relationships are defined by a constraint on the mean of the distribution, Equation (5.6). In order for them to be useful in practice, this constraint must be effective, which in practice means that it must be easy to relate the parameters of the probability distribution to its expectation.

In this section, we examine a very general class of probability distributions for which this is the case.

6.1 The Exponential Dispersion Family

6.1.1 Definition

An *exponential dispersion family (EDF)* of probability distributions has the following form:

$$P(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right] \tag{6.1}$$

where

- $\theta \in \mathbb{R}$ is the *natural* parameter,
- $\phi > 0$ is the *dispersion* parameter; in many settings this is not of direct interest, and thus may be referred to as a "nuisance" parameter.
- $b : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$ are both functions. The function b is known as the "log normaliser" for reasons that will become clear.

6.1.2 Examples

6.1.2.1 Poisson

The Poisson distribution for $y \in \mathbb{N}$ is

$$\begin{aligned} P(y|\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} && (6.2)\\ &= \exp[y\log\lambda - \lambda - \log y!] && (6.3) \end{aligned}$$

It is thus an EDF with

- $\theta = \log \lambda$
- $\phi = 1$;
- $b(\theta) = \lambda = e^{\theta}$;
- $c(y, \phi) = -\log y!$

6.1.2.2 Bernoulli

The Bernoulli distribution has $y \in \{0, 1\}$. Its distribution is

$$\begin{aligned} P(y|\pi) &= \pi^y (1-\pi)^{1-y} && (6.4) \\ &= \exp\left[y \log \frac{\pi}{1-\pi} + \log(1-\pi)\right] && (6.5) \end{aligned}$$

It is thus an EDF with

- $\theta = \log \frac{\pi}{1-\pi}$
- $\phi = 1$;
- $b(\theta) = -\log(1-\pi) = \log(1+e^{\theta})$;
- $c(y, \phi) = 0$.

6.1.2.3 Gaussian

The normal distribution has $y \in \mathbb{R}$. Its density is

$$\begin{aligned} P\left(y|\mu, \sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] && (6.6) \\ &= \exp\left[\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] && (6.7) \end{aligned}$$

It is thus an EDF with

- $\theta = \mu$,
- $\phi = \sigma^2$;
- $b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2$;
- $c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$.

6.1.2.4 Further Example Distributions

- Exponential
- Gamma
- Inverse Gamma
- Binomial
- Chi-squared
- Beta

The t-distribution is not an EDF distribution.

6.2 Properties of EDFs

- In order that a probability distribution be normalised, we must have

$$\int P(y|\theta,\phi)\,dy = 1 \tag{6.8}$$

- Making use of Equation (6.8) in the context of an EDF distribution (Equation (6.1)), we have

$$\int \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right] dy = 1 \tag{6.9}$$

$$\implies \quad \exp\left(-\frac{b(\theta)}{\phi}\right) \int \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right)\, dy = 1 \tag{6.10}$$

$$\implies \quad \frac{b(\theta)}{\phi} = \log \int \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right)\, dy \tag{6.11}$$

- Equation (6.11) determines b in terms of ϕ and the function c. It is therefore not an independently variable quantity. The function b will be known as the "log normaliser", although note that the factor of ϕ is necessary too.

6.2.1 Mean

- If we differentiate Equation (6.11) with respect to θ (using the chain rule[1] and Leibniz rule[2]), we find

$$\frac{b'(\theta)}{\phi} \;=\; \frac{\int \frac{y}{\phi} \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy}{\int \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy} \tag{6.12}$$

where Equation (6.11) can now be used to substitute the denominator on the right hand side of Equation (6.12) to give

$$\begin{aligned}
\frac{b'(\theta)}{\phi} &= \int \frac{y}{\phi} \frac{\exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right)}{\exp\left(\frac{b(\theta)}{\phi}\right)} dy \\
&= \int \frac{y}{\phi} \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right) dy \qquad (6.13) \\
&= \int \frac{y}{\phi} P(y|\theta,\phi)\,dy \\
&= \frac{1}{\phi}\mathrm{E}[Y|\theta,\phi] \qquad (6.14)
\end{aligned}$$

so that

$$b'(\theta) = \mathrm{E}[Y|\theta,\phi] \tag{6.15}$$

[1] with $u = \int \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy$

[2] Leibniz rule allows the exchangeability of integral and derivative.

- Now, it turns out that b' is almost always invertible for finite parameter values, because $b'' > 0$ except when the variance of the distribution is zero (see Section 6.2.2). Thus,

$$\mu \triangleq \mathrm{E}[Y|\theta,\phi] = b'(\theta) \tag{6.16}$$

means that

$$\theta = (b')^{-1}(\mu) \tag{6.17}$$

- Notice that the EDF distribution can therefore be parameterised in terms of θ or in terms of μ. Equations (6.16) and (6.17) are the crucial equations from the point of view of models of functional relationships (and GLMs in particular), as they relate the parameterisation of the distribution to its expectation in a bijective way.

6.2.2 Variance

- From Equation (6.13), we have that

$$b'(\theta) = \exp\left(-\frac{b(\theta)}{\phi}\right)\int y \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy \tag{6.18}$$

- We can differentiate again (using the product rule) to obtain:

$$\begin{aligned} b''(\theta) &= -\frac{b'(\theta)}{\phi}b'(\theta) + \exp\left(-\frac{b(\theta)}{\phi}\right)\int \frac{y^2}{\phi}\exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy \\ &= -\frac{\mu^2}{\phi} + \frac{1}{\phi}\mathrm{E}[Y^2|\theta,\phi] \\ &= \frac{1}{\phi}\mathrm{Var}[Y|\theta,\phi] \end{aligned} \tag{6.19}$$

Note that Equation (6.19) shows that $b'' \geq 0$, with equality only if the variance is zero or the dispersion is infinite.

- We can now reparameterise in terms of μ

$$\begin{aligned} \mathrm{Var}[Y|\theta,\phi] &= \phi\, b''(\theta) \\ &= \phi\, b''((b')^{-1}(\mu)) \\ &= \phi\, \mathcal{V}(\mu) \end{aligned} \tag{6.20}$$

- The function $\mathcal{V}(\cdot) = b''((b')^{-1}(\cdot))$ is called the *variance function*. Equations (6.16) and (6.20) make it clear why ϕ is called the "dispersion". It's value does not affect $\mu = \mathrm{E}[Y|\theta,\phi]$, but it scales $\mathrm{Var}[Y|\theta,\phi]$.

6.2.3 Examples

We now look at the results of the preceding section as applied to our three examples.

6.2.3.1 Poisson

We have

- $\theta = \log \lambda$,
- $\phi = 1$,
- $b(\theta) = e^{\theta}$.

Thus,

$$\mu = b'(\theta) = e^{\theta} \tag{6.21}$$
$$\mathcal{V}(\mu) = b''(\theta) = e^{\theta} \tag{6.22}$$

meaning that

$$\mathrm{E}[Y|\theta,\phi] = e^{\log \lambda} = \lambda \tag{6.23}$$
$$\mathrm{Var}[Y|\theta,\phi] = \phi\, e^{\log \lambda} = \lambda \tag{6.24}$$

as expected.

6.2.3.2 Bernoulli

We have

- $\theta = \log(\pi/(1-\pi))$,
- $\phi = 1$,
- $b(\theta) = \log(1 + e^{\theta})$.

Thus,

$$\mu = b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} \tag{6.25}$$
$$\mathcal{V}(\mu) = b''(\theta) = \frac{e^{\theta}}{(1+e^{\theta})^2} \tag{6.26}$$

meaning that

$$\mathrm{E}[Y|\theta,\phi] = \frac{\pi}{1-\pi}\frac{1-\pi}{1} = \pi \tag{6.27}$$
$$\mathrm{Var}[Y|\theta,\phi] = \phi\,\frac{\pi}{1-\pi}\frac{(1-\pi)^2}{1} = \pi(1-\pi) \tag{6.28}$$

as expected.

6.2.3.3 Gaussian

We have

- $\theta = \mu$
- $\phi = \sigma^2$,
- $b(\theta) = \frac{1}{2}\theta^2$.

Thus

$$\mu = b'(\theta) = \theta \tag{6.29}$$
$$\mathcal{V}(\mu) = b''(\theta) = 1 \tag{6.30}$$

meaning that

$$\mathrm{E}[Y|\theta,\phi] = \theta = \mu \tag{6.31}$$
$$\mathrm{Var}[Y|\theta,\phi] = \phi = \sigma^2 \tag{6.32}$$

as expected.

Chapter 7

Generalised Linear Models

7.1 Setting The Scene

We are given:

- Predictors $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^p$: recall that in general these will be a function of the actual covariates $\boldsymbol{z}$, for example $\boldsymbol{x} = \boldsymbol{\varphi}(\boldsymbol{z})$. The predictors are numerical; either naturally, or because they are an encoding of categorical variables.
- A response $y \in \mathcal{Y} \subseteq \mathbb{R}$. This may be numerical; continuous or discrete, or it may be binary. For now, we do not consider nominal responses with more than two values, for example $Y \in \{\text{red, green, blue}\}$[1].

We suppose that there is some functional relationship between the predictors and the response, i.e. that $P(y|\boldsymbol{x}, f) = P(y|f(\boldsymbol{x}))$, with $\mathrm{E}[Y|\boldsymbol{x}, f] = f(\boldsymbol{x})$, for some $f \in \mathcal{F}$, a suitable set of possible functions.

One class of models of functional relationships, defined via a set of possible functions $\mathcal{F}$ and a set of possible probability distributions whose means will be controlled by those functions, are the *generalised linear models (GLMs)*, which we now define.

7.2 Definition

A GLM is specified through the following components:

- A *linear predictor*:
$$\eta = \boldsymbol{\beta}^T \boldsymbol{x} \tag{7.1}$$
- An *injective response function* h, such that
$$\mu = \mathrm{E}[Y|\boldsymbol{x}, \boldsymbol{\beta}] = h(\eta) = h(\boldsymbol{\beta}^T \boldsymbol{x}) \tag{7.2}$$
Equivalently, one can write
$$g(\mu) = \boldsymbol{\beta}^T \boldsymbol{x} \tag{7.3}$$
where $g = h^{-1}$ is the *link* function.

[1] You *may* consider these next term.

- The *distributional assumption*: our knowledge of Y given $\boldsymbol{x}$ and $\boldsymbol{\beta}$ is described by an *EDF* with parameters that depend on $\boldsymbol{x}$ and $\boldsymbol{\beta}$:

$$P(y|\boldsymbol{x},\boldsymbol{\beta}) = P(y|\theta(\boldsymbol{x},\boldsymbol{\beta}),\phi(\boldsymbol{x},\boldsymbol{\beta})) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right) \tag{7.4}$$

  In addition, values of Y for different $\boldsymbol{x}$ and $\boldsymbol{\beta}$ will be assumed to be independent of each other and other $\boldsymbol{x}$ values (but not $\boldsymbol{\beta}$), that is:

$$P(\{y_i\} \mid \{\boldsymbol{x}_i\}, \boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}) \tag{7.5}$$

  where $\{y_i, i = 1, ..., n\}$ are response data given the $\{\boldsymbol{x}_i, i = 1, ..., n\}$.

7.3 The Natural/Canonical Link

There is one choice of response (or equivalently link) function that greatly simplifies the formulism. This is known as the *natural link* or *canonical link*, although in practice it does not always seem that natural at all.[2]

Recall that we have both:

$$\mu = \mathrm{E}[Y|\theta,\phi] \ = b'(\theta) \tag{7.6}$$

$$\mu = \mathrm{E}[Y|\boldsymbol{x},\boldsymbol{\beta}] = h(\boldsymbol{\beta}^T\boldsymbol{x}) = h(\eta) \tag{7.7}$$

with Equation (7.6) holding as a result of $P(y|\theta,\phi)$ following an EDF distribution, and Equation (7.7) holding by definition for a GLM. Following Equations (7.6) and (7.7), we have that

$$\theta = (b')^{-1}(\mu) = (b')^{-1}(h(\boldsymbol{\beta}^T\boldsymbol{x})) \tag{7.8}$$

The *natural link* is the choice $h = b'$, or equivalently $g = (b')^{-1}$, resulting in the equation

$$\theta = \boldsymbol{\beta}^T\boldsymbol{x} = \eta \tag{7.9}$$

It is clear that this will simplify the formulism. Other links may still be used if needed, especially since the natural link often has undesirable properties.

7.3.1 Examples

We now apply the preceding theory to our three examples.

7.3.1.1 Poisson

- From Section 6.2.3.1: $\mu = b'(\theta) = e^\theta$.
- The natural response function is thus $\mu = h(\eta) = e^\eta$.
- Natural link is therefore $\eta = g(\mu) = \log\mu$.

[2] Perhaps it would be better called the *easy* link!

7.3.1.2 Bernoulli

- From Section 6.2.3.2: $\mu = b'(\theta) = e^\theta/(1+e^\theta)$.
- The natural response function is thus $\mu = h(\eta) = e^\eta/(1+e^\eta)$.
- Natural link is therefore $\eta = g(\mu) = \log(\mu/(1-\mu))$, i.e. the logit function.

7.3.1.3 Gaussian

- From Section 6.2.3.3: $\mu = b'(\theta) = \theta$.
- The natural response function is thus $\mu = h(\eta) = \eta$.
- Natural link is therefore $\eta = g(\mu) = \mu$.

7.4 Grouped Data

We have seen that for a GLM, the expectation of the response is a function of θ only, which in turn is a function of $\boldsymbol{\beta}$ and $\boldsymbol{x}$ via $\eta = \boldsymbol{\beta}^T\boldsymbol{x}$.

In principle, the dispersion ϕ may also be a function of $\boldsymbol{x}$ or otherwise vary from data point to data point. In practice, however, it is usually assumed to be constant: e.g. for Poisson, where $\phi = 1$, or Gaussian, where $\phi = \sigma^2$. The most common exception is the case of *grouped* data, where multiple replicates may be observed for a given $\boldsymbol{x}$.

7.4.1 Grouped Binary Data

Let $\pi(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}}}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x}}}$, and suppose that there are several binary values for each $\boldsymbol{x}$, that is we have data

$$\left\{(\boldsymbol{x}_i, \{y_{ir}\}_{r\in[1..m_i]})\right\}_{i\in[1..n]} \tag{7.10}$$

where m_i is the number of replicates for *group* i, indexed by r; n is the number of groups, and $M = \sum_i m_i$ is the total sample size.

If our data consists just of the counts

$$\tilde{y}_i \triangleq \sum_{r=1}^{m_i} y_{ir} \tag{7.11}$$

then while the individual y_{ir} are Bernoulli-distributed with parameter $\pi(\boldsymbol{x}_i)$, the $\tilde{y}_i$ are binomially distributed, with parameters m_i and $\pi(\boldsymbol{x}_i)$, that is

$$\tilde{Y}_i = \sum_r Y_{ir} \sim \text{Bin}(m_i, \pi(\boldsymbol{x}_i)) \tag{7.12}$$

It turns out, however, to be more convenient to model not the counts themselves, but the proportions or averages, that is

$$Y_i \triangleq \bar{Y}_i = \frac{1}{m_i}\sum_r Y_{ir} \sim \frac{1}{m_i}\text{Bin}(m_i, \pi(\boldsymbol{x}_i)) \tag{7.13}$$

where the distribution, corresponding to a binomial variable divided by the number of *trials*, is known as the *scaled* or *rescaled binomial distribution*.

The reason that it is more convenient to use the y_i instead of the $\tilde{y}_i$ is because the expectation still takes the form:

$$\mathrm{E}[Y|m, \boldsymbol{x}] = \frac{1}{m} m\pi(\boldsymbol{x}) = \pi(\boldsymbol{x}) \tag{7.14}$$

meaning that the expectation function can still be modelled in the same way as for binary regression, using, for example, the logistic function. There is thus no need to include m at the level of the expectation, only in the distribution.[3]

We have been acting as if the model resulting from this procedure is a GLM. We have seen at the level of the expectation that this is true, but what about the distribution? Is the rescaled binomial distribution an EDF? The answer is *yes*, as follows.

$$\begin{aligned}
P(y|m,\pi) &= P(my \text{ 'successes' in } m \text{ 'trials'}|m,\pi) && (7.15)\\
&= \binom{m}{my}\pi^{my}(1-\pi)^{m-my} && (7.16)\\
&= \exp\left\{my\log\pi + (m-my)\log(1-\pi) + \log\binom{m}{my}\right\} && (7.17)\\
&= \exp\left\{m(y(\log\pi - \log(1-\pi)) + \log(1-\pi)) + c(y,\frac{1}{m})\right\} && (7.18)\\
&= \exp\left\{\frac{y\log\frac{\pi}{1-\pi} + \log(1-\pi)}{\frac{1}{m}} + c(y,\frac{1}{m})\right\} && (7.19)
\end{aligned}$$

Thus the rescaled binomial distribution is an EDF, with

$$\begin{aligned}
\theta &= \log\frac{\pi}{1-\pi} && (7.20)\\
\phi &= \frac{1}{m} && (7.21)
\end{aligned}$$

Thus θ is the same as in the Bernoulli case, while $\phi = 1$ is replaced by $\phi = 1/m$. The dispersion thus depends on m, and so if different groups i have different numbers of replicates m_i, the dispersion will vary with i:

$$\phi_i = \frac{1}{m_i} \tag{7.22}$$

7.4.2 Example: Rainfall Data

We consider rainfall data recorded in Tokyo in the first 11 days of the year 1983. For each day, we know whether it rained ($z_i = 1$) or did not rain ($z_i = 0$) on that day, as given by the table in Figure 7.1.

The goal is to find a model which predicts the probability of rain on day $i + 1$ based on the rainfall on day i, which is given by z_i. Based on this binary data, we previously defined the response $y_i = z_{i+1}$, $i = 1, \ldots, 10$, so that we could write the data in the form given by the table in Figure 7.2.

[3]Note that there is still a conditioning on known m in the expression for expectation, so this statement is different from knowing the proportion without knowing the number of replicates, which would require treating m as a nuisance parameter, either via profile likelihoods, or by assigning a prior and integrating it out.

Day i	1	2	3	4	5	6	7	8	9	10	11
z_i	0	1	1	0	1	1	0	0	0	0	0

Figure 7.1: Rainfall data in Tokyo in the first 11 days of the year 1983.

Day i	1	2	3	4	5	6	7	8	9	10
z_i	0	1	1	0	1	1	0	0	0	0
y_i	1	1	0	1	1	0	0	0	0	0

Figure 7.2: Rainfall data in Tokyo for the first 10 days of the year 1983 along with data about whether it rained the following day.

Instead of treating the response variable as a binary response to a binary covariate, with the data consisting of the values of this covariate for rainfall each day and the values of the response for rainfall the day after, we now treat the response as a rescaled binomial variable giving the proportion of response values 1 corresponding to each of the two possible values of the covariate.

The data therefore corresponds to the table in Figure 7.3.

Day i, Day $i+1$	No	Yes	
No	4	2	$m_1 = 6$
Yes	2	2	$m_2 = 4$

Figure 7.3: Grouped rainfall data.

The sample size is thus $m_1 + m_2 = M = 10$, while the number of groups $n = 2$. We can denote the two values of the covariate corresponding to the two groups as $z_1 = 0$ and $z_2 = 1$. We then have the grouped data:

$$z_1 = 0 \qquad y_1 = \frac{2}{6} = \frac{1}{3} \qquad m_1 = 6 \tag{7.23}$$

$$z_2 = 1 \qquad y_2 = \frac{2}{4} = \frac{1}{2} \qquad m_2 = 4 \tag{7.24}$$

The model will be

$$Y_i \sim \frac{1}{m_i}\text{Bin}(m_i, \pi_i) \tag{7.25}$$

where $\pi_i = \pi(x_i) = \frac{e^{\beta_1+\beta_2 z_i}}{1+e^{\beta_1+\beta_2 z_i}}$, with $i \in \{1, 2\}$. This gives rise to the probability of the data:

$$P(\{y_i\} \mid \{m_i\}, \{x_i\}, \boldsymbol{\beta}) = \prod_i \binom{m_i}{m_i y_i} \pi_i^{m_i y_i} (1-\pi_i)^{m_i - m_i y_i} \tag{7.26}$$

Estimation for this case is examined in Question 7-1.

7.4.3 General Case

If $P(y_r|\theta,\phi)$ is an EDF for each $r \in [1..m]$, with natural parameter θ, log normalizer b, dispersion ϕ, and function c, then the *grouped* data,

$$y \triangleq \bar{y} = \frac{1}{m}\sum_r y_r \tag{7.27}$$

has a probability distribution that is also an EDF. This EDF has the same natural parameter θ and log-normalizer b as the original distribution, but ϕ is replaced by $\frac{\phi}{m}$ and the function c may be different and a function of m.

In general, it is advisable to group data as far as possible because:

- it simplifies the equations.
- it improves speed of convergence and hence computation time.
- some theory only holds when $m \gg 1$.

However, grouping may be impossible for continuous predictors.

Part III

Practical Classes

Chapter 8

Practical Sheets

8.1 Practical 1 - Contingency Tables

In this practical, we consider exploring practical application of some of the techniques learnt in the lectures in R. This practical may seem long, however, some of it should be a nice refresher of R techniques already learnt, and some other parts are hopefully gentle and straightforward to read about and follow (although the odd more challenging exercise is thrown in!). You are encouraged to work through to the end in your own time. I would also like to note that *Data Science and Statistical Computing II* (DSSC II) was not a prerequisite for this course, and hence is not necessary. However, if you did take that course, you are welcome to play around with using some of the exciting skills picked up in that course in conjunction with what you learn in this course (particularly surrounding the visualisation side of things).

Finally, whilst solutions to the practical sheets will be provided, it is by attempting to answer the exercises yourself first that your practical coding skills will be developed!

8.1.1 Construction of Contingency Tables

8.1.1.1 Construction from Matrices

We here consider entering 2×2 contingency tables manually into R. We do this as follows for the data in Table 2.3.

```
DR_data <- matrix( c(41, 9,
                     37, 13 ), byrow = TRUE, ncol = 2 )
dimnames( DR_data ) <- list( Dose = c("High", "Low"),
                             Result = c("Success", "Failure") )

DR_data
```

```
##       Result
## Dose   Success Failure
##   High      41       9
##   Low       37      13
```

We can add row and column sum margins as follows

```
DR_contingency_table <- addmargins( DR_data )
DR_contingency_table
```

```
##       Result
## Dose   Success Failure Sum
##   High      41       9  50
##   Low       37      13  50
##   Sum       78      22 100
```

8.1.1.2 Contingency Tables of Proportions

Proportions can be obtained using `prop.table`.

```
DR_prop <- prop.table( DR_data )
DR_prop_table <- addmargins( DR_prop )
DR_prop_table
```

```
##       Result
## Dose   Success Failure Sum
##   High    0.41    0.09 0.5
##   Low     0.37    0.13 0.5
##   Sum     0.78    0.22 1.0
```

The row conditional proportions are derived by `prop.table(DR_data, 1)`. Analogously, column conditional proportions can be obtained using `prop.table(DR_data, 2)`.
The `addmargins` function also provides versatility for specifying margins for particular dimensions. We here specify that margins only be added for each row (trivially yielding 1 in each case as expected).

```
DR_prop_1 <- prop.table( DR_data, 1 )
DR_prop_1_table <- addmargins( DR_prop_1, margin = 2 )
DR_prop_1_table
```

```
##       Result
## Dose   Success Failure Sum
##   High    0.82    0.18   1
##   Low     0.74    0.26   1
```

We can amend the function argument `FUN` so that an alternative operation is performed. Since both rows contain the same total number of subjects, we can obtain the overall proportion in each column by amending the function argument `FUN` as follows[1]:

```
DR_prop_2_table <- addmargins( DR_prop_1_table, margin = 1, FUN = mean )
DR_prop_2_table
```

```
##       Result
```

[1]This example serves to demonstrate how to change the function `FUN` - I would like to reiterate that the `mean` function works for our purposes here (to get overall proportions for each column) *only* because the row totals are both the same.

```
## Dose   Success Failure Sum
##   High    0.82    0.18   1
##   Low     0.74    0.26   1
##   mean    0.78    0.22   1
```

Notice that

```
DR_prop_3_table <- addmargins( DR_prop_1,
                               margin = c(1,2),
                               FUN = list(mean, sum) )
```

```
## Margins computed over dimensions
## in the following order:
## 1: Dose
## 2: Result
```

```
DR_prop_3_table
```

```
##       Result
## Dose   Success Failure sum
##   High    0.82    0.18   1
##   Low     0.74    0.26   1
##   mean    0.78    0.22   1
```

yields the same result. Argument `margin` dictates the order of dimensions over which operations are applied over, and the function list `FUN` dictates which function should be applied in each case. So here, R first performs `mean` over dimension 1 (the rows, in this case `Dose`), and then `sum` over dimension 2 (the columns, in this case `Result`), as is confirmed by the additional information R fed back (setting argument `quiet=TRUE` would remove this).

8.1.1.3 Construction from Dataframes

Here we suppose we have our data collated in a dataframe that we wish to cross-classify into a contingency table. We demonstrate doing this on the penguins dataset in library `palmerpenguins` (remember to install this library using `install.packages("palmerpenguins")` if you have not already done so).

```
library(palmerpenguins)
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

You will notice that the `penguins` data is in the dataframe format known in R lingo as a `tibble`. For those of you that did not take DSSC II, a `tibble` is a user-friendly type of `dataframe` that works with the user-friendly R set of libraries `tidyverse`. For our purposes, a tibble can just be viewed as any other dataframe.

Firstly, find out information about the dataset using `?penguins`.

We are interested in whether different types of penguin typically reside on different islands, hence we wish to tabulate the dataframe as follows.

```
penguins_data <- table( Species = penguins$species, Island = penguins$island )
penguins_data
```

```
##            Island
## Species     Biscoe Dream Torgersen
##   Adelie        44    56        52
##   Chinstrap      0    68         0
##   Gentoo       124     0         0
```

In this case we may be fairly certain that there is a connection between these variables... but we can test out some techniques using this contingency table nonetheless.

8.1.1.4 Exercises

These questions involve using the contingency table from the penguin data introduced in Section 8.1.1.3.

a) Use `addmargins` to add row and column sum totals to the contingency table of penguin data.

b) Use `prop.table` to obtain a contingency table of proportions.

c) Display the column-conditional probabilities, and use `addmargins` to add the column sums as an extra row at the bottom of the matrix (note: this should be a row of 1's).

d) Suppose I want the overall proportions of penguin specie to appear in a final column on the right of the table. How would I achieve this?

8.1.2 Chi-Square Test of Independence

Run the command `chisq.test` on `DR_data` with argument `correct` set to `FALSE`.

```
chisq.test( DR_data, correct = FALSE )
```

```
##
##  Pearson's Chi-squared test
##
## data:  DR_data
## X-squared = 0.9324, df = 1, p-value = 0.3342
```

`chisq.test` runs a χ^2 test of independence, and setting the argument `correct` to `FALSE` tells R not to use continuity correction. Look at the help file for `chisq.test`, and you will see that the default is for R to use Yates' continuity correction (see Section 2.4.3.5).

```
chisq.test( DR_data )
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  DR_data
## X-squared = 0.52448, df = 1, p-value = 0.4689
```

Notice that the p-value with continuity correction is larger, as expected. The ML estimates of the expected cell frequencies can be obtained by running

```
chisq.test( DR_data )$expected
```

```
##       Result
## Dose   Success Failure
##   High      39      11
##   Low       39      11
```

8.1.2.1 Exercise

For the penguin data, apply the χ^2 test of independence between penguin specie and island of residence, and interpret the results.

8.1.3 Data Visualisation

Here we introduce several types of data visualisation methods for categorical data presented in the form of contingency tables, and apply them to the Dose-Result contingency table.

8.1.3.1 Barplots

The exercises in this section should introduce you to (or refresh your memory of, if you have seen it before) the `barplot` function, as well as refresh your memory of generic R plotting function arguments.

a) Run `barplot( DR_prop )`. What do the plots show?

b) Investigate the `density` argument of the function `barplot` by both running the commands below, and also looking in the help file.

```
barplot( DR_prop, density = 70 )
barplot( DR_prop, density = 30 )
barplot( DR_prop, density = 0 )
```

c) Add a title, and x- and y-axis labels, to the plot above.

d) Use the help file for `barplot` to find out how to add a legend to the plot.

e) How would we alter the call to `barplot` in order to view dose proportion levels conditional on result (instead of the overall proportions corresponding to each cell). You may wish to use some of the table manipulation commands from Section 8.1.1.

f) Suppose instead that we wish to display each dose level in a bar, with the proportion of successes and failures illustrated by the shading in each bar. How would we do that?

8.1.3.2 Fourfold plots

Try running the following to obtain the plot shown in Figure 8.1.

```
fourfoldplot( DR_data )
```

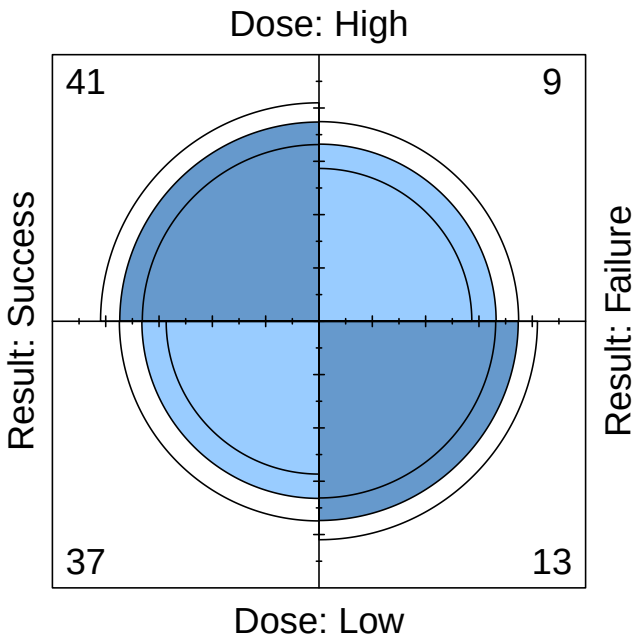


Figure 8.1: Fourfoldplot of the Dose-Result contingency table data.

A fourfold plot provides a graphical expression of the association in a 2×2 contingency table, visualising the odds ratio. Each cell entry is represented as a quarter-circle (denoted by the middle of the three rings).

We see that the shaded diagonal areas are represented by quarter-circles with greater area than the off-diagonal areas, hence the association between the two binary classification variables in positive, that is, the odds ratio r_{12} is greater than 1. The strength of association can be visually strengthened by choice of colour (although it is subjective which colour scheme is best...). For example, to obtain a red/blue colour scheme, we can run the following to obtain the plot shown in Figure 8.2

```
fourfoldplot( DR_data, color = c("red", "blue") )
```

The inner and outer rings of the quarter-circles correspond to confidence rings. The observed frequencies support the null hypothesis of no association between the variables if the rings for adjacent quarters overlap (we will explore this hypothesis test in the lectures...).

8.1.3.3 Sieve Diagrams

We here investigate the `sieve` plotting function of library `vcd`. Remember to look at the help file to help you understand the various arguments for this function.

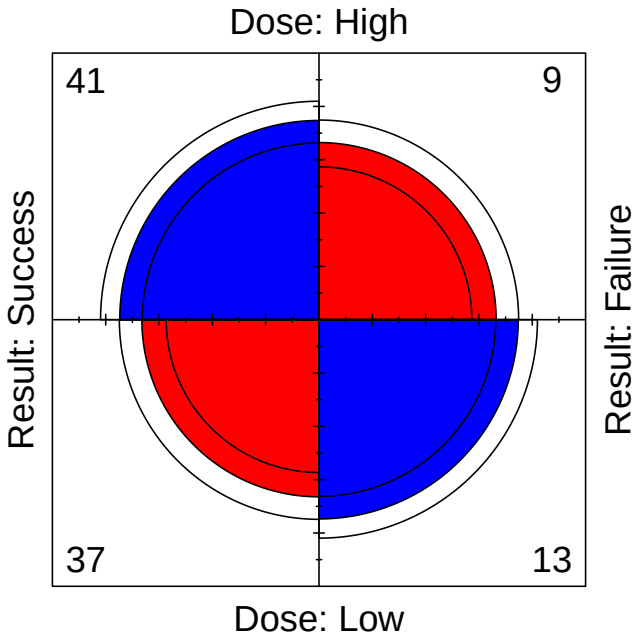


Figure 8.2: Fourfoldplot of the Dose-Result contingency table data.

a) Run

```
library(vcd)
sieve( DR_data )
```

What is shown?

b) Now run

```
library(vcd)
sieve( DR_data, shade = T )
```

Does this make the data easier or harder to visualise?

c) Finally, run

```
sieve( DR_data, sievetype = "expected", shade = T )
```

What is shown now?

8.1.3.4 Mosaic Plots

Run

```
mosaic( DR_data )
```

Mosaic plots for two-way tables display graphically the cells of a contingency table as rectangular areas of size proportional to the corresponding observed frequencies. Were the classification variables independent, then the areas would be perfectly aligned in rows and columns. The worse the alignment is, the stronger the lack of fit for independence. Furthermore, specific locations of the table that deviate from independence the most may be identified and thus the pattern of underlying association attempt to be explained.

8.1.4 Odds Ratios in R

This section seeks to test your understanding of odds ratios for 2×2 contingency tables, as well as your ability to write simple functions in R.

a) Write a function to compute the odds ratio of the success of event A with probability `pA` against the sucess of event B with probability `pB`.

b) Write a function to compute the odds ratio for a 2×2 contingency table. Test it on the Dose-response data above.

c) Will there be an issue running your function from part (b) if exactly one of the cell counts of the supplied matrix is equal to zero?

d) What about if both cells of a particular row or column of the supplied matrix are equal to zero?

e) We consider two possible options for amending the function in this case.

- i) First option: ensure that your function terminates and returns a clear error message of what has gone wrong and why when a zero would be found to be in both the numerator and denominator of the odds ratio. Hint: The command `stop` can be used to halt execution of a function and display an error message.

- ii) Second option: in the case that a row or column of zeroes is found, add 0.5 to each cell of the table before calculating the odds ratio in the usual way. Make sure that your function returns a clear warning (as opposed to error) message explaining that an alteration to the supplied table was made before calculating the odds ratio because there was a row or column of zeroes present. Hint: The command `warning()` can be used to display a warning message (but not halt execution of the function).

8.1.5 Further Exploration: Mushrooms

Consider the mushrooms data in Table 2.7 of Section 2.4.3.4 (note that this is the table **after combining cells**). Explore further the topics covered in this practical session. You will need to manually enter the data from the lecture notes as a matrix to get started.

8.2 Practical 2 - Contingency Tables

This practical gives you the opportunity to develop the techniques learnt in Practical 1 by utilising the functions presented there, as well as providing a base for exploration of new ones.

Each of the three sections below considers exploring one of three datasets. You are encouraged to explore each of these datasets using the array of techniques previously discussed, as well as utilising the presented questions and suggestions to help you learn new ones.

8.2.1 Mushrooms

We begin by returning to the mushrooms data (Table 2.7, introduced in Section 2.4.3.4). This dataset was the basis for the open-ended final exercise of Practical 1. You may have explored

this dataset in some, or all, of the following ways, amongst others:

- Manually input the data into R.
- Investigated adding relevant margins to the tables and exploring corresponding contingency tables of proportions.
- Performed a χ^2 test of independence.
- Generated visual representations of the data in the form of barplots, sieve diagrams and mosaic plots.

You are encouraged to keep exploring this dataset. In particular, the following sections prompt analysis of residuals, a GLR test, and investigation into nominal odds ratios.

8.2.1.1 Residual Analysis

Pearson and Adjusted residuals can be obtained from the output of `chisq.test()`. Use the help file for this function to find out how.

8.2.1.2 GLR Test

The function below uses the appropriate parameters provided from the output of a call to `chisq.test()` to perform a GLR test.

- Read through the code and try to understand what each line does.
- What information/results are being returned at the end of the function?
- Apply the function on the mushroom data and interpret the results.

```
G2 <- function( data ){
  # computes the G2 test of independence
  # for a two-way contingency table of
  # data: IxJ matrix
  X2 <- chisq.test( data )
  Ehat <- X2$expected
  df <- X2$parameter

  term.G2 <- data * log( data / Ehat )
  term.G2[data==0] <- 0 # Because if data == 0, we get NaN

  Gij2 <- 2 * term.G2 # Individual cell contributions to G2 statistic.
  dev_res <- sign( data - Ehat ) * sqrt( abs( Gij2 ) )
  G2 <- sum( Gij2 ) # G2 statistic
  p <- 1 - pchisq( G2, df )
  return( list( G2 = G2, df = df, p.value = p,
                Gij2 = Gij2, dev_res = dev_res ) )
}
```

8.2.1.3 Nominal Odds Ratios

In Section 8.1.4, you were encouraged to write a function to calculate the odds ratio for a 2×2 contingency table. Such a function (without concern for zeroes occurring) may be given by

```
ORmat <- function( M ){ ( M[1,1] * M[2,2] ) / ( M[1,2] * M[2,1] ) }
```

Based on this, what does the following function do? Test the function out on the mushrooms data and interpret the results.

```
nominal_OR <- function( data, ref_x = nrow( data ), ref_y =  ncol( data ) ){

  # I and J
  I <- nrow(data)
  J <- ncol(data)

  # Odds ratio matrix.
  OR_reference_IJ <- matrix( NA, nrow = I, ncol = J )
  for( i in 1:I ){
    for( j in 1:J ){
      OR_reference_IJ[i,j] <- ORmat( M = data[c(i,ref_x), c(j,ref_y)] )
    }
  }

  OR_reference <- OR_reference_IJ[-ref_x, -ref_y, drop = FALSE]

  return(OR_reference)

}
```

8.2.1.4 Visualisation and Residuals

We can add residual information to mosaic plots as follows:

```
mosaic( mushroom_data,
        gp = shading_hcl,
        residuals_type = "Pearson" )
```

Alternatively, residuals can be directly specified using the `residuals` argument (see the help file).

8.2.2 Dose-Result

Manually input into R the hypothetical data presented in Table 2.15. You may wish to attempt some or all of the following.

- Utilise previously introduced skills on this dataset.
- Amend the function presented in Section 8.2.1.3 to write two new functions

  - one to compute the set of $(I-1)\times(J-1)$ local odds ratios for $i=1,...,I-1$ and $j=1,...,J-1$; and
  - one to compute the set of $(I-1)\times(J-1)$ global odds ratios for $i=1,...,I-1$ and $j=1,...,J-1$.

- Write a function that produces an $(I-1)\times(J-1)$ matrix of fourfold plots, each corresponding to the submatrices associated with each of the $(I-1)\times(J-1)$ local odds ratios for $i=1,...,I-1$ and $j=1,...,J-1$.

- Perform a linear trend test on the data, either by writing your own code to calculate the relevant quantities, or by utilising the following function, courtesy of Kateri [2014].

```
linear.trend <- function( table, x, y ){
  # linear trend test for a 2-way table
  # PARAMETERS:
  # freq: vector of the frequencies, given by rows
  # NI: number of rows
  # NJ: number of columns
  # x: vector of row scores
  # y: vector of column scores
  # RETURNS:
  # r: Pearson's sample correlation
  # M2: test statistic
  # p.value: two-sided p-value of the asymptotic M2-test
  NI <- nrow( table )
  NJ <- ncol( table )

  rowmarg <- addmargins( table )[,NJ+1][1:NI]
  colmarg <- addmargins( table )[NI+1,][1:NJ]
  n <- addmargins( table )[NI+1,NJ+1]

  xmean <- sum( rowmarg * x ) / n
  ymean <- sum( colmarg * y ) / n
  xsq <- sqrt( sum( rowmarg * ( x - xmean )^2 ) )
  ysq <- sqrt( sum( colmarg * ( y - ymean )^2 ) )

  r <- sum( ( x - xmean ) %*% table %*% ( y - ymean ) ) / ( xsq * ysq )
  M2 = (n-1)*r^2
  p.value <- 1 - pchisq( M2, 1 )
  return( list( r = r, M2 = M2, p.value = p.value ) )
}
```

8.2.3 Titanic

Type `?Titanic` into R to learn about the `Titanic` dataset. You may wish to do this in conjuntion with one or more of the following

```
Titanic
dim( Titanic )
dimnames( Titanic )
```

We will explore this dataset further in Practical 3. For now, you are encouraged to explore associations between the variables in the contingency table. Some ideas and questions to get you started are:

- Generate some partial tables.
- Generate some marginal tables. You can do this using the function `margin.table` (look at the help file).
- Calculate partial and marginal odds ratios, and interpret the results.
- Perform a χ^2-test of independence between `Class` and `Survival`, marginalising over `Sex` and `Age`.
- Can we perform a linear trend test between `Class` and `Survival`, having marginalised over `Sex` and `Age`? If you think we can, give it a go!
- Produce a sieve or mosaic plot of the Titanic data and interpret.

8.3 Practical 3 - Contingency Tables and LLMs

8.3.1 Contingency Tables: Titanic

We revisit the `Titanic` dataset introduced in Section 8.2.3.

a) Run the following command to generate an alternative $2 \times 2 \times 4$ Titanic contingency table, called `TitanicA`. What does the command do?

```
TitanicA <- aperm( margin.table( Titanic, margin = c(1,2,4) ), c(2,3,1) )
```

b) Use the command `mantelhaen.test` to perform a Mantel-Haenszel Test on the `TitanicA` dataset for the conditional independence of `Sex` and `Survived` given `Class`.

c) Use `fourfoldplot()` to produce a matrix of fourfoldplots for the `TitanicA` dataset, each corresponding to one of the K layers for `Class`. Hint: look in the help file for `fourfoldplot` and you may find that this question is easier than you think!

d) Run the following command to obtain another alternative Titanic contingency table, called `TitanicB`. What does the command do?

```
TitanicB <- margin.table( Titanic[3:4,,,], c(1,2,4) )
```

e) Using the table `TitanicB`, calculate the marginal (over `Sex`) odds ratio between `Class` and `Survived`. Interpret the result.

f) Using the table `TitanicB`, calculate the conditional (on each level of `Sex`) odds ratio between `Class` and `Survived`. Interpret the result and compare with the result to part (e). Do you notice anything initially counter-intuitive? Explain how this comes about.

8.3.2 LLMs: Dose-Result Data

The package for fitting LLMs in R is `MASS`:

```
library(MASS)
```

We will learn how to implement LLMs in R using the Dose-Result data from Table 2.15, manually input into R in Section 9.2.2.

We fit the independence LLM to the Dose-Result data as shown. Note that this function assumes use of zero-sum constraints.

```
( I.fit <- loglm( ~ Dose + Result, data = DoseResult ) )
```

```
## Call:
## loglm(formula = ~Dose + Result, data = DoseResult)
##
## Statistics:
##                       X^2 df     P(> X^2)
## Likelihood Ratio 25.96835  4 3.211303e-05
## Pearson          25.17856  4 4.631705e-05
```

- `Likelihood Ratio` shows the G^2 GLR goodness-of-fit test against the fully saturated model of dimension q, where q is the number of variables (in this case $q = 2$).
- `Pearson` χ^2 test: in this case that arising from the χ^2 test of independence such as derived for 2-way tables in Section 2.4.3.3.

In this case, the goodness-of-fit tests suggest to reject the independence model in favour of the fully saturated model. We thus conclude that there is an association between the Dose level and treatment Result. We therefore decide to fit the saturated model, which we do as follows (note that we only need the highest level interaction terms included when we define the model, as it is assumed the LLM is hierarchical):

```
( sat.fit <- loglm( ~ Dose*Result, data = DoseResult ) )
```

```
## Call:
## loglm(formula = ~Dose * Result, data = DoseResult)
##
## Statistics:
##                   X^2 df P(> X^2)
## Likelihood Ratio   0  0        1
## Pearson            0  0        1
```

where clearly now the statistics will be 0 (as the model is being compared with itself).

The parameters' ML estimates, satisfying the zero-sum constraints, are saved in the fitted model objects under `param`.

```
( I_param <- I.fit$param )
```

```
## $`(Intercept)`
## [1] 3.493314
```

```
##
## $Dose
##       High     Medium        Low
## -0.1729852 -0.2730687  0.4460540
##
## $Result
##     Success     Partial     Failure
##  0.17502768  0.02757495 -0.20260263
```

```
( sat_param <- sat.fit$param )
```

```
## $`(Intercept)`
## [1] 3.437124
##
## $Dose
##       High     Medium        Low
## -0.2524803 -0.2488126  0.5012929
##
## $Result
##     Success     Partial     Failure
##  0.27862253  0.03096394 -0.30958647
##
## $Dose.Result
##         Result
## Dose        Success      Partial     Failure
##   High    0.3868817  0.003268524 -0.39015024
##   Medium  0.1165854 -0.128232539  0.01164718
##   Low    -0.5034671  0.124964015  0.37850305
```

- λ (`(Intercept)`) is equal to the mean log expected cell count (under the specified model).
- λ_i^X (`Dose`) is the difference between the mean log expected cell count for category i over j in $1, ..., J$ and the overall mean log expected cell count (under the specified model).
- λ_j^Y (`Result`) is the difference between the mean log expected cell count for category j over i in $1, ..., I$ and the overall mean log expected cell count (under the specified model).
- λ_{ij}^{XY} (`Dose.Result`) correspond to the differences between log expected cell count for cell (i, j) and the sum of the three contributions above (under the two-way interaction model).

For example, we can verify that

```
exp( I_param$`(Intercept)` + I_param$Dose[2] + I_param$Result[1])
```

```
##   Medium
## 29.82278
```

is equal to the expected-under-independence count for cell $(2, 1)$ of the contingency table

```
DoseResultTable[2,4] * DoseResultTable[4,1] / DoseResultTable[4,4]
```

```
## [1] 29.82278
```

which can incidentally be calculated for all cells via

```
fitted(I.fit)
```

```
## Re-fitting to get fitted values
```

```
##         Result
## Dose      Success  Partial  Failure
##   High   32.96203 28.44304 22.59494
##   Medium 29.82278 25.73418 20.44304
##   Low    61.21519 52.82278 41.96203
```

For the saturated model, we can verify that

```
exp( sat_param$`(Intercept)` + sat_param$Dose[2]
     + sat_param$Result[1] + sat_param$Dose.Result[2,1] )
```

```
## Medium
##     36
```

is precisely equal to cell $(2, 1)$ of the observed table

```
DoseResult[2,1]
```

```
## [1] 36
```

Recall that the local odds ratios can be directly derived from the interaction parameters (Equation (4.12)). For example, for cell $(2, 1)$, we have that

```
lambdaXY <- sat_param$Dose.Result
lambdaXY[2,1] + lambdaXY[3,2] - lambdaXY[2,2] - lambdaXY[3,1]
```

```
## [1] 0.873249
```

which we can verify is equivalent to cell $(2, 1)$ of the following log local OR matrix for the contingency table (obtained using the function from Solutions Section 9.2.2).

```
log( local_OR( DoseResult ) )
```

```
##           [,1]      [,2]
## [1,] 0.1387953 0.5332985
## [2,] 0.8732490 0.1136593
```

8.3.3 LLMs: Extended Dose-Result Data

Consider an extended Dose-Result dataset that now cross-classifies Treatment (`Success` or `Failure`) against presence of a prognostic factor (`Yes` or `No`) at each of six possible clinics (`A-F`), which we enter into R as follows:

```
dat <- array(c(79, 68, 5, 17,
               89, 221, 4, 46,
               141, 77, 6, 18,
               45, 26, 29, 21,
               81, 112, 3, 11,
               168, 51, 13, 12), c(2,2,6))
dimnames(dat) <- list(Treatment=c("Success","Failure"),
                      Prognostic_Factor=c("Yes","No"),
                      Clinic=c("A","B","C","D","E","F"))
```

We fit the homogeneous associations and conditional independence (of `Prognostic_Factor` and `Treatment` on `Clinic`) models as follows:

```
( hom.assoc <- loglm(~ Treatment*Prognostic_Factor
                    + Prognostic_Factor * Clinic
                    + Treatment * Clinic,
                    data=dat )  )
```

```
## Call:
## loglm(formula = ~Treatment * Prognostic_Factor + Prognostic_Factor *
##     Clinic + Treatment * Clinic, data = dat)
##
## Statistics:
##                       X^2 df  P(> X^2)
## Likelihood Ratio 7.949797  5 0.1590242
## Pearson          7.894439  5 0.1621501
```

```
( cond.ind.TF <- loglm(~ Prognostic_Factor * Clinic
                       + Treatment * Clinic, data=dat ) )
```

```
## Call:
## loglm(formula = ~Prognostic_Factor * Clinic + Treatment * Clinic,
##     data = dat)
##
## Statistics:
##                       X^2 df     P(> X^2)
## Likelihood Ratio 42.79483  6 1.280709e-07
## Pearson          41.07145  6 2.803322e-07
```

The GLR test statistic of a model against the fully saturated model can also be obtained by

```
hom.assoc$deviance
```

```
## [1] 7.949797
```

By subtracting the test statistic for $\mathcal{M}_1$ from $\mathcal{M}_0$, where $\mathcal{M}_0$ is nested in $\mathcal{M}_1$, we obtain the GLR test statistic between $\mathcal{M}_0$ and $\mathcal{M}_1$ (See Equation (4.61)).

```
( DG2 <- cond.ind.TF$deviance - hom.assoc$deviance )
```

```
## [1] 34.84504
```

The p-value for testing $\mathcal{H}_0 : \mathcal{M}_0$ against alternative $\mathcal{H}_1 : \mathcal{M}_1$ is then given by

```
( p.value <- 1 - pchisq(DG2, 1) )
```

```
## [1] 3.570184e-09
```

Finally, backwards stepwise model selection using AIC is obtained using

```
sat <- loglm(~ Prognostic_Factor * Clinic * Treatment, data=dat )
step( sat, direction="backward", test = "Chisq" )
```

```
## Start:  AIC=48
## ~Prognostic_Factor * Clinic * Treatment
##
##                                        Df   AIC    LRT Pr(>Chi)
## - Prognostic_Factor:Clinic:Treatment  5 45.95 7.9498    0.159
## <none>                                  48.00
##
## Step:  AIC=45.95
## ~Prognostic_Factor + Clinic + Treatment + Prognostic_Factor:Clinic +
##     Prognostic_Factor:Treatment + Clinic:Treatment
##
##                                Df     AIC     LRT  Pr(>Chi)
## <none>                             45.950
## - Prognostic_Factor:Treatment  1  78.795  34.845 3.570e-09 ***
## - Prognostic_Factor:Clinic     5 117.315  81.366 4.346e-16 ***
## - Clinic:Treatment             5 217.461 181.511 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Call:
## loglm(formula = ~Prognostic_Factor + Clinic + Treatment + Prognostic_Factor:Clinic +
##     Prognostic_Factor:Treatment + Clinic:Treatment, data = dat,
##     evaluate = FALSE)
##
## Statistics:
##                       X^2 df  P(> X^2)
## Likelihood Ratio 7.949788  5 0.1590247
## Pearson          7.895029  5 0.1621165
```

where inclusion of `test = "Chisq"` results in the GLR test statistic and corresponding p-value information being included in the output. In this case we stop after dropping just the three-way interaction term.

8.3.4 Log linear Models: Titanic

a) Explicitly fit several models (i.e. without using the `step` function) to the `Titanic` dataset, including the independence model and the fully saturated model. Comment on your findings.

b) Pick two distinct nested models, and calculate the G^2 likelihood ratio test statistic testing the hypothesis $\mathcal{H}_0 : \mathcal{M}_0$ against alternative $\mathcal{H}_1 : \mathcal{M}_1$.

c) Use backwards selection to fit a log-linear model using AIC. Which model is chosen? What does this tell you?

d) Use Forwards selection to fit a log-linear model using BIC (look in the help file for `step` if you are unsure how to do this). Do you get a different model to that found in part (c)? If so, is this due to the change in model selection criterion or stepwise selection direction?

8.4 Practical 4 - Binary Regression

In this practical, there are two tasks. The first tests your understanding of binary regression as discussed in the lectures, both in terms of R coding and also general understanding. The second question goes further into the R coding side of things, touching upon topics we haven't yet covered in lectures but which will be a useful taster for things to come. Before you start this practical, you may wish to review the practical example of Section 5.3.5.1.

You can find the datasets required for this practical on Ultra under *Other Material*.

8.4.1 Space Shuttle Flights

The dataset `shuttle` describes, for $n = 23$ space shuttle flights before the Challenger disaster in 1986, the temperature in $^\circ F$ at the time of take-off, as well as the occurrence or non-occurrence of a thermal overstrain of a certain component (the famous *O-ring*). The data set contains the variables:

- `flight` - flight number
- `temp` - temperature in $^\circ F$
- `td` - thermal overstrain (1 =yes, 0 =no)

You can read in the `Shuttle` data as follows.

```
## shuttle <- read.table("shuttle.asc", header=TRUE)
```

where we need *shuttle.asc* to be in the working directory (or else we set as argument to `read.table` the location of the file).

a) Fit a linear model (with the temperature as covariate) using the R function `lm`. Produce a plot which displays the data, as well as the fitted probabilities of thermal overstrain. Why does this model not provide a satisfactory description of the data?

b) Fit a logistic regression model to these data, with logit link function, using function `glm`.

c) Extract the model parameters from the logistic regression model. Write a function for the expected probability of thermal overstrain `td` in terms of `temp` explicitly (that is, $\hat{P}(\text{td}|\text{temp})$) using your knowledge of logistic regression models, and add its graph to the plot created in part (a).

d) According to this model, at what temperature is the probability of thermal overstrain exactly equal to 0.5?

e) On the day of the Challenger disaster, the temperature was $31°F$. Use your model to predict the probability of thermal overstrain on that day.

f) Replace the `logit` link by a `probit` link, and repeat the above analysis for this choice of link function.

8.4.2 US Graduate School Admissions

The dataset *binary.csv* contains data on the chances of admission to US Graduate School. We are interested in how GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution (rated from 1 to 4, highest prestige to lowest prestige), affect admission into graduate school. The response variable is a binary variable, 0 for don't admit, and 1 for admit.

The question is to find a model which relates the response to the regressor variables.

a) First read the data in as follows...

```
graduate <- read.csv("binary.csv")
```

where we need *binary.csv* to be in the working directory (or else we set as argument to `read.csv` the location of the file). Start to examine it with the following commands

```
head(graduate)
summary(graduate)
```

b) As the response is binary, a natural first step is to try logistic regression. We must first tell R that `rank` is a categorical variable (for example, we must tell R that `rank=2` is not twice as big as `rank=1`).

```
graduate$rank <- factor(graduate$rank)
graduate.glm <- glm(admit ~ gre + gpa + rank, data = graduate,
                                              family = "binomial")
```

Look at `summary(graduate.glm)`. You should see that all the terms are significant, as well as some information about the deviance and the AIC.

The logistic regression coefficients in the R output give the change in the log odds of the outcome for a one unit increase in the predictor variable. For example, for a one unit increase in GPA, the log odds of being admitted to graduate school increases by 0.804. As rank is a factor (categorical variable) the coefficients for rank have a slightly different interpretation. For example, having attended an undergraduate institution with rank of 2, versus an institution with a rank of 1, changes the log odds of admission by -0.675.

c) Use the command `exp(coef(graduate.glm))` to get the odds ratios. This will tell you, for example, that a one point increase in GRE increases the odds of admittance to Grad school by a factor of 1.0023.

d) Use the following command in order to use this model to predict the probability of admission for a student with GRE 700, GPA 3.7, to a rank 2 institution.

```
predData <- with(graduate, data.frame(gre = 700, gpa = 3.7, rank = factor(2)))
predData$prediction <- predict(graduate.glm, newdata = predData,
                               type = "response")
```

Look at the `predData` dataframe to find your predicted value.

e) We haven't formally covered AIC in the context of logistic regression, but try removing terms from the model to see if it results in a significantly worse AIC value?

f) Repeat the analysis of parts a), b), d) and e) above for the `probit` link function (note that `exp(coef(graduate.glm))` will no longer give us the odds ratios in this case). Is the model better or worse? Does the predicted admission probability of a student with GRE 700, GPA 3.7, to a rank 2 institution change?

Chapter 9

Practical Sheet Solutions

9.1 Practical 1 - Contingency Tables

9.1.1 Contingency Table Construction

We here provide solutions to the practical exercises of Section 8.1.1.4.

These questions involve using the contingency table from the penguin data introduced in Section 8.1.1.3

a) Use `addmargins` to add row and column sum totals to the contingency table of penguin data.

```
penguins_table <- addmargins( penguins_data )
penguins_table
```

```
##            Island
## Species     Biscoe Dream Torgersen Sum
##   Adelie        44    56        52 152
##   Chinstrap      0    68         0  68
##   Gentoo       124     0         0 124
##   Sum          168   124        52 344
```

b) Use `prop.table` to obtain a contingency table of proportions.

```
penguins_prop <- prop.table( penguins_data )
penguins_prop_table <- addmargins( penguins_prop )
penguins_prop_table
```

```
##            Island
## Species        Biscoe     Dream Torgersen       Sum
##   Adelie    0.1279070 0.1627907 0.1511628 0.4418605
##   Chinstrap 0.0000000 0.1976744 0.0000000 0.1976744
##   Gentoo    0.3604651 0.0000000 0.0000000 0.3604651
##   Sum       0.4883721 0.3604651 0.1511628 1.0000000
```

c) Display the column-conditional probabilities, and use `addmargins` to add the column sums as an extra row at the bottom of the matrix (note: this should be a row of 1's).

```
penguins_prop_1 <- prop.table( penguins_data, 2 )
penguins_prop_1
```

```
##            Island
## Species         Biscoe     Dream Torgersen
##   Adelie    0.2619048 0.4516129 1.0000000
##   Chinstrap 0.0000000 0.5483871 0.0000000
##   Gentoo    0.7380952 0.0000000 0.0000000
```

```
# Add margin sums only over the rows...
penguins_prop_1_table <- addmargins( penguins_prop_1, margin = 1 )
penguins_prop_1_table
```

```
##            Island
## Species         Biscoe     Dream Torgersen
##   Adelie    0.2619048 0.4516129 1.0000000
##   Chinstrap 0.0000000 0.5483871 0.0000000
##   Gentoo    0.7380952 0.0000000 0.0000000
##   Sum       1.0000000 1.0000000 1.0000000
```

```
# sum is 1 in each column as expected.
```

d) Suppose I want the overall conditional proportions of penguin specie to appear in a final column on the right of the table. How would I achieve this?

```
# One way is as follows...

# Take the row sums (sum over the columns) on the initial
# contingency table data.
penguins_prop_2_table <- addmargins( penguins_data, 2 )

# Calculate column-conditional proportions from this table.
penguins_prop_2_table <- prop.table( penguins_prop_2_table, 2 )

# Add column sums at the bottom of the table.
penguins_prop_2_table <- addmargins( penguins_prop_2_table, 1 )

penguins_prop_2_table
```

```
##            Island
## Species         Biscoe     Dream Torgersen       Sum
##   Adelie    0.2619048 0.4516129 1.0000000 0.4418605
##   Chinstrap 0.0000000 0.5483871 0.0000000 0.1976744
##   Gentoo    0.7380952 0.0000000 0.0000000 0.3604651
##   Sum       1.0000000 1.0000000 1.0000000 1.0000000
```

9.1.2 Chi-Square Test of Independence

We here provide solutions to the practical exercise of Section 8.1.2.1.

For the penguin data, apply the χ^2 test of independence between penguin specie and island of residence, and interpret the results.

```
# Apply the above tests on the Penguins data.
chisq.test( penguins_data )
```

```
##
##  Pearson's Chi-squared test
##
## data:  penguins_data
## X-squared = 299.55, df = 4, p-value < 2.2e-16
```

```
# Unsurprisingly the p-value is miniscule.
```

9.1.3 Barplots

We here provide solutions to the practical exercises of Section 8.1.3.1.

a) Run `barplot( DR_prop )`. What does the plot show?

```
barplot( DR_prop )
```

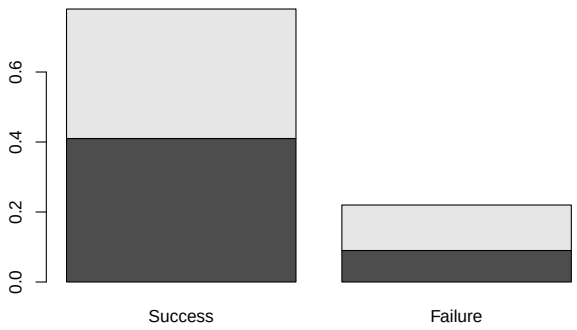


Figure 9.1: Barplots of the Dose-Result contingency table data.

We see the pair of barplots shown in Figure 9.1. In R, a matrix input to boxplot results in the column categories (in this case, `Result`) defining the bars while the row categories (`Dose`) form the stacked levels.

b) Investigate the `density` argument of the function `barplot` by both running the commands below, and also looking in the help file.

```
par(mfrow=c(1,3))
barplot( DR_prop, density = 70 )
barplot( DR_prop, density = 30 )
barplot( DR_prop, density = 0 )
```

As illustrated in Figure 9.2, we can see that the `density` argument of `boxplot` changes the level of shading for the (two) categories defined by the rows (`Dose`) of the contingency table

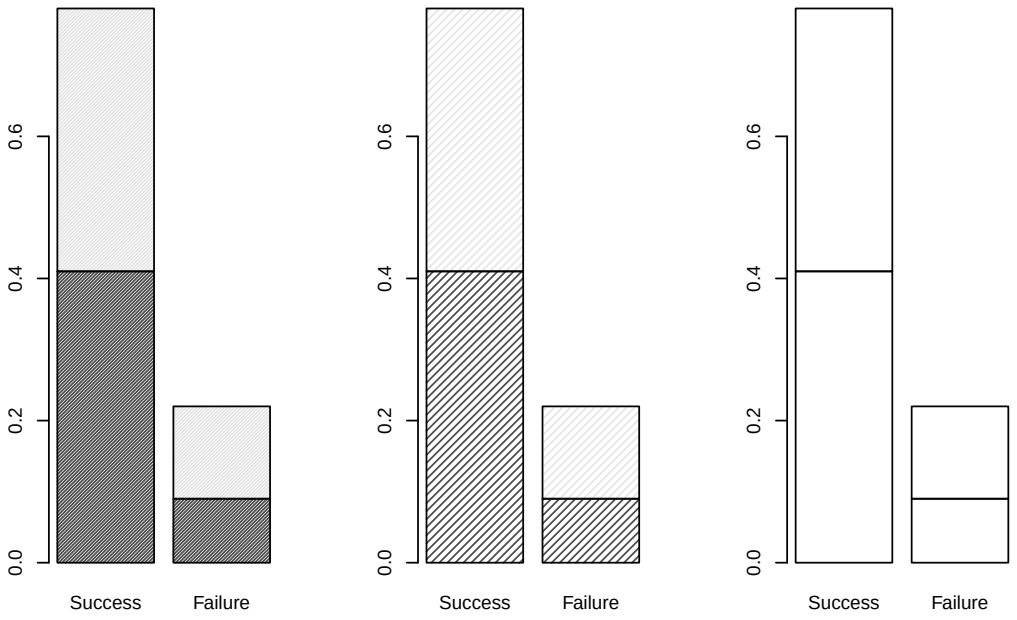


Figure 9.2: Barplots of the Dose-Result contingency table data, investigating the density parameter.

matrix.

c) Add a title, and x- and y-axis labels, to the plot above.

We can achieve this with the code below, which yields the plot shown in Figure 9.3.

```
barplot( DR_prop, density = 30, main = "Comparison of Dose by Response",
         xlab = "treatment outcome", ylab = "proportions" )
```

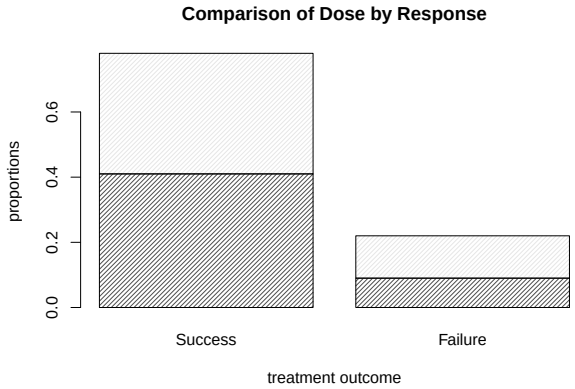


Figure 9.3: Barplots of the Dose-Result contingency table data, now with title and axis labels.

d) Use the help file for `barplot` to find out how to add a legend to the plot.

We set the argument `legend.text` to `T`. The code below thus yields the plot shown in Figure 9.4.

```
barplot( DR_prop, density = 30, legend.text = T,
         main = "Comparison of Dose by Response",
         xlab = "treatment outcome", ylab = "proportions" )
```

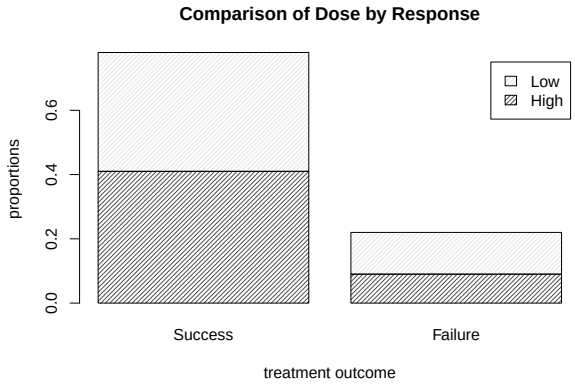


Figure 9.4: Barplots of the Dose-Result contingency table data, now with a legend.

e) How would we alter the call to `barplot` in order to view dose proportion levels conditional on result (instead of the overall proportions corresponding to each cell). You may wish to use some of the table manipulation commands from Section 8.1.1.

One possible solution is shown below, which yields the plot shown in Figure 9.5. Note that we can move the `legend` box using `args.legend` and setting the usual `x` and `y` location arguments for a legend within a list (that is, if we don't want it covering up part of the plot itself). The exact necessary values of `x` and `y` depend on your plotting window.

```
barplot( prop.table( DR_data, 2 ), density = 30, legend.text = T,
         main = "Comparison of Dose by Response",
         xlab = "treatment outcome", ylab = "proportions",
         args.legend = list( x = 2.5, y = 1.25 )  )
```

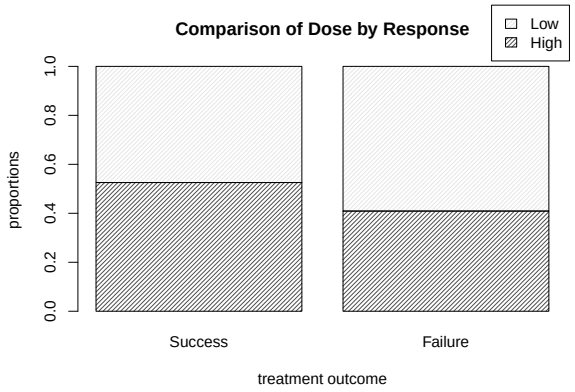


Figure 9.5: Barplots of dose level proportions corresponding to each treatment outcome.

f) Suppose instead that we wish to display each dose level in a bar, with the proportion of successes and failures illustrated by the shading in each bar. How would we do that?

We can alter the code as shown below, which runs `prop.table()` on the transposed version of the matrix `DR_data`, and yields the plot shown in Figure 9.6.

```
barplot( prop.table( t( DR_data ), 2 ), density = 30, legend.text = T,
         main = "Comparison of Response by Dose",
         xlab = "dose level", ylab = "proportions",
```

```
        args.legend = list( x = 2.6, y = 1.25 ) )
```

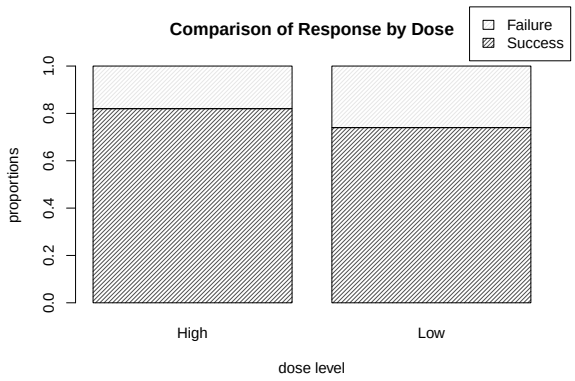


Figure 9.6: Barplots of treatment outcome proportions corresponding to each dose level.

This is of course the more informative way to display the data, as we are likely more interested in the relative proportion of `high` dose successes to `low` dose successes, rather than, say, the proportion of `successes` which happened to be `high` dose, for example.

9.1.4 Sieve Diagrams

We here provide solutions to the practical exercises of Section 8.1.3.3.

a) Run

```
library(vcd)
sieve( DR_data )
```

What is shown?

We see the plot given in Figure 9.7. A sieve (or parquet) diagram represents for an $I \times J$ table the expected frequencies under independence as a collection of IJ rectangles, each containing a set of smaller squares/rectangles. Each of the larger rectangles has height and width proportional to the corresponding row and column marginal frequencies of the contingency table respectively. This way, the area of each rectangle is proportional to the expected-under-independence frequency for the corresponding cell. The number of smaller squares/rectangles in each larger rectangle corresponds to the observed frequency, hence a larger number of smaller squares indicate that the observed frequency was larger.

b) Now run

```
sieve( DR_data, shade = T )
```

Does this make the data easier or harder to visualise?

We now see the plot given in Figure 9.8, which is the same plot as Figure 9.7, however now the observed frequency squares are coloured blue (undashed) if the observed frequency is larger than the expected frequency, and red (dashed) if the observed frequency is smaller than the expected frequency.

c) Finally, run

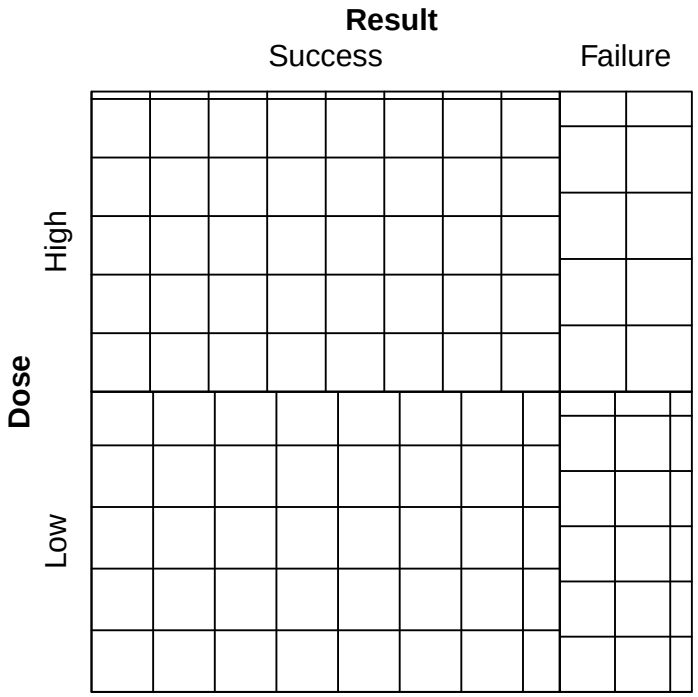


Figure 9.7: Sieve diagram comparing observed frequencies in relation to expected-under-independence frequencies for the Dose-Result data.

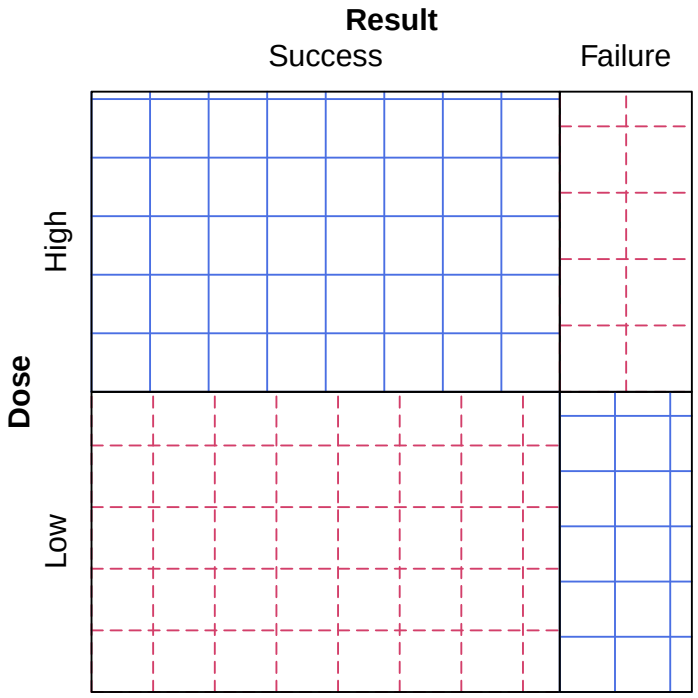


Figure 9.8: Sieve diagram comparing observed frequencies in relation to expected-under-independence frequencies for the Dose-Result data.

```
sieve( DR_data, sievetype = "expected", shade = T )
```

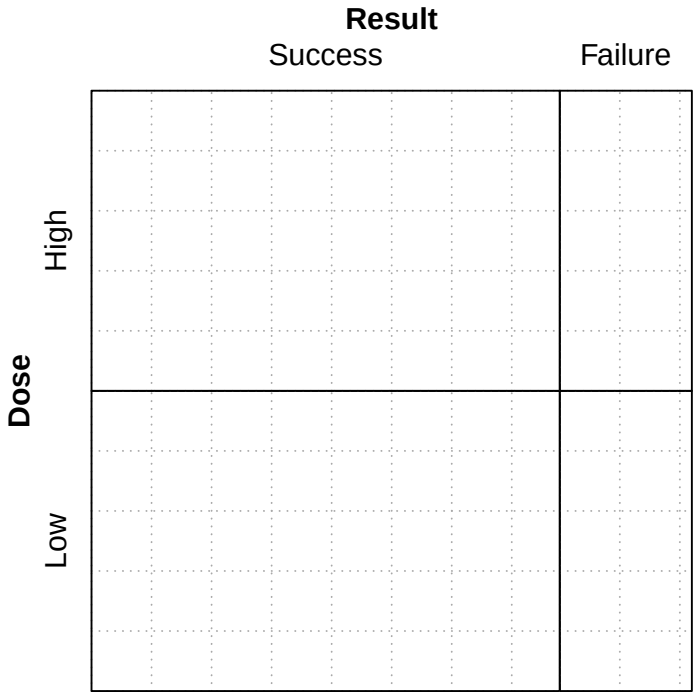


Figure 9.9: Sieve diagram showing expected-under-independence frequencies for the Dose-Result data.

What is shown now?

The plot shown in Figure 9.9, which just shows the expected frequencies under independence for each cell.

9.1.5 Odds Ratios in R

We here provide solutions to the practical exercises of Section 8.1.4.

This section seeks to test your understanding of odds ratios for 2×2 contingency tables, as well as your ability to write simple functions in R.

a) Write a function to compute the odds ratio of the success of event A with probability `pA` against the sucess of event B with probability `pB`.

```
OR <- function(pA, pB){(pA/(1-pA))/(pB/(1-pB))}
OR( 1/3, 1/2 ) # test the function.
```

```
## [1] 0.5
```

b) Write a function to compute the odds ratio for a 2×2 contingency table. Test it on the Dose-response data above.

```
OR_cont_tab <- function( M ){ ( M[1,1] * M[2,2] ) / ( M[1,2] * M[2,1] ) }
OR_cont_tab( DR_data )
```

```
## [1] 1.600601
```

```
# M is a 2 by 2 matrix representing contingency table cell counts.
```

c) Will there be an issue running your function from part (b) if exactly one of the cell counts of the supplied matrix is equal to zero?

Fine in this case as either `0` is returned if a cell count of 0 appears in the numerator of the odds ratio, or `Inf` is returned if a cell count of 0 is in the denominator of the odds ratio.

d) What about if both cells of a particular row or column of the supplied matrix are equal to zero?

We have a problem as there will be a zero in both numerator and denominator of the odds ratio, which is undefined. For example, run...

```
AB <- matrix( c(3,4,0,0), ncol = 2 ) # Define such a matrix.
OR_cont_tab( AB ) # NaN returned.
```

```
## [1] NaN
```

e) We consider two possible options for amending the function in this case.

- i) First option: ensure that your function terminates and returns a clear error message of what has gone wrong and why when a zero would be found to be in both the numerator and denominator of the odds ratio. Hint: The command `stop` can be used to halt execution of a function and display an error message.

```
OR_cont_tab_1 <- function( M ){

  M_row_sum <- rowSums( M )
  M_col_sum <- colSums( M )

  if( prod( M_row_sum ) == 0 | prod( M_col_sum ) == 0 ){
    stop( "At least one row sum or column sum of M is equal
          to zero, hence the odds ratio is undefined." )
  }
  else{
    ( M[1,1] * M[2,2] ) / ( M[1,2] * M[2,1] )
  }
}

OR_cont_tab_1( DR_data )  # works fine with DR_data
```

```
## [1] 1.600601
```

```
OR_cont_tab_1( AB ) # Execution is halted and our error returned.
```

```
## Error in OR_cont_tab_1(AB): At least one row sum or column sum of M is equal
##           to zero, hence the odds ratio is undefined.
```

- ii) Second option: in the case that a row or column of zeroes is found, add 0.5 to each cell of the table before calculating the odds ratio in the usual way. Make sure that

your function returns a clear warning (as opposed to error) message explaining that an alteration to the supplied table was made before calculating the odds ratio because there was a row or column of zeroes present. Hint: The command `warning()` can be used to display a warning message (but not halt execution of the function).

```
OR_cont_tab_2 <- function( M ){

  M_row_sum <- rowSums( M )
  M_col_sum <- colSums( M )

  if( prod( M_row_sum ) == 0 | prod( M_col_sum ) == 0 ){
    M <- M + 0.5 * matrix( 1, nrow = 2, ncol = 2 )
    OR <- ( M[1,1] * M[2,2] ) / ( M[1,2] * M[2,1] )
    warning( "At least one row sum or column sum of the supplied
             matrix M was equal to zero, hence an amendment of 0.5 was added
             to the value of each sum prior to calculating the odds ratio." )
    return( OR )
  }
  else{
    ( M[1,1] * M[2,2] ) / ( M[1,2] * M[2,1] )
  }
}

OR_cont_tab_2( DR_data ) # works fine with DR_data
```

```
## [1] 1.600601
```

```
OR_cont_tab_2( AB ) # Alternative odds ratio calculated and warning returned.
```

```
## Warning in OR_cont_tab_2(AB): At least one row sum or column sum of the supplied
##              matrix M was equal to zero, hence an amendment of 0.5 was added
##              to the value of each sum prior to calculating the odds ratio.
```

```
## [1] 0.7777778
```

9.1.6 Mushrooms

We here provide some possible analysis of the mushrooms data, with comments made in R, to get you started. The resulting barplots and sieve diagram are shown in Figures 9.10 and 9.11 respectively.

```
# We create a matrix with the data in.
mushroom_data <- matrix( c(101, 399, 57, 487,
                           12, 389, 150, 428 ), byrow = TRUE, ncol = 4 )

# Add dimension names as follows.
dimnames( mushroom_data ) <- list( Edibility = c("Edible", "Poisonous"),
                                   Cap_Shape = c("bell", "flat", "knobbed",
```

```
                                                "convex/conical") )

# Have a look.
mushroom_data
```

```
##            Cap_Shape
## Edibility   bell flat knobbed convex/conical
##   Edible     101  399      57            487
##   Poisonous   12  389     150            428
```

```
# Let's look at the proportion of each shape of mushroom that are
# edible or poisonous.
mushroom_table <- addmargins( mushroom_data, 2, FUN = mean )
mushroom_table <- prop.table( mushroom_table, 2 )
mushroom_table <- addmargins(mushroom_table, 1 )
mushroom_table
```

```
##            Cap_Shape
## Edibility        bell      flat   knobbed convex/conical      mean
##   Edible    0.8938053 0.5063452 0.2753623      0.5322404 0.5160652
##   Poisonous 0.1061947 0.4936548 0.7246377      0.4677596 0.4839348
##   Sum       1.0000000 1.0000000 1.0000000      1.0000000 1.0000000
```

```
# Let's perform a chi-square test.
chisq.test( mushroom_data )
```

```
##
##  Pearson's Chi-squared test
##
## data:  mushroom_data
## X-squared = 113.84, df = 3, p-value < 2.2e-16
```

```
# This matches the result found manually in the lectures.
```

```
# Barplot
barplot( prop.table( mushroom_data, margin = 2 ), density = 50,
         main = "Comparison of Edibility by Cap Shape", cex.main = 0.8,
         xlab = "treatment outcome", ylab = "proportions",
         legend.text = T, args.legend = list( x = 5.1, y = 1.25 ) )
```

```
# Sieve diagram
sieve( mushroom_data  )
```

```
# Much can be drawn form this diagram.  For example,
# we can see that there are far more edible bell mushrooms in our
# sample than would be expected under an assumption of independence.
```

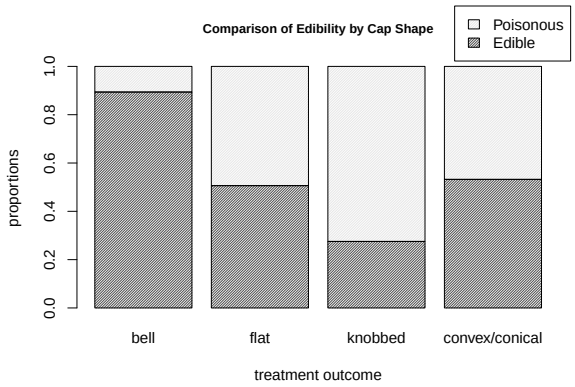


Figure 9.10: Barplots for the mushroom data of edibility proportions corresponding to each cap shape.

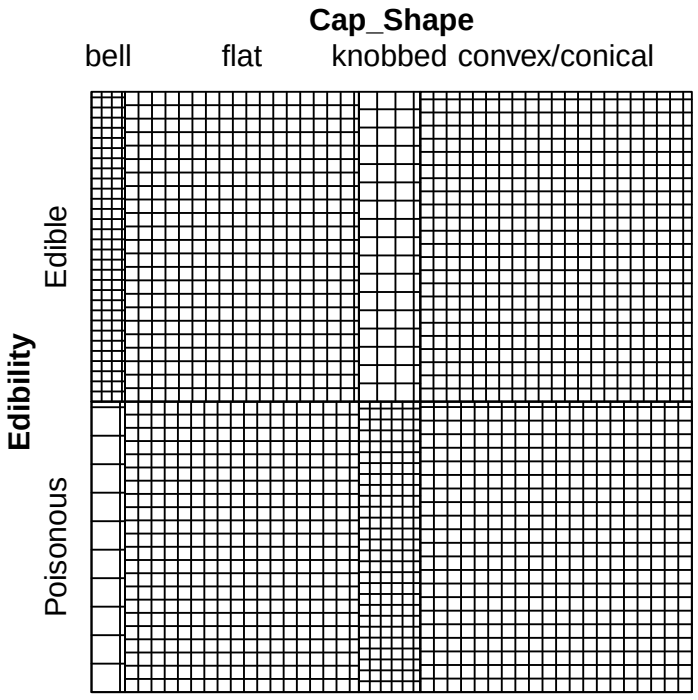


Figure 9.11: Sieve diagram showing expected-under-independence frequencies for the mushroom data.

9.2 Practical 2 - Contingency Tables

Practical 2 is designed to be exploratory, hence it is not possible to provide comprehensive *solutions*. We discuss some analysis of each dataset in accordance with the suggestions provided.

9.2.1 Mushrooms

Inputting the mushrooms data and some exploratory analysis was presented in Section 9.1.6.

We can assign the result of the χ^2 test to a named object as follows

```
chisq_Mush <- chisq.test( mushroom_data )
chisq_Mush
```

```
##
##  Pearson's Chi-squared test
##
## data:  mushroom_data
## X-squared = 113.84, df = 3, p-value < 2.2e-16
```

This matches the result found manually in the lectures. Very strong evidence to reject the null hypothesis of independence between edibility and cap shape.

9.2.1.1 Residual Analysis

We can use the χ^2 test result to obtain Pearson residuals...

```
chisq_Mush$residuals
```

```
##            Cap_Shape
## Edibility        bell       flat   knobbed convex/conical
##   Edible     5.589590 -0.3798221 -4.820746      0.6810948
##   Poisonous -5.772167  0.3922285  4.978209     -0.7033419
```

...and also Adjusted residuals.

```
chisq_Mush$stdres
```

```
##            Cap_Shape
## Edibility        bell       flat   knobbed convex/conical
##   Edible     8.269282 -0.6987975 -7.314099       1.322946
##   Poisonous -8.269282  0.6987975  7.314099      -1.322946
```

Both residuals suggest that bell-capped mushrooms and knobbed-capped mushrooms contribute most towards the X^2 statistic. For example, there are far more edible bell-capped mushrooms than would be expected if the two variables were independent.

9.2.1.2 GLR Test

We can apply the G^2 GLR test on the mushroom data as follows.

```
G2mush <- G2( mushroom_data )
G2mush$p.value
```

```
## [1] 0
```

```
G2mush$dev_res
```

```
##            Cap_Shape
## Edibility       bell      flat   knobbed convex/conical
##   Edible    10.53326 -3.895337 -8.462185       5.482674
##   Poisonous -6.03326  3.933403 11.005295      -5.394469
```

```
G2mush$Gij2
```

```
##            Cap_Shape
## Edibility        bell      flat   knobbed convex/conical
##   Edible    110.94946 -15.17365 -71.60858       30.05971
##   Poisonous -36.40022  15.47166 121.11651      -29.10030
```

The p-value is close to 0, as expected, hence strong evidence to reject $\mathcal{H}_0$. The deviance residuals are as calculated in lectures. Due to the relation between deviance residuals and the G^2 statistic, classes of individual variables with large marginal sums of deviance residual values contribute most towards the G^2 statistic. In this case, we again notice that the bell cap-shaped mushrooms have a larger contribution toward the G^2 statistic, as evidenced by the large residual values, and also the large marginal sum in residual values over the two classes of edibility for bell-shaped caps.

9.2.1.3 Nominal Odds Ratios

Odds ratios for cells $(1,1)$, $(1,2)$ and $(1,3)$ to reference cell (I,J), as calculated in lectures, are obtained by.

```
nominal_OR( mushroom_data )
```

```
##          [,1]      [,2]     [,3]
## [1,] 7.396988 0.9014427 0.333963
```

9.2.2 Dose-Result Data

Input of data can be performed as follows:

```
DoseResult <- matrix( c(47, 25, 12,
                        36, 22, 18,
                        41, 60, 55), byrow = TRUE, ncol = 3 )

dimnames( DoseResult ) <- list( Dose=c("High",
                                       "Medium",
                                       "Low"),
                                Result=c("Success",
                                         "Partial",
```

```
                                          "Failure") )

DoseResultTable <- addmargins(DoseResult)
```

Functions for local and global odds ratios respectively, then applied on the Dose-Result data to produce the results obtained in lectures, are:

```
local_OR <- function( DR ){

  # I and J
  I <- nrow(DR)
  J <- ncol(DR)

  # Odds ratio matrix.
  OR_local <- matrix( NA, nrow = I-1, ncol = J-1 )
  for( i in 1:(I-1) ){
    for( j in 1:(J-1) ){
      OR_local[i,j] <- ORmat( M = DR[c(i,i+1), c(j,j+1)] )
    }
  }

  return(OR_local)

}

global_OR <- function( DR ){

  # I and J
  I <- nrow(DR)
  J <- ncol(DR)

  # Odds ratio matrix.
  OR_global <- matrix( NA, nrow = I-1, ncol = J-1 )
  for( i in 1:(I-1) ){
    for( j in 1:(J-1) ){
      OR_global[i,j] <- ORmat( M = matrix( c( sum( DR[1:i,1:j] ),
                                              sum( DR[(i+1):I,1:j] ),
                                              sum( DR[1:i,(j+1):J] ),
                                              sum( DR[(i+1):I,(j+1):J] ) ),
                                              nrow = 2 ) )
    }
  }

  return(OR_global)

}
```

```
local_OR( DoseResult )
```

```
##          [,1]     [,2]
## [1,] 1.148889 1.704545
## [2,] 2.394678 1.120370
```

```
global_OR( DoseResult )
```

```
##          [,1]     [,2]
## [1,] 2.557038 2.754717
## [2,] 3.023440 2.359736
```

A plot of multiple fourfold plots can be obtained using the following function, then applied on the Dose-Result data to produce the plots shown in Figure 9.12.

```
ffold_local <- function ( data ){

  # I and J
  I <- nrow(data)
  J <- ncol(data)

  par( mfrow = c(I-1, J-1) )
  for( i in 1:(I-1) ){
    for( j in 1:(J-1) ){
      sub_data <- data[c(i,i+1),c(j,j+1)]
      fourfoldplot( sub_data )
    }
  }

}

# Apply fourfold local OR plot function.
ffold_local( DoseResult )
```

To compute the statistic for the linear trend test manually, we could...

```
score_x <- 1:3
score_y <- 1:3

score_xy <- crossprod( t(score_x), t(score_y) )
n <- sum(DoseResult)

sum_xy <- sum( score_xy * DoseResult )
sum_x <- sum( score_x * rowSums(DoseResult) )
sum_y <- sum( score_y * colSums(DoseResult) )
sum_x2 <- sum( score_x^2 * rowSums(DoseResult) )
sum_y2 <- sum( score_y^2 * colSums(DoseResult) )

rxy <- ( n * sum_xy - sum_x * sum_y ) / ( sqrt( n * sum_x2 - sum_x^2 ) *
```

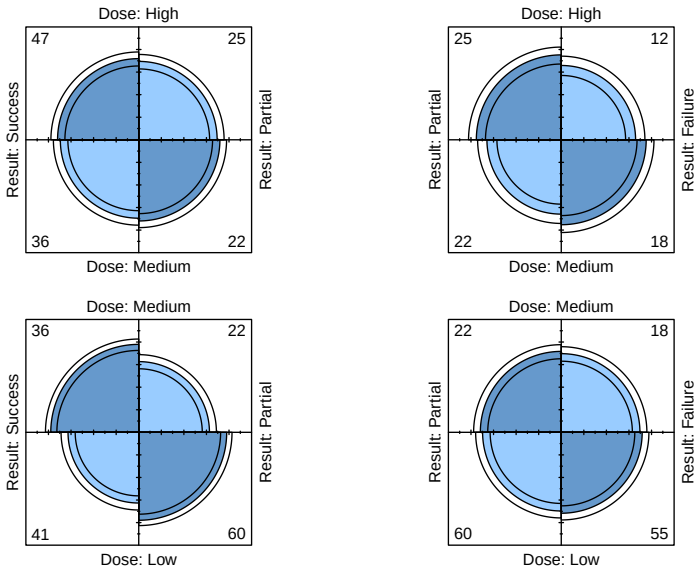


Figure 9.12: Four-fold plots of Dose-Result data

```
                                          sqrt( n * sum_y2 - sum_y^2 ) )

M2 <- (n-1) * rxy^2

p.value <- 1-pchisq(M2,1)
```

Or using the function provided we can simply do

```
linear.trend( table = DoseResult, x = score_x, y = score_y )
```

```
## $r
## [1] 0.2709127
##
## $M2
## [1] 23.11902
##
## $p.value
## [1] 1.522773e-06
```

yielding the same result. The test provides evidence to reject the null hypothesis of independence between the two variables.

9.2.3 Titanic

A partial table cross-classifying `Class` and `Sex` for `Age = Child` and `Survived = No` can be obtained in the usual way of selecting particular elements of an array.

```
Titanic[,, Age="Child", Survived="No"]
```

```
##       Sex
## Class  Male Female
```

```
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
```

Marginal tables are produced using `margin.table()` with argument `margin` denoting the variables of interest (that is, those to still be included in the table). For example, a marginal table of `Class` and `Sex` summing over margins 3 and 4 (`Age` and `Survived`, thus ignoring them), and a marginal vector for `Survived`, are obtained as follows

```
margin.table( Titanic, margin = c(1,2) )
```

```
##       Sex
## Class  Male Female
##   1st   180    145
##   2nd   179    106
##   3rd   510    196
##   Crew  862     23
```

```
margin.table( Titanic, margin = 4 )
```

```
## Survived
##   No  Yes
## 1490  711
```

We can use previous odds ratio functions on relevant 2×2 marginal/partial tables. Note that whether we can treat `Class` as ordinal is debatable, depending on how category `Crew` might fit into that. If we feel that we can treat `Class` as ordinal, then global Odds ratios might also be applicable. We calculate nominal, local and global marginal odds ratios for `Class` and `Sex` having marginalised over `Age` and `Survived`.

```
nominal_OR( margin.table( Titanic, margin = c(1,2) ) )
```

```
##            [,1]
## [1,] 0.03312265
## [2,] 0.04505757
## [3,] 0.06942800
```

```
local_OR( margin.table( Titanic, margin = c(1,2) ) )
```

```
##           [,1]
## [1,] 0.7351185
## [2,] 0.6489826
## [3,] 0.0694280
```

```
global_OR( margin.table( Titanic, margin = c(1,2) ) )
```

```
##            [,1]
## [1,] 0.26012139
## [2,] 0.22830253
## [3,] 0.05187198
```

The nominal marginal odds ratios reflect the fact the proportion of men (or more, precisely, the odds that a randomly selected person of a particular `Class` is a man) is much lower for all other levels of `Class` relative to `Crew`. Both local and global marginal odds ratios provide evidence that there is an association between `Class` and `Sex`, particularly for `Crew` in comparison with any of the other levels. For calculating global odds ratios, we here treat `Crew` as the fourth category of an ordinal variable `Class` following on from `3rd`. This position in the ordinal ranking (as opposed to being otherwise positioned in the `Class` ordering) vastly influences the global odds ratios (but less so the local ones, which are applicable for nominal variables anyway).

A Chi square test for independence of `Class` and `Survival`, marginalising over `Sex` and `Age`, is obtained by

```
chisq.test( margin.table( Titanic, margin = c(1,4) ) )
```

```
##
##  Pearson's Chi-squared test
##
## data:  margin.table(Titanic, margin = c(1, 4))
## X-squared = 190.4, df = 3, p-value < 2.2e-16
```

```
# An unsurprising overwhelming amount of evidence
# that these two are associated.
```

A linear trend test - applicability depends if we feel that we can treat `Class` as an ordinal variable.

```
linear.trend( table = margin.table( Titanic, margin = c(1,4) ),
              x = 1:4, y = 1:2 )
```

```
## $r
## [1] -0.2713954
##
## $M2
## [1] 162.042
##
## $p.value
## [1] 0
```

A sieve and mosaic plot could be obtained by running the following commands.

```
sieve( Titanic )
mosaic( Titanic )
```

9.3 Practical 3 - Contingency Tables and LLMs

9.3.1 Titanic

a) The command created a sub-table by marginalising over `Age` and then permuting the dimension order, so that we now have a $2 \times 2 \times 4$ contingency table array.

b) Mantel Haenszel Test carried out as follows (continuity correction can be applied similar to previous tests).

```
mantelhaen.test( TitanicA, correct = FALSE )
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  TitanicA
## Mantel-Haenszel X-squared = 362.67, df = 1, p-value < 2.2e-16
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##   8.232629 14.185153
## sample estimates:
## common odds ratio
##          10.80653
```

An unsurprisingly very small p-value giving evidence that `Sex` and `Survived` are not conditionally independent given `Class`.

c) The Fourfold plot given in Figure 9.13 is generated as follows.

```
par(mfrow = c(1,1))
fourfoldplot( TitanicA )
```

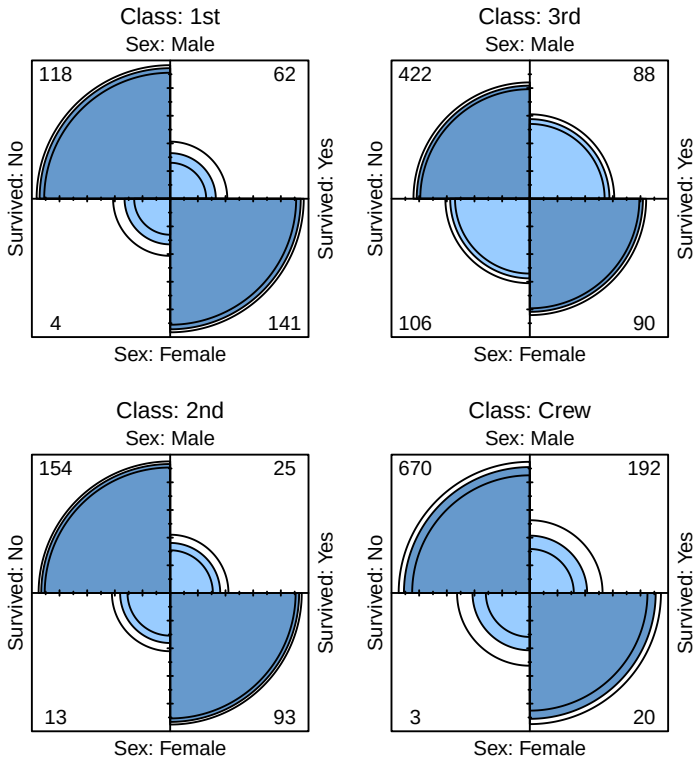


Figure 9.13: Four-fold plots of Titanic data

Visual evidence that `Sex` and `Survived` are not conditionally independent given `Class`.

d) The submatrix marginalises out `Age` and only considers interest to lie in subjects in either of the two categories `3rd` or `Crew` for `Class`. Note that we do this because parts (e) and (f) consider the question of whether subjects in 3rd class or crew had a greater probability of survival.

e) We marginalise over `Sex` as follows and calculate the odds ratio:

```
TitanicB_margin <- margin.table( TitanicB, c(1,3) )
ORmat( TitanicB_margin )
```

```
## [1] 0.9344041
```

This suggests that `3rd` class passengers had greater odds of survival than did `Crew`. (Note that the Odds Ratio is defined in terms of the odds of "Surviving: No" in this case).

f) Now condition on gender and calculate the odds ratios.

```
ORmat( TitanicB[,1,] )
```

```
## [1] 1.37422
```

```
ORmat( TitanicB[,2,] )
```

```
## [1] 7.851852
```

Both of these are positive, suggesting that both `Male` and `Female Crew` had greater odds of survival than did `Male` or `Female 3rd` class passengers respectively.

By looking at the appropriate marginal and conditional tables...

```
prop.table( TitanicB_margin, margin = 1 )
```

```
##       Survived
## Class         No       Yes
##   3rd  0.7478754 0.2521246
##   Crew 0.7604520 0.2395480
```

```
prop.table( TitanicB[,1,], margin = 1 )
```

```
##       Survived
## Class         No       Yes
##   3rd  0.8274510 0.1725490
##   Crew 0.7772622 0.2227378
```

```
prop.table( TitanicB[,2,], margin = 1 )
```

```
##       Survived
## Class         No       Yes
##   3rd  0.5408163 0.4591837
##   Crew 0.1304348 0.8695652
```

we can see that indeed the proportion of `3rd` class passengers surviving overall was greater than that for `Crew`, however, when we condition on gender, a higher proportion of `Male` and `Female Crew` survived than did `Male` or `Female 3rd` Class passengers respectively.

This apparent contradiction occurs because the relationship between survival and class is influenced by the confounding variable `Sex`, which we ignore in the table marginalising over `Sex`.
The conditional probability tables show that the survival rate was much lower for men compared to women on board the Titanic (perhaps because of the *"women and children first"* policy) when it came to getting a seat on a lifeboat. If we compare the row-proportional marginal table over survival...

```
prop.table( margin.table( TitanicB, margin = c(1,2) ), margin = 1 )
```

```
##       Sex
## Class       Male    Female
##   3rd  0.7223796 0.2776204
##   Crew 0.9740113 0.0259887
```

... this shows that there were a greater proportion of women in third class than in the crew – more than a quarter of third class passengers were women whereas less than 3% of the crew were Female. In this case, the difference in the proportion of women amongst the third class passengers and crew has a substantial impact on the overall survival rate for each group and has the unfortunate effect of masking the third class passengers vs crew effect when gender is ignored in the original table. Hence we have an example of Simpson's Paradox, where trends within sub-populations can be reversed when the data are aggregated.

9.3.2 Log linear Models: Titanic

a) Explicitly fit several models (i.e. without using the `step` function), including the independence model and the fully saturated model. Comment on your findings.

The independence model is fitted as follows:

```
( Titanic.I.fit <- loglm( ~ Class + Survived + Sex + Age, data=Titanic ) )
```

```
## Call:
## loglm(formula = ~Class + Survived + Sex + Age, data = Titanic)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 1243.663 25        0
## Pearson          1637.445 25        0
```

Unsurprisingly, very strong evidence that `Class`, `Sex` and `Survived` are not all mutually independent.

The fully saturated model is fitted as follows:

```
( Titanic.sat.fit <- loglm( ~ Class*Survived*Sex*Age, data=Titanic ) )
```

```
## Call:
## loglm(formula = ~Class * Survived * Sex * Age, data = Titanic)
##
## Statistics:
```

```
##                   X^2 df P(> X^2)
## Likelihood Ratio   0  0        1
## Pearson          NaN  0        1
```

Note the `NaN` appears under `Pearson` because the chi-squared statistic is ill-defined if there are expected values equal to zero.

b) Pick two distinct nested models, and calculate the G^2 likelihood ratio test statistic testing the hypothesis $\mathcal{H}_0 : \mathcal{M}_0$ against alternative $\mathcal{H}_1 : \mathcal{M}_1$.

Let's consider $\mathcal{H}_0 : \mathcal{M}_0$ to be the *all two-way interactions* model and the $\mathcal{H}_1 : \mathcal{M}_1$ to be the *all three-way interactions* model. Then we have that:

```
( Three.fit <- loglm( ~ Class*Survived*Sex
                      + Class*Survived*Age
                      + Class*Sex*Age
                      + Survived*Sex*Age, data = Titanic ) )
```

```
## Call:
## loglm(formula = ~Class * Survived * Sex + Class * Survived *
##     Age + Class * Sex * Age + Survived * Sex * Age, data = Titanic)
##
## Statistics:
##                            X^2 df  P(> X^2)
## Likelihood Ratio 0.0001904416  3 0.9999993
## Pearson                   NaN  3       NaN
```

```
( Two.fit <- loglm( ~ Class*Survived + Survived*Sex
                    + Survived*Age + Class*Sex
                    + Sex*Age + Class*Age, data = Titanic ) )
```

```
## Call:
## loglm(formula = ~Class * Survived + Survived * Sex + Survived *
##     Age + Class * Sex + Sex * Age + Class * Age, data = Titanic)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 116.5881 13        0
## Pearson                NaN 13      NaN
```

```
( Titanic.DG2 <- Two.fit$deviance - Three.fit$deviance )
```

```
## [1] 116.5879
```

```
# The p-value for testing H_0: M_0 against alternative H_1: M_1 is then given by
( p.value <- 1 - pchisq(Titanic.DG2, 10) )
```

```
## [1] 0
```

```
# Evidence of three-way interactions between the variables.
```

c) Use backwards selection to fit a log-linear model using AIC. Which model is chosen?

What does this tell you?

```
step( Titanic.sat.fit, direction="backward", test = "Chisq" )
```

```
## Start:  AIC=64
## ~Class * Survived * Sex * Age
##
##                           Df AIC        LRT Pr(>Chi)
## - Class:Survived:Sex:Age  3  58 0.00019044        1
## <none>                       64
##
## Step:  AIC=58
## ~Class + Survived + Sex + Age + Class:Survived + Class:Sex +
##     Survived:Sex + Class:Age + Survived:Age + Sex:Age + Class:Survived:Sex +
##     Class:Survived:Age + Class:Sex:Age + Survived:Sex:Age
##
##                       Df     AIC    LRT  Pr(>Chi)
## - Survived:Sex:Age     1  57.685  1.685    0.1942
## <none>                    58.000
## - Class:Sex:Age        3  61.783  9.783    0.0205 *
## - Class:Survived:Age   3  89.263 37.262 4.049e-08 ***
## - Class:Survived:Sex   3 117.013 65.013 4.984e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=57.69
## ~Class + Survived + Sex + Age + Class:Survived + Class:Sex +
##     Survived:Sex + Class:Age + Survived:Age + Sex:Age + Class:Survived:Sex +
##     Class:Survived:Age + Class:Sex:Age
##
##                       Df     AIC    LRT  Pr(>Chi)
## <none>                    57.685
## - Class:Sex:Age        3  71.953 20.268 0.0001494 ***
## - Class:Survived:Age   3  95.899 44.214 1.359e-09 ***
## - Class:Survived:Sex   3 126.904 75.219 3.253e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Call:
## loglm(formula = ~Class + Survived + Sex + Age + Class:Survived +
##     Class:Sex + Survived:Sex + Class:Age + Survived:Age + Sex:Age +
##     Class:Survived:Sex + Class:Survived:Age + Class:Sex:Age,
##     data = Titanic, evaluate = FALSE)
##
## Statistics:
##                       X^2 df  P(> X^2)
## Likelihood Ratio 1.685426  4 0.7933632
## Pearson               NaN  4       NaN
```

We can see the backwards stepwise process of sequentially removing terms from the saturated model one-by-one until the model listed under `Call:` is finally chosen. In this case, we can see that we have all two-way interaction terms, and all three-way interaction terms except that corresponding to `Survived`, `Age` and `Sex`, thus suggesting that these three variable exhibit homogeneous associations amongst each other when conditioning on `Class`.

d) Use Forwards selection to fit a log-linear model using BIC (look in the help file for `step` if you are unsure how to do this). Do you get a different model to that found in part (c)? If so, is this due to the change in model selection criterion or stepwise selection direction?

```
n <- sum( Titanic )
step( Titanic.I.fit, scope = ~ Class*Survived*Sex*Age, direction="forward",
      test = "Chisq", k = log(n) )
```

```
## Start:  AIC=1297.54
## ~Class + Survived + Sex + Age
##
##                   Df     AIC    LRT  Pr(>Chi)
## + Survived:Sex     1  870.77 434.47 < 2.2e-16 ***
## + Class:Sex        3  908.03 412.60 < 2.2e-16 ***
## + Class:Survived   3 1139.73 180.90 < 2.2e-16 ***
## + Class:Age        3 1172.30 148.33 < 2.2e-16 ***
## + Sex:Age          1 1281.95  23.28 1.398e-06 ***
## + Survived:Age     1 1285.68  19.56 9.746e-06 ***
## <none>               1297.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=870.77
## ~Class + Survived + Sex + Age + Survived:Sex
##
##                   Df    AIC    LRT  Pr(>Chi)
## + Class:Sex        3 481.26 412.60 < 2.2e-16 ***
## + Class:Survived   3 712.96 180.90 < 2.2e-16 ***
## + Class:Age        3 745.53 148.33 < 2.2e-16 ***
## + Sex:Age          1 855.18  23.28 1.398e-06 ***
## + Survived:Age     1 858.90  19.56 9.746e-06 ***
## <none>               870.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=481.26
## ~Class + Survived + Sex + Age + Survived:Sex + Class:Sex
##
##                   Df    AIC     LRT  Pr(>Chi)
## + Class:Age        3 356.02 148.327 < 2.2e-16 ***
## + Class:Survived   3 398.27 106.075 < 2.2e-16 ***
```

```
## + Sex:Age         1 465.67  23.284 1.398e-06 ***
## + Survived:Age    1 469.39  19.561 9.746e-06 ***
## <none>              481.26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=356.02
## ~Class + Survived + Sex + Age + Survived:Sex + Class:Sex + Class:Age
##
##                   Df    AIC     LRT  Pr(>Chi)
## + Class:Survived  3 273.03 106.075 < 2.2e-16 ***
## + Survived:Age    1 352.71  11.008 0.0009071 ***
## <none>              356.02
## + Sex:Age         1 357.63   6.086 0.0136259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=273.03
## ~Class + Survived + Sex + Age + Survived:Sex + Class:Sex + Class:Age +
##     Class:Survived
##
##                       Df    AIC    LRT  Pr(>Chi)
## + Class:Survived:Sex  3 230.94 65.180 4.591e-14 ***
## + Survived:Age        1 255.15 25.583 4.237e-07 ***
## <none>                  273.03
## + Sex:Age             1 274.64  6.086   0.01363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=230.94
## ~Class + Survived + Sex + Age + Survived:Sex + Class:Sex + Class:Age +
##     Class:Survived + Class:Survived:Sex
##
##                 Df    AIC    LRT  Pr(>Chi)
## + Survived:Age  1 213.06 25.583 4.237e-07 ***
## <none>            230.94
## + Sex:Age       1 232.56  6.086   0.01363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=213.06
## ~Class + Survived + Sex + Age + Survived:Sex + Class:Sex + Class:Age +
##     Class:Survived + Survived:Age + Class:Survived:Sex
##
##                       Df    AIC     LRT  Pr(>Chi)
## + Class:Survived:Age  3 206.94 29.2063 2.027e-06 ***
## <none>                  213.06
```

```
## + Sex:Age             1 220.62  0.1356    0.7127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=206.94
## ~Class + Survived + Sex + Age + Survived:Sex + Class:Sex + Class:Age +
##     Class:Survived + Survived:Age + Class:Survived:Sex + Class:Survived:Age
##
##           Df    AIC     LRT Pr(>Chi)
## <none>       206.94
## + Sex:Age  1 214.37 0.26868   0.6042
```

```
## Call:
## loglm(formula = ~Class + Survived + Sex + Age + Survived:Sex +
##     Class:Sex + Class:Age + Class:Survived + Survived:Age + Class:Survived:Sex +
##     Class:Survived:Age, data = Titanic, evaluate = FALSE)
##
## Statistics:
##                       X^2 df    P(> X^2)
## Likelihood Ratio 22.22167  8 0.004521358
## Pearson               NaN  8         NaN
```

We can see the forwards stepwise process of sequentially adding terms from the first-order linear model one-by-one until the model listed under `Call:` is finally chosen. In this case, we can see that we have all two-way interaction terms except that between `Sex` and `Age`, and all possible three-way interactions based on this given that the model must be hierarchical. This model suggests that `Sex` and `Age` are conditionally independent given `Class` and `Survived`. It is likely that this change in model selection choice is due to the change of criteria (i.e. BIC rather than AIC), although the selection method can have an effect.

9.4 Practical 4 - Binary Regression

9.4.1 Space Shuttle Flights

The dataset `shuttle` describes, for $n = 23$ space shuttle flights before the Challenger disaster in 1986, the temperature in $^\circ F$ at the time of take-off, as well as the occurrence or non-occurrence of a thermal overstrain of a certain component (the famous *O-ring*). The data set contains the variables:

- `flight` - flight number
- `temp` - temperature in $^\circ F$
- `td` - thermal overstrain (1 =yes, 0 =no)

a) Fit a linear model (with the temperature as covariate) using the R function `lm`. Produce a plot which displays the data, as well as the fitted probabilities of thermal overstrain. Why does this model not provide a satisfactory description of the data?

We fit the linear model, and produce the required plot (shown in Figure 9.14), as follows:

```
shuttle.lm <- lm(td~temp, data=shuttle)
plot(td~temp, data=shuttle, ylab="overstrain")
abline(shuttle.lm)
```

The model does not provide a satisfactory desciption of the data because the fitted function could produce probabilities for overstrain that do not lie in the range $[0, 1]$ - these being meaningless.

b) Fit a logistic regression model to these data, with logit link function, using function `glm`.

We fit a logistic regression model as follows:

```
( shuttle.glm <- glm(td~temp, data=shuttle, family=binomial(link="logit")) )
```

```
##
## Call:  glm(formula = td ~ temp, family = binomial(link = "logit"), data = shuttle)
##
## Coefficients:
## (Intercept)         temp
##     15.0429      -0.2322
##
## Degrees of Freedom: 22 Total (i.e. Null);  21 Residual
## Null Deviance:       28.27
## Residual Deviance: 20.32     AIC: 24.32
```

c) Extract the model parameters from the logistic regression model. Write a function for the expected probability of thermal overstrain `td` in terms of `temp` explicitly (that is, $\hat{P}(\text{td}|\text{temp})$) using your knowledge of logistic regression models, and add its graph to the plot created in part (a).

We extract the model parameters as follows:

```
( coefficients <- as.numeric(shuttle.glm$coef) )
```

```
## [1] 15.0429016 -0.2321627
```

where the `as.numeric` part just allows us to use the coefficients exactly and easily in the function below.

We hence have that

$$\text{logit}(\text{td}|\text{temp}) = \log(\hat{P}(\text{td}|\text{temp})/(1 - \hat{P}(\text{td}|\text{temp}))) = \beta_1 + \beta_2 \text{temp} \tag{9.1}$$

with $\beta_1 = 15.0429$ and $\beta_2 = -0.2322$.

Each additional degree farenheit will reduce the odds of thermal overstrain by the factor $e^{-0.2322} = 0.7928$.
The intercept parameter would be interpreted as the estimated odds of thermal overstrain for a temperature of 0 degrees farenheit.

To write a function that returns $\hat{P}(\text{td}|\text{temp})$, we first rearrange Equation (9.1) as follows:

$$\hat{P}(td) = \frac{\exp(\beta_1 + \beta_2 \text{temp})}{1 + \exp(\beta_1 + \beta_2 \text{temp})} \tag{9.2}$$

Now we need to write a function that returns this probability for any value of the covariate temperature.

```
shuttle.fit <- function(temp){
  exp(coefficients[1] + coefficients[2] * temp) /
             ( 1 + exp(coefficients[1] + coefficients[2] * temp) )
}
```

where `coefficients` is here assumed to be taken from the global environment. To avoid needing to make this assumption (which has risks), we could require the model to be explicitly specified as a function argument as well, that is

```
shuttle.fit2 <- function(temp, model){
  coefficients <- as.numeric(model$coef)
  exp(coefficients[1] + coefficients[2] * temp) /
           ( 1 + exp(coefficients[1] + coefficients[2] * temp) )
}
```

or as a third option we could write a function explicitly using the specific values of beta found above.

Using the second option, we use the following lines of code to add the model fit to the plot:

```
lines(30:90, shuttle.fit2(30:90, model = shuttle.glm), lty = 2)
legend(70, 0.8, c("Linear model", "Logistic model"), lty=c(1,2))
```

d) According to this model, at what temperature is the probability of thermal overstrain exactly equal to 0.5?

Take the inverse of the expression for $\hat{P}(\text{td}|\text{temp})$:

```
shuttle.fit.inverse <- function(prob, model){
  coefficients <- as.numeric(model$coef)
  (log( prob / ( 1- prob ) ) - coefficients[1]) / coefficients[2]
}
shuttle.fit.inverse(0.5, model = shuttle.glm )
```

```
## [1] 64.79464
```

For this model, a temperature of 64.795 has a predicted probability of thermal overstrain of 0.5.

e) On the day of the Challenger disaster, the temperature was $31°F$. Use your model to predict the probability of thermal overstrain on that day.

```
shuttle.fit2(31, model = shuttle.glm)
```

```
## [1] 0.9996088
```

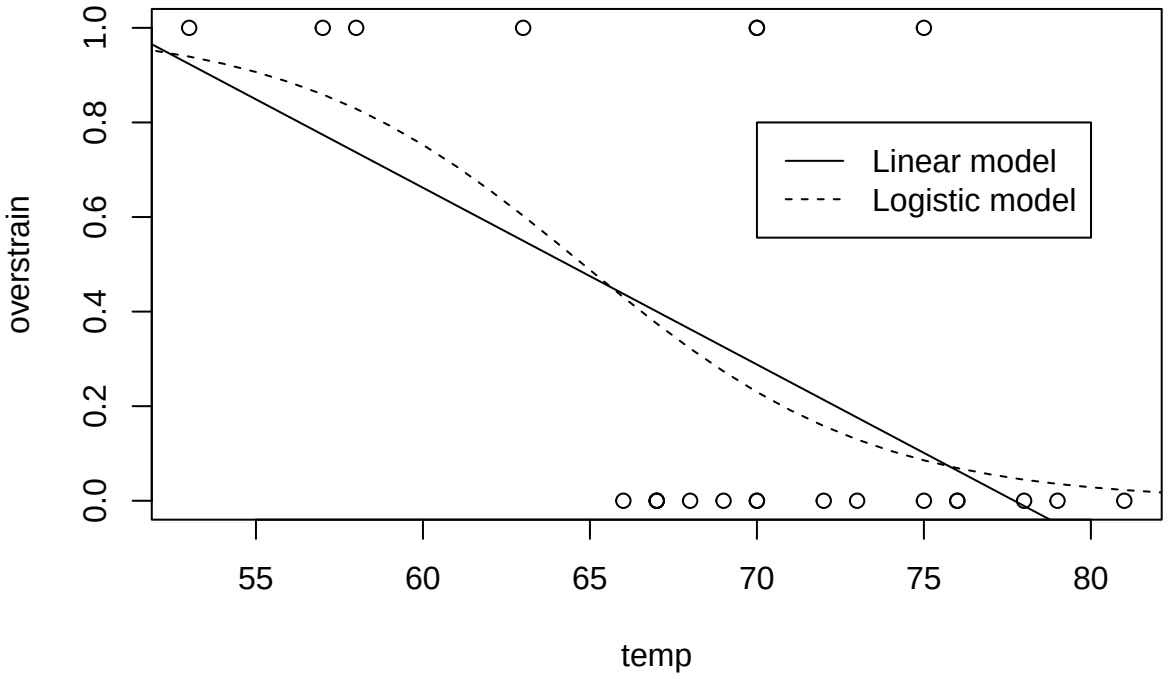


Figure 9.14: Shuttle data and required model fits for Practical 4.

The probability of thermal overstrain is predicted to be very high. As an aside, the commission that investigated the cause of the accident indeed found that a thermal overstrain in the so-called *O-Rings* was responsible for the disaster.

f) Replace the `logit` link by a `probit` link, and repeat the above analysis for this choice of link function.

Fit a probit model and find the coefficient values as follows:

```
( shuttle.glm2 <- glm(td~temp, data=shuttle, family=binomial(link="probit")) )
```

```
##
## Call:  glm(formula = td ~ temp, family = binomial(link = "probit"),
##     data = shuttle)
##
## Coefficients:
## (Intercept)         temp
##      8.7749      -0.1351
##
## Degrees of Freedom: 22 Total (i.e. Null);  21 Residual
## Null Deviance:       28.27
## Residual Deviance: 20.38     AIC: 24.38
```

```
shuttle.glm2$coef
```

```
## (Intercept)        temp
##   8.7749032  -0.1350958
```

In this case we have that

$$\hat{P}(\text{td}|\text{temp}) = \Phi(\hat{\beta}_1 + \hat{\beta}_2\text{temp}) \tag{9.3}$$
$$\hat{\beta}_1 + \hat{\beta}_2\text{temp} = \Phi^{-1}(\hat{P}(\text{td}|\text{temp})) \tag{9.4}$$
$$\text{temp} = \frac{\Phi^{-1}(\hat{P}(\text{td}|\text{temp}) - \hat{\beta}_1}{\hat{\beta}_2} \tag{9.5}$$

We can code these expressions into R functions as follows.

```
shuttle.fit.probit <- function( temp, model ){
  coefficients <- as.numeric(model$coef)
  pnorm(coefficients[1] + coefficients[2] * temp)
}
shuttle.fit.probit.inverse <- function( prob, model ){
  coefficients <- as.numeric(model$coef)
  ( qnorm( prob ) - coefficients[1] ) / coefficients[2]
}
```

We then use these new functions to predict the relevant quantities as follows:

```
shuttle.fit.probit.inverse( 0.5, model = shuttle.glm2 )
```

```
## [1] 64.95321
```

```
shuttle.fit.probit( 31, model = shuttle.glm2 )
```

```
## [1] 0.9999978
```

9.4.2 US Graduate School Admissions

Here we present the code, along with the output, related to Section 8.4.2.

Having read the data in, we get the following...

```
head(graduate)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

```
summary(graduate)
```

```
##      admit             gre             gpa             rank
##  Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu.:2.000
##  Median :0.0000   Median :580.0   Median :3.395   Median :2.000
##  Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean   :2.485
```

```
##  3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.   :4.000
```

We fit and summarise the model as follows.

```
graduate$rank <- factor(graduate$rank)
( graduate.glm <- glm(admit ~ gre + gpa + rank, data = graduate,
                                          family = "binomial") )
```

```
##
## Call:  glm(formula = admit ~ gre + gpa + rank, family = "binomial",
##     data = graduate)
##
## Coefficients:
## (Intercept)          gre          gpa        rank2        rank3        rank4
##   -3.989979     0.002264     0.804038    -0.675443    -1.340204    -1.551464
##
## Degrees of Freedom: 399 Total (i.e. Null);  394 Residual
## Null Deviance:       500
## Residual Deviance: 458.5     AIC: 470.5
```

```
summary(graduate.glm)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = "binomial",
##     data = graduate)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## gre          0.002264   0.001094   2.070 0.038465 *
## gpa          0.804038   0.331819   2.423 0.015388 *
## rank2       -0.675443   0.316490  -2.134 0.032829 *
## rank3       -1.340204   0.345306  -3.881 0.000104 ***
## rank4       -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
```

```
## Number of Fisher Scoring iterations: 4
```

Then, to get the odds ratios, we run

```
exp(coef(graduate.glm))
```

```
## (Intercept)         gre         gpa       rank2       rank3       rank4
##   0.0185001   1.0022670   2.2345448   0.5089310   0.2617923   0.2119375
```

To predict admission for a student with GRE 700, GPA 3.7, to a rank 2 institution, we run

```
( predData <- with(graduate, data.frame(gre = 700, gpa = 3.7, rank = factor(2))) )
```

```
##   gre gpa rank
## 1 700 3.7    2
```

```
( predData$prediction <- predict(graduate.glm, newdata = predData,
                                 type = "response") )
```

```
## [1] 0.4736781
```

and find that the predicted probability of admission for such a student is 0.4737.

By looking at the summary, or extracting the AIC value from the summary, for the original model, we have

```
summary(graduate.glm)$aic
```

```
## [1] 470.5175
```

If we remove one term at a time from the model, we get the following

```
graduate.glm2 <- glm(admit ~ gre + gpa, data = graduate,
                     family = "binomial")
summary(graduate.glm2)$aic
```

```
## [1] 486.344
```

```
graduate.glm3 <- glm(admit ~ gre + rank, data = graduate,
                     family = "binomial")
summary(graduate.glm3)$aic
```

```
## [1] 474.5318
```

```
graduate.glm4 <- glm(admit ~ gpa + rank, data = graduate,
                     family = "binomial")
summary(graduate.glm4)$aic
```

```
## [1] 472.8753
```

Removing `rank` seems to make the AIC quite a bit worse, whereas removal of either terms `gpa` or `gre` results in a much smaller difference. This is consistent with the levels of significance for each term shown in the summary of the original model.

We now repeat the analysis fitting this time with a `probit` link function:

```
graduate.glm.probit <- glm(admit ~ gre + gpa + rank, data = graduate,
                    family = "binomial"(link="probit"))
summary(graduate.glm.probit)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = binomial(link = "probit"),
##     data = graduate)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6163  -0.8710  -0.6389   1.1560   2.1035
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.386836   0.673946  -3.542 0.000398 ***
## gre          0.001376   0.000650   2.116 0.034329 *
## gpa          0.477730   0.197197   2.423 0.015410 *
## rank2       -0.415399   0.194977  -2.131 0.033130 *
## rank3       -0.812138   0.208358  -3.898 9.71e-05 ***
## rank4       -0.935899   0.245272  -3.816 0.000136 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.41  on 394  degrees of freedom
## AIC: 470.41
##
## Number of Fisher Scoring iterations: 4
```

The model has a slightly lower AIC value.

The prediction for a student with GRE 700, GPA 3.7, to a rank 2 institution, is given by

```
predData$prediction2 <- predict(graduate.glm.probit, newdata = predData,
                                type = "response")
```

The prediction for the new data point is similar for both models, at approximately 0.47.

Part IV

Problems

Chapter 10

Problems

Chapter 1

Q1-1: Introductory Questions

a) What is a model?
b) What is a statistical model?
c) What is an advanced statistical model?
d) What is a variable?
e) What is a random variable?

Q1-2: Types of Variable

Decide whether the following variables are categorical or numerical, and classify further if possible:

a) gender of the next lamb born at the local farm.

b) number of times a dice needs to be thrown until the first 6 is observed.

c) amount of fluid (in ounces) dispensed by a machine used to fill bottles with lemonade.

d) thickness of penny coins in millimetres.

e) assignment grades of a 3H Mathematics course (from A to D).

f) marital status of some random sample of citizens.

Q1-3: Properties of Probability Distributions

a) Prove the results for the expectation and covariance structure of the multinomial distribution stated in Section 1.4.3.1.

b) Suppose $X_1^2, ..., X_m^2$ are m independent variables with chi-squared distributions $X_i^2 \sim \chi^2(n_i)$. Show that

$$\sum_{i=1}^{m} X_i^2 = \chi^2(\sum_{i=1}^{m} n_i) \tag{10.1}$$

Table 10.1: Data on restraint use and fatality of road traffic accidents.

	Injury: Fatal	**Non-Fatal**	**Sum**
Restraint Use: No	433	8049	**8482**
Restraint Use: Yes	570	554883	**555453**
Sum	**1003**	**562932**	**563935**

c) Suppose X^2 has the distribution $\chi^2(n)$. Prove that

$$\mathrm{E}[X^2] = n \tag{10.2}$$
$$\mathrm{Var}[X^2] = 2n \tag{10.3}$$

Hint: Your solution may require you to assume or show that $\mathrm{E}[Z^4] = 3$, where $Z \sim \mathcal{N}(0, 1)$.

Chapter 2

Q2-1: Quickfire Questions

a) What is the difference between a Poisson, Multinomial and product Multinomial sampling scheme?
b) What does an odds of 1.8 mean relative to success probability π?

Q2-2: Sampling Schemes

Write down statements for the expectation and variance of a variable following a product Multinomial sampling scheme as described in Section 2.3.3.

Q2-3: Fatality of Road Traffic Accidents

Table 10.1 shows fatality results for drivers and passengers in road traffic accidents in Florida in 2015, according to whether the person was wearing a shoulder and lap belt restraint versus not using one. Find and interpret the odds ratio.

Q2-4: Difference of Proportions or Odds Ratios?

A 20-year study of British male physicians (Doll and Peto [1976]) noted that the proportion who died from lung cancer was 0.00140 per year for cigarette smokers and 0.00010 per year for non-smokers. The proportion who died from heart disease was 0.00669 for smokers and 0.00413 for non-smokers.

a) Describe and compare the association of smoking with lung cancer and with heart disease using the difference of proportions.

b) Describe and compare the association of smoking with lung cancer and heart disease using the odds ratio.

c) Which response (lung cancer or heart disease) is more strongly related to cigarette smoking, in terms of increased proportional risk to the individual?

d) Which response (lung cancer or heart disease) is more strongly related to cigarette smoking, in terms of the reduction in deaths that could occur with an absence of smoking?

Q2-5: Asymptotic Distribution of X^2

a) Show that Equation (2.50) in Section 2.4.2 holds.

b) Show that Equation (2.51) in Section 2.4.2 holds.

Q2-6: Maximum Likelihood by Lagrange Multipliers

a) Consider a Multinomial sampling scheme, and that X and Y are independent. We need to find the MLE of $\boldsymbol{\pi}$, but now we have that

$$\pi_{ij} = \pi_{i+}\pi_{+j} \tag{10.4}$$

The log likelihood is

$$\begin{aligned} l(\boldsymbol{\pi}) &\propto \sum_{i,j} n_{ij} \log(\pi_{ij}) && (10.5) \\ &= \sum_{i} n_{i+} \log(\pi_{i+}) + \sum_{j} n_{+j} \log(\pi_{+j}) && (10.6) \end{aligned}$$

Use the method of Lagrange multipliers to show that

$$\begin{aligned} \hat{\pi}_{i+} &= \frac{n_{i+}}{n_{++}} && (10.7) \\ \hat{\pi}_{+j} &= \frac{n_{+j}}{n_{++}} && (10.8) \end{aligned}$$

Hint: You may wish to go through the following steps to achieve this:

- *Write down the constraints.*
- *Write down the Lagrange function.*
- *State the equations satisfied at the local optima.*
- *Rearrange for $\hat{\lambda}_1, \hat{\lambda}_2$ and thus $\hat{\pi}_{i+}$ and $\hat{\pi}_{+j}$.*

b) Using the method of Lagrange multipliers, show that Equation (2.100) of Section 2.4.4.1.1 holds.

Q2-7: Second-Order Taylor Expansion

Show that Approximation (2.116) of Section 2.4.5.1 holds.

Q2-8: Relative Risk

a) In 1998, A British study reported that "*Female smokers were 1.7 times more vulnerable than Male smokers to get lung cancer.*" We don't investigate whether this is true or not, but is 1.7 the odds ratio or the relative risk? Briefly (one sentence maximum) explain your answer.

b) A National Cancer institute study about tamoxifen and breast cancer reported that the women taking the drug were 45% less likely to experience invasive breast cancer than were women taking a placebo. Find the relative risk for (i) those taking the drug compared with those taking the placebo, and (ii) those taking the placebo compared with those taking the drug.

Q2-9: The Titanic

For adults who sailed on the *Titanic* on its fateful voyage, the odds ratio between gender (categorised as *Female* or *Male*), and survival (categorised as *yes* or *no*) was 11.4 (Dawson [1995]).

a) It is claimed that "*The Probability of survival for women was 11.4 times that for men*". i) What is wrong with this interpretation? ii) What should the correct interpretation be? iii) When would the quoted interpretation be approximately correct?

b) The odds of survival for women was 2.9. Find the proportion of each gender who survived.

Q2-10: Test and Reality

For a diagnostic test of a certain disease, let π_1 denote the probability that the diagnosis is positive given that a subject has the disease, and let π_2 denote the probability that the diagnosis is positive given that a subject does not have the disease. Let τ denote the probability that a subject has the disease.

a) More relevant to a patient who has received a positive diagnosis is the probability that they truly have the disease. Given that a diagnosis is positive, show that the probability that a subject has the disease (called the *positive predictive value*) is

$$\frac{\pi_1\tau}{\pi_1\tau + \pi_2(1-\tau)} \tag{10.9}$$

b) Suppose that a diagnostic test for a disease has both sensitivity and specificity equal to 0.95, and that $\tau = 0.005$. Find the probability that a subject truly has the disease given a positive diagnostic test result.

c) Create a 2×2 contingency table of cross-classified probabilities for presence or absence of the disease and positive or negative diagnostic test result.

d) Calculate the odds ratio and interpret.

Table 10.2: Data for Q2-11

	Happiness: Not too Happy	Pretty Happy	Very Happy
Income: Above Average	21	159	110
Income: Average	53	372	221
Income: Below Average	94	249	83

Q2-11: Happiness and Income

Table 10.2 shows data from a General Social Survey cross-classifying a person's perceived happiness with their family income.

a) Perform a χ^2 test of independence between the two variables.

b) Calculate and interpret the adjusted residuals for the four corner cells of the table.

Q2-12: Tea! (Fisher's Exact Test of Independence)

This is a quote from Fisher [1937][1]:

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. For this purpose let us first lay down a simple form of experiment with a view to studying its limitations and its characteristics, both those which appear to be essential to the experimental method, when well developed, and those which are not essential but auxiliary.

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

a) From the text, we know that there are 4 cups with milk added first and 4 with tea infusion added first. How many distinct orderings can these 8 cups to be tasted take, in terms of type.

b) Note that the lady also knows that there are four cups of each type, and must group them into two sets of four (those she thinks had milk added first, and those she thinks had tea infusion added first). Given that the lady guesses milk first three times when indeed the milk was added first, cross-classify the lady's guesses against the truth in a 2×2 contingency table.

[1] It is possible we do not need the entire quotation presented here for our purposes, but I think there is something to be gained from seeing Statistical writing of many years ago...

Table 10.3: Data on presidential votes for Q2-13.

	Vote 2012: Obama	**Romney**
Vote 2008: Obama	802	53
Vote 2008: McCain	34	494

c) Fisher presented an *exact*[2] solution for testing the null hypothesis

$$\mathcal{H}_0 : r_{12} = 1 \tag{10.10}$$

against the one-sided alternative

$$\mathcal{H}_1 : r_{12} > 1 \tag{10.11}$$

for contingency tables with fixed row and column sums.
What hypothesis does $\mathcal{H}_0$ correspond to in the context of the tea tasting test described above? Write down an expression for $P(N_{11} = t)$ under $\mathcal{H}_0$. Thus, perform a (Fisher's exact[3]) hypothesis test to test the lady's claim that she can indeed discriminate whether the milk or tea infusion was first added to the cup.

d) Suppose the lady had correctly classified all eight cups as having either milk or tea infusion added first. Would Fisher's exact hypothesis test provide evidence of her ability now?

Q2-13: US Presidential Elections

Table 10.3 cross-classifies a sample of votes in the 2008 and 2012 US Presidential elections. Test the null hypothesis that vote in 2008 was independent from vote in 2012 by estimating, and finding a 95% confidence interval for, the population odds ratio.

Q2-14: Job Security and Happiness

Consider the table presented in Figure 10.1 summarising the responses for extent of agreement to the statement "job security is good" (JobSecOK, or Job Security) and general happiness (Happiness) from the 2018 US General Social Surveys. Additional possible responses of *don't know* and *no answer* are omitted here.

a) Calculate a minimal set of local odds ratios. Interpret and discuss.

b) Calculate a minimal set of odds ratios, treating job security locally and happiness cumulatively. Interpret and discuss.

c) Calculate a minimal set of odds ratios, treating job security as a nominal variable (taking *very true* as the reference category) and happiness as a cumulative variable. Interpret and discuss.

[2] *Exact* in the sense that the probabilities of any possible outcome can be calculated exactly.

[3] You don't need to specifically worry about it being Fisher's exact test; just perform a logical exact hypothesis test and it will most likely correspond with Fisher's!

	Not too happy	Pretty happy	Very happy
Not at all true	15	25	5
Not too true	21	47	21
Somewhat true	64	248	100
Very true	73	474	311

Figure 10.1: Contingency Table for Q2-14

d) Calculate a minimal set of global odds ratios, that is, treating both job security and happiness as cumulative variables.

e) Calculate a 95% confidence interval for the global odds ratio, with the first two categories of each variable being grouped together into one category in each case.

f) Perform a linear trend test to assess whether there is any evidence of association between Happiness and Job Security.

Chapter 3

Q3-1: A Question of Notation

a) Based on the notation of Section 3.1, state what table (in terms of dimension, as well as conditioning and marginalising of variables) is being described by:

i. $(n_{(i)jk})$

ii. $(n_{(i)+k(l)})$

iii. $(n_{i_1 i_2 + (i_4) i_5 (i_6) +})$

b) Use the notation of Section 3.1, in the context of Section 3.1.1.1, to notate and construct the conditional (on drug treatment $X = B$) table, having marginalised over clinic Z.

c) Assuming a total of 8 variables, use the notation of Section 3.1 to notate the partial marginal table obtained by summing over all levels of variables i_3, i_5 and i_7, for a fixed level of variables i_1 and i_4.

Q3-2: Conditional and Marginal Local Odds Ratios

Show that Equation (3.17) of Section 3.3.4.2 holds for local odds ratios under the conditional independence of X and Y given Z. Recall that such conditional independence means that Equation (3.16) holds.

Q3-3: Types of Independence

In Section 3.3, we discussed different types of independence, such as mutual, joint, marginal and conditional, as well as homogeneous associations. Throughout, I illustrated what the different independence assumptions might mean in the context of the three variables Drug Treatment (X), Response (Y) and Clinic (Z) of the example introduced in Section 3.1.1.1.

Think of another example with three (or more...) categorical variables (with at least two levels each), and consider what the different types of independence would imply about the associations (or lack thereof) between the three variables. Note that I am not interested in any numbers here - it is about the implications of any hypothesised independence scenario - so all we need are three variable names with possible categories. You may wish to use a hypothetical example of your own construction, or one that you have found from other sources.

You may wish to hypothesise possible reasoning for each situation, such as I did in Section 3.3.5 with the *Example Potential Explanations*. Note that the purpose of this is to get you thinking (like an expert might) of possible explanations in real-life scenarios that would lead to the various hypotheses about independence. It does not mean that strong evidence of any particular scenario of association necessarily implies that the explanation or reasoning is correct[4].

You should discuss your answers to this question with your fellow students, and you are also welcome to discuss your answer with me during the weekly office hour.

I am well aware that many will be tempted to skip this question, but being able to convey the ideas reflected in our models or tests into practical examples is a very important part of being a Statistician (are we really doing Statistics at all if we don't have this skill?). Even if this holistic reasoning does not motivate you, this skill will be tested in the end-of-year exam, and is often the source of lost marks among our students.

Q3-4: Conditional and Joint Independence

Consider an $I \times J \times K$ contingency table, with classification variables X, Y, Z. By writing down the relevant probability forms involved, or otherwise...

a) Prove that if
   - X and Y are conditionally independent given Z; and
   - X and Z are conditionally independent given Y,

then Y and Z are jointly independent from X.

b) Prove that if Y and Z are jointly independent from X, then
   - X and Y are conditionally independent given Z; and
   - X and Z are conditionally independent given Y.

Q3-5: 1988 General Social Survey

The 1988 General Social Survey compiled by the National Opinion Research Center asked:

[4]We all know (I hope...) that *correlation* (or in this case, *association*) *does not imply causation*.

- "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms."

The table in Figure 10.2 summarizes opinions about health care costs (H) and the information program (I), classified also by the respondent's gender (G).

Gender (G)	Information Opinion (I)	Health Opinion (H)	
		Support	Oppose
Male	Support	76	160
	Oppose	6	25
Female	Support	114	181
	Oppose	11	48

Figure 10.2: Source: 1988 General Social Survey, National Opinion Research Center.

a) Compute the marginal GH-table.

b) For the marginal GH-table, compute the MLE of the marginal odds ratio, along with a 95% confidence interval. Interpret the result.

c) Perform a Mantel Haenszel Chi-Square test, at the 5% level of significance, in order to test if the Information Opinion and the Health Opinion are independent at each level of Gender.

d) Compute the partial (conditional) IH odds ratio at each level of Gender. Interpret the result.

Q3-6: Alien Spacejets

The aliens are designing spacejets to enable them to travel between two planets. In particular, they need spacejets that can withstand the atmospheric pressure encountered whilst travelling through space between the two planets. They therefore cross-classify the survival of different spacejets on test runs in the atmosphere (Survives, Y: yes or no) against Type, Z: A or B and Colour, X: red or green. The dataset is presented in the table shown in Figure 10.3.

a) Calculate the marginal XY contingency table of the observed counts.

b) Calculate an estimate of, along with a 90% confidence interval for, the marginal odds ratio of Colour and Survives.

c) What does the estimated odds ratio tell us about the relation between Colour and Survives? Infer whether there is evidence that Colour and Survives are dependent or not.

d) Based on the results of Parts (a)-(c), the aliens are convinced that the dataset provides at least some evidence that, regardless of Type, green spacejets have a greater chance of surviving the intense atmospheric pressure than red spacejets. Perform reasonable analyses to:

Table 1: Dataset.

Colour (X)	Survives (Y)	Type (Z)	counts
red	yes	A	20
red	no	A	3
green	yes	A	50
green	no	A	14
red	yes	B	32
red	no	B	47
green	yes	B	5
green	no	B	15

Figure 10.3: Table of Data for Q3-6.

i. illustrate to the aliens why they have jumped to the wrong conclusion, and
ii. discuss alternative inferences.

Supposing this was an excerpt from an exam question (as it may have been in the past), credit would be awarded for the clarity of your answer.

Q3-7 The Aliens in Durham

Suppose a horde of 216 aliens (beings from a far-off planet) arrive in Durham and are surveyed about the suitability of building structures were they to make Durham their home (they need to report back to the rest of their species). Suppose the rest of their kind is interested in knowing whether the opinion about suitability is associated with alien height (short, average-height, tall) and alien type (A,B). They therefore cross-classify these variables in the contingency table presented in Figure 10.4.

		Suitability (Y)	
Type (Z)	Height (X)	Suitable	Unsuitable
A	Short	12	5
	Average	34	13
	Tall	25	22
B	Short	12	12
	Average	60	3
	Tall	14	4

Figure 10.4: Contingency Table of Data for Q3-7.

a) Calculate an estimate of, along with a 95% confidence interval for, the marginal odds ratio of Suitability and Type.

b) Calculate a minimal set of marginal (over Type) global odds ratios between Height and Suitability, along with a reasonable 95% confidence interval for each. What assumption about variable Height is being made in order to do this? Interpret and explain the results in terms of associations between Height and Suitability.

Chapter 4

Q4-1: Equations from the Notes

a) By rearranging Equation (4.11), show that Equations (4.16) - (4.19) hold for the 2-way saturated LLM using zero-sum constraints.

b) Assuming Equation (4.26) holds, for the 2-way independence LLM, show that the ML estimates for the λ parameters are as given by Equations (4.27) - (4.29).

Q4-2: Two-way Mutual Independence Model

Assume a LLM with dependency structure $[X, Y]$ under a Poisson sampling scheme with corner point constraints.

a) Write down the appropriate LLM expression corresponding to dependency structure $[X, Y]$.

b) Rearrange the appropriate LLM expression for the corresponding λ parameters.

c) Write down the log-likelihood equations, and hence use the method of Lagrange multipliers to show that

$$\begin{aligned} \hat{E}_{++} = E_{++}(\hat{\boldsymbol{\lambda}}) &= n_{++} \\ \hat{E}_{i+} = E_{i+}(\hat{\boldsymbol{\lambda}}) &= n_{i+} \qquad i = 1, ..., I \\ \hat{E}_{+j} = E_{+j}(\hat{\boldsymbol{\lambda}}) &= n_{+j} \qquad j = 1, ..., J \end{aligned}$$

d) Using the results of parts (b) and (c), or otherwise, find the MLEs of the LLM λ coefficients.

Q4-3: Hierarchical Log Linear Models

a) Identify if the following models are hierarchical, and for the ones that are, use the square bracket [] notation (e.g. $[X, Y]$ for the two-way independence model) presented throughout Sections 3 and 4 to represent them, and discuss what assumptions about the associations between the variables they convey:

$$\begin{aligned} \text{i.} \quad \log E_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} \qquad \forall i, j, k \\ \text{ii.} \quad \log E_{ijkl} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{il}^{XW} + \lambda_{kl}^{ZW} + \lambda_{ikl}^{XZW} \qquad \forall i, j, k, l \\ \text{iii.} \quad \log E_{ijkl} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ik}^{XZ} + \lambda_{il}^{XW} + \lambda_{kl}^{ZW} + \lambda_{ikl}^{XZW} \qquad \forall i, j, k, l \end{aligned}$$

b) Write down the hierarchical log-linear models corresponding to

(i) $[XY, XZ]$

(ii) $[XY, XZ, XW]$

(iii) $[X_1X_3, X_2X_3X_4, X_4X_5, X_5X_6]$

and discuss what assumptions about the associations between the variables they convey.

c) For the model discussed in part b-iii above, are variables X_1 and X_6 necessarily independent? Explain your answer.

Q4-4: 1988 General Social Survey

The 1988 General Social Survey compiled by the National Opinion Research Center asked:

- "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms."

The table in Figure 10.2 summarizes opinions about health care costs (H) and the information program (I), classified also by the respondent's gender (G).

Regarding the questions below, any inference based on hypothesis tests should be performed at the 5% level of significance.

a) By using appropriate statistical tools, compare the following associations of the classification variables.

- mutual independence of X, Y and Z, against
- conditional independence of X and Y on Z.

Justify the way you address the problem.

b) By using appropriate statistical tools, compare the following associations of the classification variables

- joint independence of X and Z on Y, against
- joint independence of X and Y on Z.

Justify the way you address the problem.

Q4-5: $2 \times 2 \times 2$ Contingency Table

Consider a $2 \times 2 \times 2$ Contingency Table with classification variables X, Y and Z.

a) State the equation of the Log-linear model describing the dependency type $[XY, XZ]$.

b) Apply the two types of the non-identifiability constraints (corner points and sum-to-zero).

c) Write down the number of free parameters, and explain how you calculated them.

For parts (d) and (e), consider the corner point constraints only.

d) Express $\log(\pi_{1|jk}/\pi_{2|jk})$ as a function of the linear model λ coefficients.

e) Express $\log(r_{ij(k)}^{XY})$, that is, the log conditional (on Z) odds ratio, as a function of the linear model λ coefficients. Give a short interpretation of λ_{11}^{XY} based on this.

Q4-6: Exam-Style Question

All possible hierarchical log-linear models were fitted to a contingency table, based on a sample of 312 individuals, for three factors: X (2 levels), Y (3 levels) and Z (5 levels). The sampling scheme that was implemented was the Poisson sampling scheme. The results are shown in the table in Figure 10.5.

Model	Log-likelihood
$[X, Y, Z]$	-71.75
$[Z, XY]$	-71.69
$[Y, XZ]$	-70.37
$[X, YZ]$	-48.13
$[XY, XZ]$	-70.30
$[XY, YZ]$	-47.83
$[XZ, YZ]$	-39.83
$[XY, XZ, YZ]$	-39.30
$[XYZ]$	-38.00

Figure 10.5: Table for Q4-5.

a) Calculate the number of free parameters resulting after applying corner point non-identifiability constraints for each model in the table.

b) Define the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC), then compute both AIC and BIC for each model in the table.

c) Identify which model is selected by each criterion and give a short explanation of what each criterion is intended to achieve.

d) Explain precisely what model $[XY, YZ]$ says about the dependence of factor Y on the other two factors.

Chapter 5

Q5-1: Tokyo Rainfall

We consider rainfall data recorded in Tokyo in the first 11 days of the year 1983, as given by the table in Figure 10.6. For each day, we know whether it rained ($z_i = 1$) or did not rain ($z_i = 0$) on that day:

Day i	1	2	3	4	5	6	7	8	9	10	11
z_i	0	1	1	0	1	1	0	0	0	0	0

Figure 10.6: Rainfall data in Tokyo in the first 11 days of the year 1983.

The goal is to find a model which predicts the probability of rain on day $i+1$ based on the rainfall on day i, which is given by z_i. Therefore, we define the response $y_i = z_{i+1}$, $i = 1, \ldots, 10$, so that we can write the data in the form given by the table in Figure 10.7

Day i	1	2	3	4	5	6	7	8	9	10
z_i	0	1	1	0	1	1	0	0	0	0
y_i	1	1	0	1	1	0	0	0	0	0

Figure 10.7: Rainfall data in Tokyo for the first 10 days of the year 1983 along with data about whether it rained the following day.

a) Formulate a logistic regression model for $\pi(z)$ (consider the elements stated in Section 5.3.2), using the linear predictor $\eta = \beta_1 + \beta_2 z$, that is, take the predictors to be $x_1 \equiv 1$ and $x_2 = z$.

b) Formulate the score function $S(\boldsymbol{\beta})$.

c) Solve the estimating equation $S(\hat{\boldsymbol{\beta}}) = 0$.

d) According to BBC Weather, it didn't rain in Tokyo on Thursday November 17, 2022. Use your model to predict the rainfall probability in Tokyo on Friday November 18, 2022.

Chapter 6

Q6-1: Properties of the Exponential Dispersion Family

Let the distribution for Y be an exponential dispersion family:

$$P(y|\theta(\mu), \phi) = \exp\Big[\frac{y\theta(\mu) - b(\theta(\mu))}{\phi} + c(y, \phi)\Big] \tag{10.12}$$

where $\mu(\theta) = \mathrm{E}[Y|\theta, \phi]$.

a) Use the fact that $\mu(\theta)$ and $\theta(\mu)$ are inverses to show that $\theta'(\mu) = 1/\mathcal{V}(\mu)$, where $\mathcal{V}(\mu)$ is the variance function.

b) Show that with one data point $(\boldsymbol{x}, y)$, the maximum likelihood estimator for μ is y.

Chapter 7

Q7-1: The Gamma Distribution

We consider the exponential dispersion family (EDF)

$$P(y|\theta,\phi) = \exp\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\} \qquad y > 0 \tag{10.13}$$

a) Show that the expectation of a distribution belonging to a EDF is given by $b'(\theta)$.

b) Show that the Gamma distribution with density

$$P(y|\nu,\alpha) = \frac{\alpha^\nu}{\Gamma(\nu)} y^{\nu-1} e^{-\alpha y} \tag{10.14}$$

(with shape parameter ν and rate parameter α) is a member of the EDF.

Note: The Gamma-function is defined by $\Gamma(\nu) = \int_0^\infty e^{-t} t^{\nu-1}\, dt \quad (\nu > 0)$.

c) Exploiting properties of the EDF, find the mean μ and the variance of the Gamma distribution.

d) Reparametrize the density function so that it depends on the parameters ν and μ.

e) For gamma-distributed response, identify the natural link for use in a generalized linear model. Why is this link function often unsuitable in practice? Suggest an alternative link function.

Q7-2: Coronary Heart Disease

We are given data collected in the framework of a study of coronary heart disease in the table presented in Figure 10.8. It shows 1329 patients cross-classified by the level of their serum cholesterol (below or above 260) and the presence or absence of heart disease.

Serum	Heart Disease		
Cholesterol	present	absent	
< 260	51	992	$n_1 = 1043$
≥ 260	41	245	$n_2 = 286$
Total	92	1237	$N = 1329$

Figure 10.8: Table of data cross-classifying cholesterol (treated as a binary covariate) and presence or absence of heart disease.

We can consider this as a data set which is grouped with respect to a binary covariate, with values (say) $z_1 = 0$ (if < 260) and $z_2 = 1$ (if ≥ 260), and response defined through group-wise heart disease presence rates, that is $y_1 = 51/1043$ and $y_2 = 41/286$. We model this data through a Binomial logit model, that is

$$\pi(z_i) = \frac{\exp(\beta_1 + \beta_2 z_i)}{1 + \exp(\beta_1 + \beta_2 z_i)} \tag{10.15}$$

where $y_i|z_i \sim B(n_i, \pi(z_i))/n_i$ (*rescaled* Binomial distribution).

a) Verify through differentiation of the log-likelihood that the score-function is given by

$$S(\beta_1, \beta_2) = \sum_{i=1}^{2} n_i \begin{pmatrix} 1 \\ z_i \end{pmatrix} \Big(y_i - \frac{\exp(\beta_1 + \beta_2 z_i)}{1 + \exp(\beta_1 + \beta_2 z_i)} \Big) \tag{10.16}$$

b) Solve the score equation by hand (i.e. without R).

c) Comment on the relationship between this and the solution to Q5-1.

Q7-3: Exam-Style Question

Consider the following one-parameter family of probability densities:

$$P(y|a) = \frac{\cos(a)}{e^{-(a-\frac{\pi}{2})y} + e^{-(a+\frac{\pi}{2})y}} \tag{10.17}$$

where $y \in \mathbb{R}$ and $a \in (-\frac{\pi}{2}, \frac{\pi}{2})$.

a) Show that the above family of distributions forms an exponential dispersion family of distributions. Be careful to define all the elements of an exponential dispersion family.

b) Derive the mean and the variance as a function of the natural parameter, and express the variance in terms of the mean. What happens as a reaches the limits of its range?

c) If you were building a GLM using this distribution, what would be the natural link?

Chapter 11

Solutions

Chapter 1

Q1-1: Introductory Questions

a) What is a model?
b) What is a statistical model?
c) What is an advanced statistical model?
d) What is a variable?
e) What is a random variable?

Answer

a) Quite broad. The Cambridge dictionary lists the following definitions of model (noun)

- something that a copy can be based on because it is an extremely good example of its type.
- a person who wears clothes so that they can be photographed or shown to possible buyers, or a person who is employed to be photographed or painted.
- a physical object, usually smaller than the real object, that is used to represent something.
- **a simple description of a system or process that can be used in calculations or predictions of what might happen.**

b) A statistical model is a set (also known as a family) of (probability) distributions.

c) Advanced is clearly a subjective word here...

- Can a model be advanced? If so, what does that mean?
- Perhaps it is the modelling that is advanced? Requiring more in-depth, expert knowledge, for example.
- Let's look at the **Aim of the Course**: To provide advanced methodological and practical knowledge in the field of statistical modelling, covering a wide range of modelling techniques which are essential for the professional statistician.

d) and e) A *variable* in statistics is the same as a variable in mathematics more generally: an unspecified element of a given set whose elements are the possible values of the variable. The difference is that in statistics we place a probability distribution on this

space of values. A variable with a probability distribution on its space of values is sometimes called a *random variable* (although the usual mathematical definition is not in these terms). However, I believe this terminology can be redundant and confusing, and would exercise caution using this terminology, for the following two reasons:

- Firstly, nothing has happened to the variable. Rather, an extra structure (the probability distribution) has been added to the space of values, and it is this structure that we must focus on for statistics. As a result, just as in mathematics more generally, we can, and will, get away with not distinguishing between a variable and its value, using x, for example, for both. Nevertheless, just as in mathematics more generally, such a distinction is sometimes useful. If it is useful, then a variable will be denoted in upper case: the usual notation x will be replaced by the notation $X = x$, both of them indicating that the variable X has the value x.

- Secondly, the terminology encourages the notion that there are two types of quantity *in the world*: "random" quantities, to be represented by "random" variables; and "non-random" quantities, to be represented by "ordinary" variables. With the possible exception of quantum mechanics, which is irrelevant to our purposes, this is false, or, rather, meaningless: try thinking of an empirical test to decide whether a quantity is "random" or "non-random"; there are only quantities, represented by variables. The purpose of probability distributions is to describe our *knowledge of the values of these variables*.

Q1-2: Types of Variable

Decide whether the following variables are categorical or numerical, and classify further if possible:

a) gender of the next lamb born at the local farm.

- **Answer:** categorical, unranked, binary.

b) number of times a dice needs to be thrown until the first 6 is observed.

- **Answer:** numerical, discrete.

c) amount of fluid (in ounces) dispensed by a machine used to fill bottles with lemonade.

- **Answer:** numerical, continuous.

d) thickness of penny coins in millimetres.

- **Answer:** numerical, continuous.

e) assignment grades of a 3H Mathematics course (from A to D).

- **Answer:** categorical, ranked.

f) marital status of some random sample of citizens.

- **Answer:** categorical, unranked.

Q1-3: Properties of Probability Distributions

a) Prove the results for the expectation and covariance structure of the multinomial distribution stated in Section 1.4.3.1.

Answer: Consider first a Multinoulli random variable $\boldsymbol{W}$, that is, a random variable following a Multinomial distribution with $n = 1$, parameterised by probability vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)$. Note that $\boldsymbol{W}$ is a vector of $k-1$ 0's and a single 1, with the W_j taking value 1 with probability π_j.

Almost by definition, we have that $\mathrm{E}[\boldsymbol{W}] = \boldsymbol{\pi}$.

Now, for $j = 1, ..., k$, we have

$$\begin{aligned} \mathrm{Var}[W_j] &= \mathrm{E}[W_j^2] - (\mathrm{E}[W_j])^2 && (11.1)\\ &= \mathrm{E}[W_j] - (\mathrm{E}[W_j])^2 && (11.2)\\ &= \pi_j(1 - \pi_j) && (11.3) \end{aligned}$$

For $j \neq j'$:

$$\begin{aligned} \mathrm{Cov}[W_j, W_{j'}] &= \mathrm{E}[W_j W_{j'}] - \mathrm{E}[W_j]\mathrm{E}[W_{j'}] && (11.4)\\ &= 0 - \pi_j \pi_{j'} && (11.5)\\ &= -\pi_j \pi_{j'} && (11.6) \end{aligned}$$

where we have used the fact that if $j \neq j'$, at least one of W_j and $W_{j'}$ is equal to 0. It therefore follows that

$$\mathrm{Var}[\boldsymbol{W}] = \Sigma = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T \quad (11.7)$$

Now, $\boldsymbol{Y} = \sum_{l=1}^{n} \boldsymbol{Y}_l$, where $\boldsymbol{Y}_l, l = 1, ..., n$ are i.i.d. multinoulli. Therefore

$$\begin{aligned} \mathrm{E}[\boldsymbol{Y}] &= \mathrm{E}[\sum_{l=1}^{n} \boldsymbol{Y}_l] && (11.8)\\ &= \sum_{l=1}^{n} \mathrm{E}[\boldsymbol{Y}_l] && (11.9)\\ &= n\boldsymbol{\pi} && (11.10) \end{aligned}$$

with the third line from the identically distributed nature of $\boldsymbol{Y}_l$.

Also,

$$\begin{aligned} \mathrm{Var}[\boldsymbol{Y}] &= \mathrm{Var}[\sum_{l=1}^{n} \boldsymbol{Y}_l] && (11.11)\\ &= \sum_{l=1}^{n} \mathrm{Var}[\boldsymbol{Y}_l] && (11.12)\\ &= n\Sigma && (11.13) \end{aligned}$$

with the second line coming from the independence of $\boldsymbol{Y}_l$ and the third line from the identically distributed nature of $\boldsymbol{Y}_l$.

b) Suppose $X_1^2, ..., X_m^2$ are m independent variables with chi-squared distributions $X_i^2 \sim \chi^2(n_i)$. Show that

$$\sum_{i=1}^{m} X_i^2 = \chi^2(\sum_{i=1}^{m} n_i) \tag{11.14}$$

Answer: This simply comes from noticing that each X_i^2 is a sum of n_i squared standard normal variables, which when all summed together just becomes a bigger sum of $\sum_{i=1}^{m} n_i$ squared standard normal variables.

$$\sum_{i=1}^{m} X_i^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} Z_{ij}^2 = \chi^2(\sum_{i=1}^{m} n_i) \tag{11.15}$$

c) Suppose X^2 has the distribution $\chi^2(n)$. Prove that

$$\mathrm{E}[X^2] = n \tag{11.16}$$
$$\mathrm{Var}[X^2] = 2n \tag{11.17}$$

Hint: Your solution may require you to assume or show that $\mathrm{E}[Z^4] = 3$, where $Z \sim \mathcal{N}(0, 1)$.

Answer: Note that $X^2 = \sum_{i=1}^{n} Z_i^2$ with $Z_i \sim \mathcal{N}(0, 1)$. Therefore

$$\mathrm{E}[X^2] = \sum_{i=1}^{n} \mathrm{E}[Z_i^2] = n \tag{11.18}$$

since $\mathrm{E}[Z_i^2] = \mathrm{Var}[Z_i] + (\mathrm{E}[Z_i])^2 = 1$.

Also,

$$\mathrm{Var}[X^2] = \mathrm{Var}[\sum_{i=1}^{n} Z_i^2] = \sum_{i=1}^{n} \mathrm{Var}[Z_i^2] = n\mathrm{Var}[Z_i^2] \tag{11.19}$$

with

$$\mathrm{Var}[Z_i^2] = \mathrm{E}[Z_i^4] - (\mathrm{E}[Z_i^2])^2 = 3 - 1^2 = 2 \tag{11.20}$$

To show that $\mathrm{E}[Z_i^4] = 3$, we choose to show that $\mathrm{E}[Z^{n+1}] = n\mathrm{E}[Z^{n-1}]$, which clearly yields the required result. With $\phi(z) = \frac{1}{2\pi} e^{-z/2}$, we have

$$\begin{aligned}
\mathrm{E}[Z^{n+1}] &= \int_{-\infty}^{\infty} z^{n+1} \phi(z) dz && (11.21)\\
&= \int_{-\infty}^{\infty} z^n z \phi(z) dz && (11.22)\\
&= -\int_{-\infty}^{\infty} z^n \phi'(z) dz && (11.23)\\
&= -z^n \phi(z)\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} n z^{n-1} \phi(z) dz && (11.24)\\
&= 0 + n\mathrm{E}[Z^{n-1}] && (11.25)
\end{aligned}$$

where the third line comes from the fact that $\phi'(z) = -z\phi(z)$, and the fourth line comes from integration by parts with $u = z^n$ and $v' = \phi'(z)$.

Chapter 2

Q2-1: Quickfire Questions

a) What is the difference between a Poisson, Multinomial and product Multinomial sampling scheme?

Answer: See Sections 2.1.2 and 2.3.

b) What does an odds of 1.8 mean relative to success probability π?

Answer: An odds of 1.8 means that the probability of success is 1.8 times greater than the probability of failure.

Q2-2: Sampling Schemes

Write down statements for the expectation and variance of a variable following a product Multinomial sampling scheme as described in Section 2.3.3.

Answer: Let $\boldsymbol{N}_i = (N_{i1}, ..., N_{iJ})$ and $\boldsymbol{\pi}_i^\star = (\pi_{1|i}, ..., \pi_{J|i})$ for $i = 1, ..., I$. Then

$$\begin{aligned} \mathrm{E}[\boldsymbol{N}_i] &= n_{i+}\boldsymbol{\pi}_i^\star & (11.26)\\ \mathrm{Var}[\boldsymbol{N}_i] &= n_{i+}(\mathrm{diag}(\boldsymbol{\pi}_i^\star) - \boldsymbol{\pi}_i^\star \boldsymbol{\pi}_i^{\star,T}) & (11.27)\\ \mathrm{Cov}[\boldsymbol{N}_i, \boldsymbol{N}_{i'}] &= 0_M \qquad i \neq i' & (11.28) \end{aligned}$$

where 0_M is a matrix of zeroes.

Q2-3: Fatality of Road Traffic Accidents

Table 10.1 shows fatality results for drivers and passengers in road traffic accidents in Florida in 2015, according to whether the person was wearing a shoulder and lap belt restraint versus not using one. Find and interpret the odds ratio.

Answer: The sample odds ratio is calculated as

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{433 \times 554883}{8049 \times 570} = 52.37 \qquad (11.29)$$

Thus, the odds of a road traffic accident being fatal are *estimated to be*[1] 52.37 times greater if no restraint is used relative to if one is used.

Q2-4: Difference of Proportions or Odds Ratios?

A 20-year study of British male physicians (Doll and Peto [1976]) noted that the proportion who died from lung cancer was 0.00140 per year for cigarette smokers and 0.00010 per year for non-smokers. The proportion who died from heart disease was 0.00669 for smokers and 0.00413 for non-smokers.

a) Describe and compare the association of smoking with lung cancer and with heart disease using the difference of proportions.

[1] Crucially, using the data, we estimate...

b) Describe and compare the association of smoking with lung cancer and heart disease using the odds ratio.

c) Which response (lung cancer or heart disease) is more strongly related to cigarette smoking, in terms of increased proportional risk to the individual?

d) Which response (lung cancer or heart disease) is more strongly related to cigarette smoking, in terms of the reduction in deaths that could occur with an absence of smoking?

Answer:

a) Difference of proportions:

- Lung cancer: 0.0014 - 0.0001 = 0.0013;
- Heart disease: 0.00669 - 0.00413 = 0.00256.

Using the difference of proportions, the data suggests that cigarette smoking has a bigger impact on heart disease.

b) Odds ratio:

- Lung cancer (L): The sample odds for smokers (S) is given by $\omega_{L,S} = 0.0014/0.0086$, and the sample odds for non-smokers (N) is $\omega_{L,N} = 0.0001/0.9999$. The sample odds ratio is therefore $\omega_{L,S}/\omega_{L,N} = 14.02$.
- Heart disease (H): The sample odds for smokers (S) is given by $\omega_{H,S} = 0.00669/0.99331$, and the sample odds for non-smokers (N) is $\omega_{H,N} = 0.00413/0.99587$. The sample odds ratio is therefore $\omega_{H,S}/\omega_{H,N} = 1.624$.

The odds of dying from lung cancer are *estimated to be* 14.02 times higher for smokers than for non-smokers, whilst the odds of dying from heart disease are *estimated to be* 1.624 times higher for smokers than non-smokers. Thus, using the sample odds ratio, the data suggests that cigarette smoking has a bigger impact on lung cancer.

c) For measure based on increased proportional risk, we use the sample odds ratios above; lung cancer has an odds ratio of 14.02 compared to heart disease with an odds ratio of 1.624. Therefore, increased proportional risk to the individual smoker is much higher for lung cancer than heart disease relative to the corresponding risks for a non-smoker.

d) The difference of proportions describes excess deaths due to smoking. That is, if $N =$ *number of smokers in population*, *we predict* there would be $0.00130N$ fewer deaths per year from lung cancer if they had never smoked, and $0.00256N$ fewer deaths per year from heart disease. Thus (based on this study), elimination of cigarette smoking is predicted to have the biggest impact on deaths due to heart disease.

Q2-5: Asymptotic Distribution of X^2

a) Show that Equation (2.50) in Section 2.4.2 holds.

Answer: To verify Equation (2.50), let

$$W = \Sigma^{\star} \left(\Sigma^{\star}\right)^{-1} \tag{11.30}$$

Then we can verify that

$$\begin{aligned} W_{jj} &= \pi_j(1-\pi_j)\left(\frac{1}{\pi_j}+\frac{1}{\pi_k}\right) - \sum_{i=1,i\neq j}^{k-1}\frac{\pi_i\pi_j}{\pi_k} && (11.31)\\ &= 1-\pi_j+\frac{\pi_j}{\pi_k}(1-\pi_j-\sum_{i=1,i\neq j}^{k-1}\pi_i) && (11.32)\\ &= 1-\pi_j+\frac{\pi_j}{\pi_k}\pi_k && (11.33)\\ &= 1 && (11.34)\\ W_{jl} &= \pi_j(1-\pi_j)\frac{1}{\pi_k} - \pi_j\pi_l\left(\frac{1}{\pi_l}+\frac{1}{\pi_k}\right) - \frac{1}{\pi_k}\sum_{i=1,i\neq j}^{k-1}\pi_i\pi_j && (11.35)\\ &= \frac{\pi_j}{\pi_k}(1-\pi_j-\pi_l-\sum_{i=1,i\neq j}^{k-1}\pi_i) - \pi_j\frac{\pi_l}{\pi_l} && (11.36)\\ &= \frac{\pi_j}{\pi_k}\pi_k - \pi_j && (11.37)\\ &= 0 && (11.38) \end{aligned}$$

b) Show that Equation (2.51) in Section 2.4.2 holds.

Answer: To verify Equation (2.51) we have that

$$\begin{aligned} m(\bar{\boldsymbol{Y}}-\pi^\star)^T(\Sigma^\star)^{-1}(\bar{\boldsymbol{Y}}-\pi^\star) &= m\frac{1}{\pi_k}\sum_{i,j=1}^{k-1}(\bar{X}_i-\pi_i)(\bar{X}_j-\pi_j) && (11.39)\\ &\quad + m\sum_{i=1}^{k-1}\frac{1}{\pi_i}(\bar{X}_i-\pi_i)^2 && (11.40)\\ & && (11.41) \end{aligned}$$

where we have that

$$\begin{aligned} \sum_{i,j=1}^{k-1}(\bar{X}_i-\pi_i)(\bar{X}_j-\pi_j) &= \sum_{i=1}^{k-1}\left((\bar{X}_i-\pi_i)\sum_{j=1}^{k-1}(\bar{X}_j-\pi_j)\right) && (11.42)\\ &= -(\bar{X}_k-\pi_k)\sum_{i=1}^{k-1}(\bar{X}_i-\pi_i) && (11.43)\\ &= (\bar{X}_k-\pi_k)^2 && (11.44) \end{aligned}$$

thus verifying the result.

Q2-6: Maximum Likelihood by Lagrange Multipliers

a) Consider a Multinomial sampling scheme, and that X and Y are independent. We need to find the MLE of $\boldsymbol{\pi}$, but now we have that

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad (11.45)$$

The log likelihood is

$$l(\boldsymbol{\pi}) \propto \sum_{i,j} n_{ij} \log(\pi_{ij}) \tag{11.46}$$

$$= \sum_i n_{i+} \log(\pi_{i+}) + \sum_j n_{+j} \log(\pi_{+j}) \tag{11.47}$$

Use the method of Lagrange multipliers to show that

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n_{++}} \tag{11.48}$$

$$\hat{\pi}_{+j} = \frac{n_{+j}}{n_{++}} \tag{11.49}$$

Answer: The Lagrange function is

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\lambda}) = l(\boldsymbol{\pi}) - \lambda_1\Big(\sum_i \pi_{i+} - 1\Big) - \lambda_2\Big(\sum_j \pi_{+j} - 1\Big) \tag{11.50}$$

$$= \sum_i n_{i+} \log(\pi_{i+}) + \sum_j n_{+j} \log(\pi_{+j}) - \lambda_1\Big(\sum_i \pi_{i+} - 1\Big) - \lambda_2\Big(\sum_j \pi_{+j} - 1\Big)$$

Local optima $\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}}$ will satisfy:

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})}{\partial \pi_{i+}} = 0 \qquad i = 1, ..., I \tag{11.51}$$

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})}{\partial \pi_{+j}} = 0 \qquad j = 1, ..., J \tag{11.52}$$

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})}{\partial \lambda_1} = 0 \tag{11.53}$$

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})}{\partial \lambda_2} = 0 \tag{11.54}$$

which in this case means satisfy:

$$\frac{n_{i+}}{\pi_{i+}} - \lambda_1 = 0 \qquad i = 1, ..., I \tag{11.55}$$

$$\frac{n_{+j}}{\pi_{+j}} - \lambda_2 = 0 \qquad j = 1, ..., J \tag{11.56}$$

$$\sum_i \pi_{i+} = 1 \tag{11.57}$$

$$\sum_j \pi_{+j} = 1 \tag{11.58}$$

and hence

$$n_{i+} = \hat{\lambda_1} \hat{\pi}_{i+} \tag{11.59}$$

$$\implies \sum_i n_{i+} = \hat{\lambda}_1 \sum_i \hat{\pi}_{i+} \tag{11.60}$$

$$\implies \hat{\lambda}_1 = n_{++} \tag{11.61}$$

and similarly that

$$\hat{\lambda}_2 = n_{++} \tag{11.62}$$

Thus

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n_{++}} \tag{11.63}$$

$$\hat{\pi}_{+j} = \frac{n_{+j}}{n_{++}} \tag{11.64}$$

b) Using the method of Lagrange multipliers, show that Equation (2.100) of Section 2.4.4.1.1 holds.

Answer:

$$l(\boldsymbol{\pi}) = \sum_{i,j} n_{ij} \log(\pi_{j|i}) \tag{11.65}$$

$$= \sum_{i,j} n_{ij} \log(\pi_{+j}) \tag{11.66}$$

$$= \sum_{j} n_{+j} \log(\pi_{+j}) \tag{11.67}$$

Since $\sum_j \pi_{+j} = 1$, the Lagrange function is

$$L(\boldsymbol{\pi}, \lambda) = l(\boldsymbol{\pi}) - \lambda(\sum_j \pi_{+j} - 1) \tag{11.68}$$

$$= \sum_j n_{+j} \log(\pi_{+j}) - \lambda(\sum_j \pi_{+j} - 1) \tag{11.69}$$

Local optima $\hat{\boldsymbol{\pi}}, \hat{\lambda}$ will satisfy:

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\lambda})}{\partial \pi_{+j}} = 0 \qquad j = 1, ..., J \tag{11.70}$$

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\lambda})}{\partial \lambda} = 0 \tag{11.71}$$

which implies

$$\frac{n_{+j}}{\hat{\pi}_{+j}} - \hat{\lambda} = 0 \qquad j = 1, ..., J \tag{11.72}$$

$$\sum_j \hat{\pi}_{+j} - 1 = 0 \tag{11.73}$$

and hence

$$\sum_{j=1}^{J} n_{+j} = \hat{\lambda} \tag{11.74}$$

$$\implies \hat{\lambda} = n_{++} \tag{11.75}$$

and thus

$$\hat{\pi}_{+j} = \frac{n_{+j}}{n_{++}} \tag{11.76}$$

Q2-7: Second-Order Taylor Expansion

Show that Approximation (2.116) of Section 2.4.5.1 holds.

Answer: $f(x)$ as stated and its first two derivatives are given as

$$f(x) = x \log \frac{x}{x_0} \tag{11.77}$$

$$f'(x) = \log \frac{x}{x_0} + 1 \tag{11.78}$$

$$f''(x) = \frac{1}{x} \tag{11.79}$$

Taking a second-order Taylor expansion around x_0 leads to

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \tag{11.80}$$

However, $f(x_0) = 0$. Also $f'(x_0) = 0$ by definition of the fact that we assume the point x_0 we are taking a Taylor expansion around is a maximum, so

$$f(x) \approx \frac{1}{2x_0}(x - x_0)^2 \tag{11.81}$$

Under $\mathcal{H}_0$, we assume n_{ij} will be close to $\hat{E}_{ij}$. Therefore, we can utilise the above second-order Taylor expansion approximation, with $x = n_{ij}$ and $x_0 = \hat{E}_{ij}$, to give that

$$f(n_{ij}) \approx \frac{(n_{ij} - \hat{E}_{ij})^2}{2\hat{E}_{ij}} \tag{11.82}$$

so that

$$G^2 = 2 \sum_{i,j} f(n_{ij}) = \sum_{i,j} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{11.83}$$

Q2-8: Relative Risk

a) In 1998, A British study reported that "*Female smokers were 1.7 times more vulnerable than Male smokers to get lung cancer.*" We don't investigate whether this is true or not, but is 1.7 the odds ratio or the relative risk[2]? Briefly (one sentence maximum) explain your answer.

Answer: I would say that 1.7 is the relative risk, as it seems that the statement is claiming that the probability of one event happening is 1.7 times greater than the probability of the other. This by assuming that the definition of *vulnerability* is with regard to the absolute probability of the event occurring (for one group to another) rather than a ratio of the odds of it occurring. Having said this, there is vaguery in the wording - there is potential for it to be interpreted differently. Such vaguery in this one line alone highlights how we need to be careful to be precise in the phrasing of our results, so as not to confuse, or (deliberately...?) mislead people.

[2]Relative risk was introduced in Section 2.1.3.2.

b) A National Cancer institute study about tamoxifen and breast cancer reported that the women taking the drug were 45% less likely to experience invasive breast cancer than were women taking a placebo. Find the relative risk for (i) those taking the drug compared with those taking the placebo, and (ii) those taking the placebo compared with those taking the drug.

Answer: i) Defining π_D and π_P to be the probabilities of invasive breast cancer given the drug and the placebo respectively, we have that

$$\pi_D = (1 - 0.45)\pi_P \implies \pi_D = 0.55\pi_P \implies \pi_D/\pi_P = 0.55 \tag{11.84}$$

ii) From part (i), we have that

$$\pi_P/\pi_D = 1/0.55 = 1.82 \tag{11.85}$$

Q2-9: The Titanic

For adults who sailed on the *Titanic* on its fateful voyage, the odds ratio between gender (categorised as *Female (F)* or *Male (M)*), and survival (categorised as *yes (Y)* or *no (N)*) was 11.4 (Dawson [1995]).

a) It is claimed that "*The Probability of survival for women was 11.4 times that for men*". i) What is wrong with this interpretation? ii) What should the correct interpretation be? iii) When would the quoted interpretation be approximately correct?

Answer: i) and ii) The probability of survival for women was not 11.4 times that for men. The (sample[3]) odds of survival for women (not accounting for other factors) was 11.4 times greater than for men.

iii) Let $\pi_{YF}, \pi_{NF}, \pi_{YM}, \pi_{NM}$ be the probabilities of survival or not for women and men respectively. The quoted interpretation would be approximately correct when the probabilities of success of both events are small, hence $\pi_{NF} \approx 1$ and $\pi_{NM} \approx 1$ so that:

$$\frac{\pi_{YF}/\pi_{NF}}{\pi_{YM}/\pi_{NM}} \approx \frac{\pi_{YF}}{\pi_{YM}} \tag{11.86}$$

b) The odds of survival for women was 2.9. Find the proportion of each gender who survived.

Answer: We have that

$$\frac{\pi_{YF}}{\pi_{NF}} = \frac{\pi_{Y|F}}{\pi_{N|F}} = 2.9 \tag{11.87}$$

$$\implies \quad \pi_{Y|F} = 2.9(1 - \pi_{Y|F}) \tag{11.88}$$

$$\implies \quad \pi_{Y|F} = \frac{2.9}{3.9} = \frac{29}{39} \tag{11.89}$$

[3]In brackets here as in this case we are also talking about the population of interest.

We also have that:

$$\frac{\pi_{Y|F}/\pi_{N|F}}{\pi_{Y|M}/\pi_{N|M}} = 11.4 \tag{11.90}$$

$$\implies \frac{\pi_{Y|M}}{\pi_{N|M}} = \frac{2.9}{11.4} = \frac{29}{114} \tag{11.91}$$

$$\implies \pi_{Y|M} = \frac{29}{114}(1 - \pi_{Y|M}) \tag{11.92}$$

$$\implies \frac{143}{114}\pi_{Y|M} = \frac{29}{114} \tag{11.93}$$

$$\implies \pi_{Y|M} = \frac{29}{143} \tag{11.94}$$

Q2-10: Test and Reality

For a diagnostic test of a certain disease, let π_1 denote the probability that the diagnosis is positive given that a subject has the disease, and let π_2 denote the probability that the diagnosis is positive given that a subject does not have the disease. Let τ denote the probability that a subject has the disease.

a) More relevant to a patient who has received a positive diagnosis is the probability that they truly have the disease. Given that a diagnosis is positive, show that the probability that a subject has the disease (called the *positive predictive value*) is

$$\frac{\pi_1\tau}{\pi_1\tau + \pi_2(1-\tau)} \tag{11.95}$$

Answer: As defined in the question, let

$$\pi_1 = P(\text{positive diagnosis given presence of disease}) = P(T^+|D^+) \tag{11.96}$$

$$\pi_2 = P(\text{positive diagnosis given absence of disease}) = P(T^+|D^-) \tag{11.97}$$

$$\tau = P(\text{presence of disease}) = P(D^+) \tag{11.98}$$

Then, using Bayes theorem, we have that

$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+)} \tag{11.99}$$

$$= \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} \tag{11.100}$$

$$= \frac{\pi_1\tau}{\pi_1\tau + \pi_2(1-\tau)} \tag{11.101}$$

b) Suppose that a diagnostic test for a disease has both sensitivity and specificity equal to 0.95, and that $\tau = 0.005$. Find the probability that a subject truly has the disease given a positive diagnostic test result.

Answer: Recall that

$$\text{Sensitivity:} \quad P(T^+|D^+) = 0.95 \tag{11.102}$$

$$\text{Specificity:} \quad P(T^-|D^-) = 0.95 \tag{11.103}$$

Table 11.1: Contingency table of probabilities cross-classifying presence or absence of disease against positive or negative diagnostic test.

	Test: Positive	**Negative**	**Sum**
Disease: Presence	0.00475	0.00025	**0.005**
Disease: Absence	0.04975	0.94525	**0.995**
Sum	**0.05450**	**0.94550**	**1.000**

Table 11.2: n_{ij}.

	Happiness: Not too Happy	**Pretty Happy**	**Very Happy**	**Sum**
Income: Above Average	21	159	110	**290**
Income: Average	53	372	221	**646**
Income: Below Average	94	249	83	**426**
Sum	**168**	**780**	**414**	**1362**

and we also have that $\tau = P(D^+) = 0.005$.

Then

$$P(D^+|T^+) = \frac{0.95 \times 0.005}{0.95 \times 0.005 + (1 - 0.95)(1 - 0.005)} = 0.087 \tag{11.104}$$

c) Create a 2×2 contingency table of cross-classified probabilities for presence or absence of the disease and positive or negative diagnostic test result.

Answer: See Table 11.1.

d) Calculate the odds ratio and interpret.

Answer:

$$r_{12} = \frac{0.00475 \times 0.94525}{0.00025 \times 0.04975} = 361 \tag{11.105}$$

The odds of a positive test result are 361 times higher for a subject for whom the disease is present than a subject for whom the disease is absent. Equivalently, the odds of presence of the disease are 361 times higher for a subject with a positive test result than a subject with a negative test result.

Q2-11: Happiness and Income

Table 10.2 shows data from a General Social Survey cross-classifying a person's perceived happiness with their family income.

a) Perform a χ^2 test of independence between the two variables.

Answer: Table 11.2 shows the observed data with row and column sum totals.

Table 11.3: Estimated E_{ij} values.

	Happiness: Not too Happy	**Pretty Happy**	**Very Happy**	**Sum**
Income: Above Average	35.77093	166.0793	88.14978	**290**
Income: Average	79.68282	369.9559	196.36123	**646**
Income: Below Average	52.54626	243.9648	129.48899	**426**
Sum	**168.00000**	**780.0000**	**414.00000**	**1362**

Table 11.4: X_{ij}^2 values.

	Happiness: Not too Happy	**Pretty Happy**	**Very Happy**
Income: Above Average	6.099373	0.3017620	5.416147
Income: Average	8.935086	0.0112936	3.091592
Income: Below Average	32.702862	0.1039235	16.690423

Table 11.3 shows the estimated cell values under independence.

Table 11.4 shows the X_{ij}^2 values for each cell.

We thus get that

$$X^2 = \sum_{i,j} X_{ij}^2 = 73.352... \tag{11.106}$$

Comparing to a chi-square distribution with 4 degrees of freedom, we have that

$$P(\chi_4^{2,\star} \geq 73.352) \leq 0.0005 \tag{11.107}$$

hence the test provides strong evidence to reject $\mathcal{H}_0$ that the two variables are independent.

b) Calculate and interpret the adjusted residuals for the four corner cells of the table.

Answer: Adjusted standardised residuals are presented in Table 11.5.

The top-left and bottom-right cell adjusted residuals provide evidence that fewer people are in those cells in the population than if the variables were independent. The top-right and bottom-left cell adjusted residuals provide evidence that more people are in those cells in the population than if the variables were independent. Although not required for the question, by calculating all residuals, we can see that there is also evidence of more people on average

Table 11.5: Adjusted standardised Residuals.

	Happiness: Not too Happy	**Pretty Happy**	**Very Happy**
Income: Above Average	-2.973173	-0.9472192	3.144277
Income: Average	-4.403194	0.2242210	2.906749
Income: Below Average	7.367666	0.5948871	-5.907023

Table 11.6: Cross-classification of guess versus truth in the tea tasting experiment of Fisher (1935).

	Truth: Milk first	**Tea first**	**Sum**
Guess: Milk first	3	1	**4**
Guess: Tea first	1	3	**4**
Sum	**4**	**4**	**8**

income that are very happy, and less people on average income that are not too happy, than if the variables were independent.

Q2-13: Tea! (Fisher's Exact Test of Independence)

Regarding the quote (not repeated here to save space) in the corresponding problem from Fisher [1937]:

a) From the text, we know that there are 4 cups with milk added first and 4 with tea infusion added first. How many distinct orderings can these 8 cups to be tasted take, in terms of type.

Answer: There is a total of $\frac{8!}{4!4!} = 70$ distinct orderings of these cups.

b) Note that the lady also knows that there are four cups of each type, and must group them into two sets of four (those she thinks had milk added first, and those she thinks had tea infusion added first). Given that the lady guesses milk first three times when indeed the milk was added first, cross-classify the lady's guesses against the truth in a 2×2 contingency table.

Answer: Observe that we are in a sampling scenario with fixed row and column sums. Therefore we have all the information we need to display the results of this experiment in a contingency table, as given by Table 11.6.

c) Fisher presented an *exact*[4] solution for testing the null hypothesis

$$\mathcal{H}_0 : r_{12} = 1 \tag{11.108}$$

against the one-sided alternative

$$\mathcal{H}_1 : r_{12} > 1 \tag{11.109}$$

for contingency tables with fixed row and column sums.
What hypothesis does $\mathcal{H}_0$ correspond to in the context of the tea tasting test described above? Write down an expression for $P(N_{11} = t)$ under $\mathcal{H}_0$. Thus, perform a (Fisher's exact) hypothesis test to test the lady's claim that she can indeed discriminate whether the milk or tea infusion was first added to the cup.

Answer: $\mathcal{H}_0$ corresponds to the situation where the lady is purely guessing whether milk or tea infusion was added first (with no ability to discriminate based on taste).

[4] *Exact* in the sense that the probabilities of any possible outcome can be calculated exactly.

Under $\mathcal{H}_0$, given n_{1+}, n_{+1} and n_{++}, we have that $N_{11} \sim \mathcal{H}g(N = n_{++}, M = n_{1+}, q = n_{+1})$[5] so that:

$$P(N_{11} = t) = \frac{\binom{n_{1+}}{x}\binom{n_{++}-n_{1+}}{n_{+1}-x}}{\binom{n_{++}}{n_{+1}}} \qquad \max(0, n_{+1} + n_{1+} - n_{++}) \leq x \leq \min(n_{+1}, n_{1+}) \quad (11.110)$$

A p-value is obtained by calculating the sum of the extreme probabilities, where extreme is in the direction of the alternative hypothesis. Let t_{obs} denote the observed value of N_{11}, then

$$\begin{aligned} P(N_{11} \geq t_{obs}) &= P(N_{11} \geq 3) && (11.111)\\ &= P(N_{11} = 3) + P(N_{11} = 4) && (11.112)\\ &= \frac{\binom{4}{3}\binom{8-4}{4-3}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{8-4}{4-4}}{\binom{8}{4}} && (11.113)\\ &= \frac{17}{70} && (11.114) \end{aligned}$$

Hence the test does not provide evidence of the lady's ability at any standard level of significance.

d) Suppose the lady had correctly classified all eight cups as having either milk or tea infusion added first. Would Fisher's exact hypothesis test provide evidence of her ability now?

Answer: You may repeat the above steps of part (c), and see that:

$$P(N_{11} \geq 4) = P(N_{11} = 4) = \frac{1}{70} \quad (11.115)$$

hence the test would provide evidence of the lady's ability at the 5% level of significance, but the power is such that this test could never provide evidence at the 1% level of significance. For this, we would need more cups in the test, for example, five of each type - feel free to play around with this scenario. How many cups of each type (assuming there is an even split) would be required such that if the lady misclassifies one cup of each type, the hypothesis test still provides evidence for her ability at the 1% level of significance?

Q2-13: US Presidential Elections

Table 10.3 cross-classifies a sample of votes in the 2008 and 2012 US Presidential elections. Test the null hypothesis that vote in 2008 was independent from vote in 2012 by estimating, and finding a 95% confidence interval for, the population odds ratio.

Answer:

[5]Note that this distribution follows in the context of this scenario since we view the lady as randomly guessing which $q = n_{+1} = 4$ cups had milk added first from the total of $N = n_{++} = 8$ cups and seeing how many are the type of interest, namely those for which milk really was added first $M = n_{+1} = 4$. Having explained it this way, we could also view it as $\mathcal{H}g(N = n_{++}, M = n_{+1}, q = n_{1+})$.

Table 11.7: Local Odds Ratios between successive row categories i and $i+1$ and column categories j and $j+1$, these being indicated in each case by the row and column names.

	Happiness: Not very-Pretty	**Pretty-Very**
JobSecOK: not at all-not too	1.34	2.23
JobSecOK: not too-somewhat	1.73	0.90
JobSecOK: somewhat-very	1.68	1.63

We wish to test the following hypotheses:

$$\mathcal{H}_0 : \qquad r_{12} = 1 \tag{11.116}$$
$$\mathcal{H}_0 : \qquad r_{12} \neq 1 \tag{11.117}$$

An estimate for this odds ratio is given by

$$\hat{r}_{12} = \frac{802 \times 494}{34 \times 53} = 219.8602 \tag{11.118}$$

A $(1-\alpha)$ confidence interval for $\log r_{12}$ is given by

$$\begin{aligned} \log \hat{r} \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}} &= \log 219.8602 \pm 1.96\sqrt{\frac{1}{802} + \frac{1}{34} + \frac{1}{53} + \frac{1}{494}} \\ &= (4.9480, 5.8380) \end{aligned}$$

so that a Wald confidence interval for $\hat{r}_{12}$ is given by

$$(e^{4.9480}, e^{5.8380}) = (140.9, 343.1) \tag{11.119}$$

Since $r_{0,12} = 1$ lies outside of this interval, the test rejects $\mathcal{H}_0$ at the 5% level of significance.

Q2-14: Job Security and Happiness

Consider the table presented in Figure 10.1 summarising the responses for extent of agreement to the statement "job security is good" (JobSecOK, or Job Security) and general happiness (Happiness) from the 2018 US General Social Surveys. Additional possible responses of *don't know* and *no answer* are omitted here.

a) Calculate a minimal set of local odds ratios. Interpret and discuss.

Answer: Cell $(1,1)$ corresponds to the relative difference between *not at all true* and *not too true* in the odds of *pretty happy* to *very happy*. This is estimated by

$$\hat{r}^L_{11} = \frac{15 \times 47}{21 \times 25} = 1.34. \tag{11.120}$$

Calculating the remaining odds ratios yields Table 11.7.

In general, all of the local odds ratios are greater than 1, thus suggesting a positive association between consecutive levels of Job Security and Happiness. The exception to this is cell $(2,2)$, which is slightly less than 1.

Table 11.8: Cumulative Odds Ratios between successive row categories i and $i+1$, with the cumulative split at j (this being in the lower combined category) being denoted by the column name.

	Happiness: Not very	**Pretty**
JobSecOK: not at all-not too	1.62	2.47
JobSecOK: not too-somewhat	1.68	1.04
JobSecOK: somewhat-very	1.98	1.77

Table 11.9: Cumulative Odds Ratios between row categories i and $I = 4$, with the cumulative split at j (this being in the lower combined category) being denoted by the column name.

	Happiness: Not very	**Pretty**
JobSecOK: not at all	5.38	4.55
JobSecOK: not too	3.32	1.84
JobSecOK: somewhat	1.98	1.77

b) Calculate a minimal set of odds ratios, treating job security locally and happiness cumulatively. Interpret and discuss.

Answer: Cell $(1, 1)$ corresponds to the relative difference between those who say *not too happy* and those who say either *pretty happy* and *very happy*, in the odds of saying *not at all true* to *not too true*. This is estimated by

$$\hat{r}_{11}^{C_Y} = \frac{15/(25+5)}{21/(47+21)} = \frac{15 \times (47+21)}{21 \times (25+5)} = 1.619 \tag{11.121}$$

Calculating the remaining odds ratios yields Table 11.8.

In general, all of the odds ratios are greater than 1, thus suggesting a positive association between consecutive levels of Job Security with the grouped levels of Happiness in each case (regardless of whether *Pretty happy* is grouped with *not too happy* or *very happy*).

c) Calculate a minimal set of odds ratios, treating job security as a nominal variable (taking *very true* as the reference category) and happiness as a cumulative variable. Interpret and discuss.

Answer: Cell $(1, 1)$ corresponds to the relative difference between those who say *not too happy* and those who say either *pretty happy* and *very happy*, in the odds of saying *not at all true* to *very true*. This is estimated by

$$\hat{r}_{11}^{C_Y} = \frac{15/(25+5)}{73/(474+311)} = \frac{15 \times (474+311)}{73 \times (25+5)} = 5.377 \tag{11.122}$$

Calculating the remaining odds ratios yields Table 11.9.

In general, all of the odds ratios are greater than 1, thus suggesting a positive association between the level of job security with the (grouped) level of Happiness.

Table 11.10: Global Odds Ratios, with the cumulative split at i, j (these themselves being in the corresponding lower combined categories) being denoted by the row and column name.

	Happiness: Not very	**Pretty**
JobSecOK: not at all	3.80	3.72
JobSecOK: not too	3.03	1.99
JobSecOK: somewhat	2.41	1.90

Table 11.11: Cumulative contingency table, with the categories grouped according to the row and column names.

	Happiness: Not very-Pretty	**Very**
JobSecOK: not at all-not too	108	26
JobSecOK: somewhat-very	859	411

d) Calculate a minimal set of global odds ratios, that is, treating both job security and happiness as cumulative variables.

Answer: Cell $(1, 1)$ corresponds to the odds ratio obtained by combining *pretty happy* and *very happy* into one category for Happiness, and combining *not too true*, *somewhat true* and *very true* into one category for Job Security. This is estimated by

$$\hat{r}_{11}^{G} = \frac{15 \times (47 + 21 + 248 + 100 + 474 + 311)}{(21 + 64 + 73) \times (25 + 5)} = 3.80 \tag{11.123}$$

Calculating the remaining global odds ratios yields Table 11.10.

In general, all of the odds ratios are greater than 1, thus suggesting a positive association between the grouped levels of job security with the grouped levels of Happiness.

e) Calculate a 95% confidence interval for the global odds ratio, with the first two categories of each variable being grouped together into one category in each case.

Answer: We can readily adopt the standard formula for a confidence interval of a log odds ratio for the global odds ratio. We just need to work with the 2×2 table of counts induced by any accumulation. In our case, we calculate

$$\begin{aligned} 15 + 25 + 21 + 47 &= 108 \\ 5 + 21 &= 26 \\ 64 + 248 + 73 + 474 &= 859 \\ 100 + 311 &= 411 \end{aligned}$$

so that the cumulative contingency table is given by Table 11.11.

We therefore have that

$$\hat{r}_{22}^{G} = \frac{108 \times 411}{26 \times 859} = 1.99 \tag{11.124}$$

Table 11.12: Cross-multiplied XY scores.

	Happiness: Not very	**Pretty**	**Very**
JobSecOK: not at all	1	2	3
JobSecOK: not too	2	4	6
JobSecOK: somewhat	3	6	9
JobSecOK: very	4	8	12

and the 95% confidence interval for $\log r_{22}^G$ is given by

$$\log 1.99 \pm 1.96\sqrt{\frac{1}{108}+\frac{1}{26}+\frac{1}{859}+\frac{1}{411}} = (0.24, 1.13) \tag{11.125}$$

Thus a confidence interval for r_{22}^G is given by

$$(e^{0.24}, e^{1.13}) = (1.27, 3.10) \tag{11.126}$$

hence the test rejects the null hypothesis of no association between these combined groups of job security and happiness.[6]

f) Perform a linear trend test to assess whether there is any evidence of association between Happiness and Job Security.

Answer: We wish to test:

$$\mathcal{H}_0 : \rho_{XY} = 0 \tag{11.127}$$

against

$$\mathcal{H}_1 : \rho_{XY} \neq 0 \tag{11.128}$$

We choose to utilise equally spaced scores set equal to the order of each category, with cross-multplied XY scores presented in Table 11.12.

We calculate the following:

$$\begin{aligned}
\sum_{k=1}^{n_{++}} x_k y_k = \sum_{i,j} x_i y_j n_{ij} &= 1\times 15 + 2\times 25 + 3\times 5 \\
&\quad + 2\times 21 + 4\times 47 + 6\times 21 \\
&\quad + 3\times 64 + 6\times 248 + 9\times 100 \\
&\quad + 4\times 73 + 8\times 474 + 12\times 311 = 10832 \\
\sum_{k=1}^{n_{++}} x_k = \sum_{i=1}^{4} x_i n_{i+} &= 1\times 45 + 2\times 89 + 3\times 412 + 4\times 858 = 4891 \\
\sum_{k=1}^{n_{++}} y_k &= 1\times 173 + 2\times 794 + 3\times 437 = 3072 \\
\sum_{k=1}^{n_{++}} x_k^2 = \sum_{i=1}^{4} x_i^2 n_{i+} &= 1^2\times 45 + 2^2\times 89 + 3^2\times 412 + 4^2\times 858 = 17837 \\
\sum_{k=1}^{n_{++}} y_k^2 &= 1^2\times 173 + 2^2\times 794 + 3^2\times 437 = 7282
\end{aligned}$$

[6] You should repeat this analysis for some of the other global odds ratios.

so that

$$r_{XY} = \frac{1404 \times 10832 - 4891 \times 3072}{\sqrt{1404 \times 17837 - 4891^2}\sqrt{1404 \times 7282 - 3072^2}} = 0.1948... \tag{11.129}$$

and

$$M^2 = (n-1)r_{XY}^2 = 1403 \times 0.1948...^2 = 53.24 \tag{11.130}$$

We have that

$$P(\chi_1^2 \geq 53.24) \leq 10^{-12} \tag{11.131}$$

hence the test rejects the null hypothesis of independence for any reasonable level of significance, and thus provides evidence of association between job security and happiness.

Chapter 3

Q3-1: A Question of Notation

a) Based on the notation of Section 3.1, state what table (in terms of dimension, as well as conditioning and marginalising of variables) is being described by:

i. $(n_{(i)jk})$

ii. $(n_{(i)+k(l)})$

iii. $(n_{i_1 i_2 + (i_4) i_5 (i_6) +})$

Answer:

i. The two-way partial table obtained by conditioning on $X = i$.

ii. The one-way table of counts for different categories of Z, obtained by summing over all levels of variable Y for a fixed level of (or conditioning on) variables $X = i$ and $W = l$.

iii. The three-way partial marginal table obtained by summing over all levels of variables X_3 and X_7 for a fixed level of (or conditioning on) variables $X_4 = i_4$ and $X_6 = i_6$.

b) Use the notation of Section 3.1, in the context of Section 3.1.1.1, to notate and construct the conditional (on drug treatment $X = B$) table, having marginalised over clinic Z.

Answer: This 'table' could be notated $(n_{B,y,+})$ (or $(n_{B,i_2,+})$), and is simply given by the vector:

$$(\text{Success } = 20, \text{Failure } = 40) \tag{11.132}$$

that simply states the number of successes and failures for drug treatment B (regardless of clinic) in our dataset.

c) Assuming a total of 8 variables, use the notation of Section 3.1 to notate the partial marginal table obtained by summing over all levels of variables X_3, X_5 and X_7, for a fixed level of variables X_1 and X_4.

Answer: $(n_{(i_1) i_2 + (i_4) + i_6 + i_8})$, or we may add commas in between the indices to make things (perhaps?) clearer... $(n_{(i_1), i_2, +, (i_4), +, i_6, +, i_8})$

Q3-2: Conditional and Marginal Local Odds Ratios

Show that Equation (3.17) of Section 3.3.4.2 holds for local odds ratios under the conditional independence of X and Y given Z. Recall that such conditional independence means that Equation (3.16) holds.

Answer: We have that

$$r^{XZ}_{i(j)k} = \frac{\pi_{i,j,k}\pi_{i+1,j,k+1}}{\pi_{i+1,j,k}\pi_{i,j,k+1}} \tag{11.133}$$

$$= \frac{\frac{\pi_{i,+,k}\pi_{+,j,k}}{\pi_{+,+,k}} \frac{\pi_{i+1,+,k+1}\pi_{+,j,k+1}}{\pi_{+,+,k+1}}}{\frac{\pi_{i+1,+,k}\pi_{+,j,k}}{\pi_{+,+,k}} \frac{\pi_{i,+,k+1}\pi_{+,j,k+1}}{\pi_{+,+,k+1}}} \tag{11.134}$$

$$= \frac{\pi_{i,+,k}\pi_{i+1,+,k+1}}{\pi_{i+1,+,k}\pi_{i,+,k+1}} \tag{11.135}$$

$$= r^{XZ}_{ik} \tag{11.136}$$

Q3-3: Types of Independence

My example was presented throughout Section 3.3. I am happy to discuss answers to this question during the weekly office hour, and would also encourage you to discuss possible answers with your fellow students.

Q3-4: Conditional and Joint Independence

Consider an $I \times J \times K$ contingency table, with classification variables X, Y, Z. By writing down the relevant probability forms involved, or otherwise...

a) Prove that if
   - X and Y are conditionally independent given Z; and
   - X and Z are conditionally independent given Y,

then Y and Z are jointly independent of X.

Answer: We have that

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \tag{11.137}$$

$$\pi_{ijk} = \frac{\pi_{ij+}\pi_{+jk}}{\pi_{+j+}} \tag{11.138}$$

By dividing Equation (11.137) by Equation (11.138), we have that

$$\pi_{i+k}\pi_{+j+} = \pi_{ij+}\pi_{++k} \tag{11.139}$$

Now sum both sides of Equation (11.139) over j to get that

$$\pi_{i+k} = \pi_{i++}\pi_{++k} \tag{11.140}$$

By substituting Equation (11.140) into Equation (11.137), we get that

$$\pi_{ijk} = \pi_{i++}\pi_{+jk} \tag{11.141}$$

Hence Y and Z are jointly independent of X.

b) Prove that if Y and Z are jointly independent from X, then
- X and Y are conditionally independent given Z; and
- X and Z are conditionally independent given Y.

Y and Z are jointly independent of X, so we have that

$$\pi_{ijk} = \pi_{i++}\pi_{+jk} \tag{11.142}$$

Sum both sides of Equation (11.142) over j to get that

$$\begin{aligned} \pi_{i+k} &= \pi_{i++}\pi_{++k} \\ \implies \pi_{i++} &= \frac{\pi_{i+k}}{\pi_{++k}} \end{aligned} \tag{11.143}$$

By substituting Equation (11.143) in Equation (11.142), we get

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \tag{11.144}$$

hence X and Y are conditionally independent of Z.

Sum both sides of Equation (11.142) over k to get that

$$\begin{aligned} \pi_{ij+} &= \pi_{i++}\pi_{+j+} \\ \implies \pi_{i++} &= \frac{\pi_{ij+}}{\pi_{++k}} \end{aligned} \tag{11.145}$$

By substituting Equation (11.145) in Equation (11.142), we get

$$\pi_{ijk} = \frac{\pi_{ij+}\pi_{+jk}}{\pi_{++k}} \tag{11.146}$$

hence X and Z are conditionally independent of Y.

Q3-5: 1988 General Social Survey

The 1988 General Social Survey compiled by the National Opinion Research Center asked:

- "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms."

The table in Figure 10.2 summarizes opinions about health care costs (H) and the information program (I), classified also by the respondent's gender (G).

a) Compute the marginal GH-table.

Answer: The marginal table is as given by the table in Figure 11.1.

b) For the marginal GH-table, compute the MLE of the marginal odds ratio, along with a 95% confidence interval. Interpret the result.

	Health Opinion (H)		
Gender (G)	Support	Oppose	Total
Male	82	185	267
Female	125	229	354
Total	207	414	621

Figure 11.1: GH Marginal Table.

Answer: The MLE of the marginal odds ratio is

$$\hat{r}_{12}^{GH} = \frac{n_{+11}n_{+22}}{n_{+12}n_{+21}} = \frac{82 \times 229}{185 \times 125} = 0.81202 \tag{11.147}$$

In the sample, there is a slight association between respondents being Female and supporting the healthcare proposal.

The 95% confidence interval for $\log r_{12}^{GH}$ is

$$\begin{aligned} & \log r_{12}^{GH} \pm z_{0.975}\sqrt{\sum_{i,j}\frac{1}{n_{ij}}} \\ = \; & \log 0.81202 \pm 1.96 \times \sqrt{0.029967} \\ = \; & (-0.548, 0.131) \end{aligned} \tag{11.148}$$

so that the 95% confidence interval for r_{12}^{GH} is given by

$$(e^{-0.548}, e^{0.131}) = (0.578, 1.140) \tag{11.149}$$

We would not be able to reject the null hypothesis:

$$\mathcal{H}_0 : \text{Gender and Health Opinion are independent} \tag{11.150}$$

in favour of the alternative hypothesis

$$\mathcal{H}_1 : \text{Gender and Health Opinion are not independent} \tag{11.151}$$

at the 5% level of significance.

c) Perform a Mantel Haenszel Chi-Square test, at the 5% level of significance, in order to test if the Information Opinion and the Health Opinion are independent at each level of Gender.

Answer: We wish to test the following pair of hypotheses:

$\mathcal{H}_0$: I, H are independent across the partial tables at each level of G.
$\mathcal{H}_1$: I, H are not independent across the partial tables at each level of G.

$n_{i,+,k}$	1	2
1	236	295
2	31	59

$n_{+,j,k}$	1	2
1	82	125
2	185	229

$n_{+,+,k}$	267	354

Figure 11.2: Marginal sums required for Question 3-3c.

or equivalently

$$\begin{aligned}\mathcal{H}_0: &\quad r^{IH}_{12(1)} = r^{IH}_{12(2)} = 1\\ \mathcal{H}_1: &\quad r^{IH}_{12(k)} \neq 1 \text{ for some } k.\end{aligned}$$

The relevant marginal tables are as shown in Figure 11.2.

We have that

$$\begin{aligned}\hat{E}_{111} &= \frac{n_{1+1}n_{+11}}{n_{++1}} = \frac{236 \times 82}{267} = 72.479\\ \hat{E}_{112} &= \frac{n_{1+2}n_{+12}}{n_{++2}} = \frac{295 \times 125}{354} = 104.167\\ \hat{\sigma}^2_{111} &= \frac{n_{1+1}n_{2+1}n_{+11}n_{+21}}{n^2_{++1}(n_{++1}-1)} = \frac{236 \times 31 \times 82 \times 185}{267^2 \times 266} = 5.853\\ \hat{\sigma}^2_{112} &= \frac{n_{1+2}n_{2+2}n_{+12}n_{+22}}{n^2_{++2}(n_{++2}-1)} = \frac{295 \times 59 \times 125 \times 229}{354^2 \times 353} = 11.263\end{aligned}$$

The observed test statistic is

$$T_{MH,obs} = \frac{[(n_{111} - \hat{E}_{111}) + (n_{112} - \hat{E}_{112})]^2}{\hat{\sigma}^2_{111} + \hat{\sigma}^2_{112}} = \frac{[(76 - 72.479) + (114 - 104.167)]^2}{5.853 + 11.263} = 10.419 \tag{11.152}$$

Since

$$T_{MH,obs} = 10.419 \geq \chi^2_{1,0.05} = 3.841 \tag{11.153}$$

the test rejects the null hypothesis that health opinion and information opinion are independent at each level of Gender at the 5% level of significance.

d) Compute the partial (conditional) IH odds ratio at each level of Gender. Interpret the result.

Answer: For men, the MLE of the odds ratio between I and H is

$$r^{IH}_{12(1)} = \frac{n_{111}n_{221}}{n_{121}n_{211}} = \frac{76 \times 25}{160 \times 6} = 1.979 \tag{11.154}$$

For women, the MLE of the odds ratio between I and H is

$$r^{IH}_{12(2)} = \frac{n_{112}n_{222}}{n_{122}n_{212}} = \frac{114 \times 48}{11 \times 181} = 2.748 \tag{11.155}$$

These odds ratios suggest that there is a positive association between health opinion and information opinion at each level of Gender, with the association being slightly stronger for women than for men.

Q3-6: Alien Spacejets

The aliens are designing spacejets to enable them to travel between two planets. In particular, they need spacejets that can withstand the atmospheric pressure encountered whilst travelling through space between the two planets. They therefore cross-classify the survival of different spacejets on test runs in the atmosphere (Survives, Y: yes or no) against Type, Z: A or B and Colour, X: red or green. The dataset is presented in the table shown in Figure 10.3.

a) Calculate the marginal XY contingency table of the observed counts.

Answer: The XYZ-contingency table of the data presented in Figure 10.3 is presented in Figure 11.3. The marginal XY contingency table (with totals) is presented in Figure 11.4.

		Survives (Y)	
Type (Z)	Colour (X)	yes	no
A	red	20	3
	green	50	14
B	red	32	47
	green	5	15

Figure 11.3: Contingency table of the data presented in Q3-6.

	Survives (Y)		
Colour (X)	yes	no	$Y = .$
red	52	50	102
green	55	29	84
$X = .$	107	79	186

Figure 11.4: Marginal (over Z) XY contingency table.

b) Calculate an estimate of, along with a 90% confidence interval for, the marginal odds ratio of Colour and Survives.

Answer: The (local) sample odds ratio (required only when $i = 1, j = 1$ in this case) is

$$\hat{r}_{ij} = \frac{n_{i,j,+}n_{i+1,j+1,+}}{n_{i,j+1,+}n_{i+1,j,+}} = \frac{52 \times 29}{50 \times 55} = 0.5484... \tag{11.156}$$

The corresponding 90% confidence interval for the log odds ratio is given by:

$$\log \hat{r}_{ij} \pm z_{1-\alpha/2}\sqrt{\frac{1}{n_{i,j}} + \frac{1}{n_{i+1,j}} + \frac{1}{n_{i,j+1}} + \frac{1}{n_{i+1,j+1}}} \tag{11.157}$$

$$= \log 0.5484... \pm 1.645\sqrt{\frac{1}{52} + \frac{1}{55} + \frac{1}{50} + \frac{1}{29}} \tag{11.158}$$

$$= (-1.0994, -0.1021) \tag{11.159}$$

Therefore, the 90% confidence interval for the odds ratio is given by:

$$(\exp(-1.0994), \exp(-0.1021)) = (0.3331, 0.9029) \tag{11.160}$$

c) What does the estimated odds ratio tell us about the relation between Colour and Survives? Infer whether there is evidence that Colour and Survives are dependent or not.

Answer: The odds ratio suggests a negative association between Colour and Survives, that is, green spacejets typically survive the atmosphere better than red spacejets.

Based on the 90% confidence interval for the odds ratio, there is evidence to reject the null hypothesis that Colour and Survives are independent at the 10% level of significance.

d) Based on the results of Parts (a)-(c), the aliens are convinced that the dataset provides at least some evidence that, regardless of Type, green spacejets have a greater chance of surviving the intense atmospheric pressure than red spacejets. Perform reasonable analyses to:

i. illustrate to the aliens why they have jumped to the wrong conclusion, and
ii. discuss alternative inferences.

Supposing this was an excerpt from an exam question (as it may have been in the past), credit would be awarded for the clarity of your answer.

Answer:

Type: A	Survives (Y)		
Colour (X)	yes	no	$Y = .$
red	20	3	23
green	50	14	64
$X = .$	77	17	87

Type: B	Survives (Y)		
Colour (X)	yes	no	$Y = .$
red	32	47	79
green	5	15	20
$X = .$	37	62	99

Figure 11.5: Conditional (given Z) XY contingency tables.

Conditional XY- contingency tables are presented in Figure 11.5, with corresponding conditional sample odds ratios given as follows:

$$\hat{r}_{ij}^{(A)} = \frac{20 \times 14}{50 \times 3} = 1.86... \tag{11.161}$$

$$\hat{r}_{ij}^{(B)} = \frac{32 \times 15}{47 \times 5} = 2.04... \tag{11.162}$$

from which we can see that, conditional on Type, there is evidence of a positive association between Colour and Survives, i.e. red Type A spacejets typically survive the atmosphere better than green Type A spacejets, and red Type B spacejets typically survive the atmosphere better than green Type B spacejets.

The alien's incorrect conclusion arose as a result of marginalising over the confounding factor Type, because this dataset exhibits a phenomenon known as Simpson's paradox. Simpson's paradox refers to the case that marginal (over a particular factor) associations are in the opposite direction to every single one of the conditional associations (for each level of the same factor).

	Survives (Y)		
Type (Z)	yes	no	$Y = .$
A	70	17	87
B	37	62	99
$Z = .$	107	79	186

Figure 11.6: Marginal (over X) YZ contingency table.

The marginal YZ contingency table is presented in Figure 11.6, from which we see that

$$\hat{r}_{jk} = \frac{70 \times 62}{37 \times 17} = 6.90... \tag{11.163}$$

providing evidence of a strong association between Type and Survives, i.e., Type A spacejets have a higher chance of surviving the atmosphere than Type B spacejets.

	Type (Z)		
Colour (X)	A	B	$Z = .$
red	23	79	102
green	64	20	84
$X = .$	87	99	186

Figure 11.7: Marginal (over Y) XZ contingency table.

Looking at the Marginal XZ contingency table, presented in Figure 11.7, we see that the marginal sample odds ratio is

$$\hat{r}_{ik} = \frac{23 \times 20}{64 \times 79} = 0.0909... \tag{11.164}$$

suggesting a strong negative association. That is, in this sample, there was a much larger proportion of green Type A spacejets than green Type B spacejets. As a result, when marginalising over Type, green spacejets appeared to have a better survival rate because a bigger proportion of them were Type A.

Q3-7 The Aliens in Durham

Suppose a horde of 216 aliens (beings from a far-off planet) arrive in Durham and are surveyed about the suitability of building structures were they to make Durham their home (they need to report back to the rest of their species). Suppose the rest of their kind is interested in knowing whether the opinion about suitability is associated with alien height (short, average-height, tall) and alien type (A,B). They therefore cross-classify these variables in the contingency table presented in Figure 10.4.

a) Calculate an estimate of, along with a 95% confidence interval for, the marginal odds ratio of Suitability and Type.

Answer: The Marginal YZ contingency table (with totals) is presented in Figure 11.8.

	Type (Z)		
Suitability (Y)	A	B	$Z = .$
Suitable	71	86	157
Unsuitable	40	19	59
$Y = .$	111	105	216

Figure 11.8: Marginal (over X) YZ contingency table.

The (local) sample odds ratio (required only when $j = 1, k = 1$ in this case) is

$$\hat{r}_{jk} = \frac{n_{+,j,k} n_{+,j+1,k+1}}{n_{+,j+1,k} n_{+,j,k+1}} = \frac{71 \times 19}{40 \times 86} = 0.3922... \tag{11.165}$$

which suggests a negative association between Suitability and Type, that is, Type B aliens typically find Durham more suitable than Type A aliens.

The corresponding 95% confidence interval for the log odds ratio is given by:

$$\begin{aligned} & \log \hat{r}_{jk} \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{j,k}} + \frac{1}{n_{j+1,k}} + \frac{1}{n_{j,k+1}} + \frac{1}{n_{j+1,k+1}}} \\ = \; & \log 0.3922... \pm 1.96 \sqrt{\frac{1}{71} + \frac{1}{40} + \frac{1}{86} + \frac{1}{19}} \\ = \; & (-1.566, -0.306) \end{aligned}$$

Therefore, the 95% confidence interval for the odds ratio is given by:

$$(\exp(-1.566), \exp(-0.306)) = (0.209, 0.736) \tag{11.166}$$

b) Calculate a minimal set of marginal (over Type) global odds ratios between Height and Suitability, along with a reasonable 95% confidence interval for each. What assumption about variable Height is being made in order to do this? Interpret and explain the results in terms of associations between Height and Suitability.

Answer: The Marginal XY contingency table (with totals) is presented in Figure 11.9.

	Suitability (Y)		
Height (X)	Suitable	Unsuitable	$Y = .$
Short	24	17	41
Average	94	16	110
Tall	39	26	65
$X = .$	157	59	216

Figure 11.9: Marginal (over Z) XY contingency table.

The formula for the global odds ratios are typically defined relative to cell (i, j) as

$$r_{ij}^G = \frac{\left(\sum_{m\leq i}\sum_{l\leq j}\pi_{ml}\right)\left(\sum_{m>i}\sum_{l>j}\pi_{ml}\right)}{\left(\sum_{m\leq i}\sum_{l>j}\pi_{ml}\right)\left(\sum_{m>i}\sum_{l\leq j}\pi_{ml}\right)} \qquad i = 1, ..., I-1 \quad j = 1, ..., J-1 \qquad (11.167)$$

where it is implicitly assumed that any variables with more than two categories (i.e. $I > 2$ or $J > 2$) are being treated as ordinal.

The sample global odds ratios corresponding to cells $(1, 1)$ and $(2, 1)$ are therefore given by

$$\hat{r}_{11}^G = \frac{24 \times (16 + 26)}{(94 + 39) \times 17} = 0.4458... \qquad \text{and} \qquad \hat{r}_{21}^G = \frac{(24 + 94) \times 26}{39 \times (17 + 16)} = 2.3838... \qquad (11.168)$$

respectively, where we are making the assumption that the variable Height can be treated as ordinal.

Each of these odds ratios corresponds to treating Height as a variable with two categories; either resulting from Average being grouped together with Tall (split at $(1, 1)$) or Short (split at $(2, 1)$). Therefore, following from our knowledge of confidence intervals for local odds ratios, the 95% confidence interval for the global odds ratios corresponding to cells $(1, 1)$ and $(2, 1)$ are given by

$$\log(0.4458) \pm 1.96\sqrt{\frac{1}{24} + \frac{1}{16 + 26} + \frac{1}{94 + 39} + \frac{1}{17}} = (-1.519, -0.096) \qquad (11.169)$$

and

$$\log(2.3838) \pm 1.96\sqrt{\frac{1}{24 + 94} + \frac{1}{26} + \frac{1}{39} + \frac{1}{17 + 16}} = (0.2400, 1.4974) \qquad (11.170)$$

respectively.

- The fact that $\hat{r}_{11}^G$ is less than 1 suggests a negative association, that is, group Short is less likely than group Average/Tall to be satisfied with Durham.

- The fact that $\hat{r}_{21}^G$ is greater than 1 suggests a positive association, that is, group Short/Average is more likely than group Tall to be satisfied with Durham.
- It might be reasonable to treat Height as ordinal, but the association between Height and Suitability is not linear.
- This can be verified by noticing that a greater proportion of the Average height aliens find Durham suitable relative to either Short or Tall aliens.

Chapter 4

Q4-1 Equations from the Notes

a) By rearranging Equation (4.11), show that Equations (4.16) - (4.19) hold for the 2-way saturated LLM using zero-sum constraints.

Answer:

$$\begin{aligned} \log E_{ij} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \\ \implies \sum_{i,j} \log E_{ij} &= \sum_{i,j} (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) \\ &= IJ\lambda + J\sum_i \lambda_i^X + I\sum_j \lambda_j^Y + \sum_{i,j} \lambda_{ij}^{XY} \\ &= IJ\lambda \\ \implies \lambda = \frac{1}{IJ}\sum_{i,j} \log E_{ij} & \end{aligned} \tag{11.171}$$

with the final line holding by the zero-sum constraints.

We then have that

$$\begin{aligned} \sum_i \log E_{ij} &= \sum_i (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) \\ &= I\lambda + \sum_i \lambda_i^X + I\lambda_j^Y + \sum_i \lambda_{ij}^{XY} \\ \implies \lambda_j^Y &= \frac{1}{I}\sum_i \log E_{ij} - \lambda \end{aligned} \tag{11.172}$$

and by symmetry also that

$$\lambda_i^X = \frac{1}{J}\sum_j \log E_{ij} - \lambda \tag{11.173}$$

The final equation is shown by a trivial rearrangement:

$$\lambda_{ij}^{XY} = \log E_{ij} - \lambda - \lambda_i^X - \lambda_j^Y \tag{11.174}$$

b) Assuming Equation (4.26) holds, for the 2-way independence LLM, show that the ML estimates for the λ parameters are as given by Equations (4.27) - (4.29).

Answer: For the independence model, we have that

$$\hat{E}_{ij} = \frac{\hat{E}_{i+}\hat{E}_{+j}}{n_{++}} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{11.175}$$

following Section 2.4.3.2.

Then, we can substitute these estimates into Equations (4.16) - (4.19) to obtain MLEs for the λ parameters.

$$\begin{aligned}
\hat{\lambda} &= \frac{1}{IJ}\sum_{s,t}\log \hat{E}_{st} \\
&= \frac{1}{IJ}\sum_{s,t}\log \frac{n_{s+}n_{+t}}{n_{++}} \\
&= \frac{1}{I}\sum_{s}\log n_{s+} + \frac{1}{J}\sum_{t}\log n_{+t} - \log n_{++} \\
\hat{\lambda}_{i+} &= \frac{1}{J}\sum_{t}\log \hat{E}_{st} - \hat{\lambda} \\
&= \frac{1}{J}\sum_{t}\log \frac{n_{s+}n_{+t}}{n_{++}} - \hat{\lambda} \\
&= \frac{1}{J}(J\log n_{s+} + \sum_{t}\log n_{+t} - \log n_{++}) - \hat{\lambda} \\
&= \log n_{s+} + \frac{1}{J}\sum_{t}\log n_{+t} - \frac{1}{J}\log n_{++} - \Big(\frac{1}{I}\sum_{s}\log n_{s+} + \frac{1}{J}\sum_{t}\log n_{+t} - \log n_{++}\Big) \\
&= \log n_{s+} - \frac{1}{I}\sum_{s}\log n_{s+} \qquad i = 1, ..., I
\end{aligned} \tag{11.176}$$

and by symmetry:

$$\hat{\lambda}_{+j} = \log n_{+t} - \frac{1}{J}\sum_{s}\log n_{+t} \qquad j = 1, ..., J \tag{11.177}$$

Q4-2: Two-way Mutual Independence Model

Assume a LLM with dependency structure $[X, Y]$ under a Poisson sampling scheme with corner point constraints.

a) Write down the appropriate LLM expression corresponding to dependency structure $[X, Y]$.

Answer: We have that

$$\log(E_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y \tag{11.178}$$

b) Rearrange the appropriate LLM expression for the corresponding λ parameters[7].

Answer: In this question we assume corner point constraints with reference categories I and J, so that

$$\lambda_I^X = \lambda_J^Y = 0 \tag{11.179}$$

[7]This is needed for Step 5 of the algorithm outlined in Section 4.4.

We therefore have that

$$\begin{aligned}\log(E_{IJ}) &= \lambda + \lambda_I^X + \lambda_J^Y = \lambda \\ \log(E_{Ij}) &= \lambda + \lambda_I^X + \lambda_j^Y = \lambda + \lambda_j^Y \\ \log(E_{iJ}) &= \lambda + \lambda_i^X + \lambda_J^Y = \lambda + \lambda_i^X\end{aligned}$$

so that

$$\begin{aligned}\lambda &= \log(E_{IJ}) \\ \lambda_i^X &= \log(E_{iJ}) - \lambda = \log(\frac{E_{iJ}}{E_{IJ}}) \\ \lambda_j^Y &= \log(E_{Ij}) - \lambda = \log(\frac{E_{Ij}}{E_{IJ}})\end{aligned}$$

c) Write down the log-likelihood equations, and hence use the method of Lagrange multipliers to show that

$$\begin{aligned}\hat{E}_{++} = E_{++}(\hat{\boldsymbol{\lambda}}) &= n_{++} \\ \hat{E}_{i+} = E_{i+}(\hat{\boldsymbol{\lambda}}) &= n_{i+} \qquad i = 1, ..., I \\ \hat{E}_{+j} = E_{+j}(\hat{\boldsymbol{\lambda}}) &= n_{+j} \qquad j = 1, ..., J\end{aligned}$$

Answer:

The log-likelihood expression is given by[8]

$$\begin{aligned}l(\boldsymbol{\lambda}) &\propto \sum_{i,j}(n_{ij} \log E_{ij} - E_{ij}) \\ &= \sum_{i,j}(n_{ij}(\lambda + \lambda_i^X + \lambda_j^Y) - E_{ij}(\lambda)) \\ &= n_{++}\lambda + \sum_i n_{i+}\lambda_i^X + \sum_j n_{+j}\lambda_j^Y - \sum_{i,j} E_{ij}(\lambda)\end{aligned}$$

where I define

$$E_{ij}(\lambda) = \exp(\lambda + \lambda_i^X + \lambda_j^Y)$$

The Lagrange function is then

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = n_{++}\lambda + \sum_i n_{i+}\lambda_i^X + \sum_j n_{+j}\lambda_j^Y - \sum_{i,j} E_{ij}(\lambda) - \theta^X\lambda_I^X - \theta^Y\lambda_J^Y \qquad (11.180)$$

We now need to calculate the derivative of $\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ with respect to each parameter and set

[8]This is Step 1 of the algorithm outlined in Section 4.4.

equal to 0. We have

$$\begin{aligned}
\frac{\partial \mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})}{\partial \lambda} &= n_{++} - \sum_{i,j} E_{ij}(\hat{\lambda}) = 0 \\
\implies E_{++}(\hat{\boldsymbol{\lambda}}) &= n_{++} \\
\frac{\partial \mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})}{\partial \lambda_i^X} &= n_{i+} - \sum_{j} E_{ij}(\hat{\lambda}) = 0 \qquad i = 1, ..., I-1 \\
\implies E_{i+}(\hat{\boldsymbol{\lambda}}) &= n_{i+} \qquad i = 1, ..., I-1 \\
\frac{\partial \mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})}{\partial \lambda_I^X} &= n_{I+} - \sum_{j} E_{Ij}(\hat{\lambda}) - \hat{\theta}^X = 0 \\
\implies n_{I+} &= E_{I+}(\hat{\boldsymbol{\lambda}}) + \hat{\theta}^X \\
\frac{\partial \mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})}{\partial \lambda_j^Y} &= n_{+j} - \sum_{i} E_{ij}(\hat{\lambda}) = 0 \qquad j = 1, ..., J-1 \\
\implies E_{+j}(\hat{\boldsymbol{\lambda}}) &= n_{+j} \qquad j = 1, ..., J-1 \\
\frac{\partial \mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})}{\partial \lambda_J^Y} &= n_{+J} - \sum_{i} E_{iJ}(\hat{\lambda}) - \hat{\theta}^Y = 0 \\
\implies n_{+J} &= E_{+J}(\hat{\boldsymbol{\lambda}}) + \hat{\theta}^Y \\
\frac{\partial \mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})}{\partial \theta^X} &= -\lambda_I^X = 0 \\
\frac{\partial \mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})}{\partial \theta^Y} &= -\lambda_J^Y = 0
\end{aligned}$$

From the above equations, we have that

$$\begin{aligned}
\sum_{i=1}^{I} E_{i+}(\hat{\boldsymbol{\lambda}}) + \hat{\theta}^X &= \sum_{i=1}^{I} n_{i+} \\
n_{++} &= E_{++}(\hat{\boldsymbol{\lambda}}) + \hat{\theta}^X \\
\implies \hat{\theta}^X &= 0 \\
\sum_{j=1}^{J} E_{+j}(\hat{\boldsymbol{\lambda}}) + \hat{\theta}^Y &= \sum_{j=1}^{J} n_{+j} \\
n_{++} &= E_{++}(\hat{\boldsymbol{\lambda}}) + \hat{\theta}^Y \\
\implies \hat{\theta}^Y &= 0
\end{aligned} \tag{11.181}$$

so that

$$\begin{aligned}
n_{I+} &= E_{I+}(\hat{\boldsymbol{\lambda}}) \\
n_{+J} &= E_{I+}(\hat{\boldsymbol{\lambda}})
\end{aligned} \tag{11.182}$$

as required.[9]

d) Using the results of parts (b) and (c), or otherwise, find the MLEs of the LLM λ coefficients.

[9]These expressions are required for Step 3 of the algorithm outlined in Section 4.4.

Answer: The numbering of steps below correspond to the outline algorithm of Section 4.4.

Step 1: Answer to part (b).

Step 2: Model $[X, Y]$ is defined by

$$\pi_{ij} = \pi_{i+}\pi_{+j} \implies E_{ij} = \frac{E_{i+}E_{+j}}{n_{++}} \tag{11.183}$$

Step 3: Answer to part (c).

Step 4: The MLEs are therefore

$$\hat{E}_{ij} = \frac{\hat{E}_{i+}\hat{E}_{+j}}{n_{++}} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{11.184}$$

by the results of parts (c).

Step 5: Finally, we can substitute these MLEs in to the expressions for $\boldsymbol{\lambda}$ obtained in part (b), to obtain:

$$\begin{aligned}
\hat{\lambda} &= \log(\hat{E}_{IJ}) = \log(\frac{n_{I+}n_{+J}}{n_{++}}) \\
\hat{\lambda}_i^X &= \log(\frac{\hat{E}_{iJ}}{\hat{E}_{IJ}}) = \log(\frac{n_{i+}}{n_{I+}}) \\
\hat{\lambda}_j^Y &= \log(\frac{\hat{E}_{Ij}}{\hat{E}_{IJ}}) = \log(\frac{n_{+j}}{n_{+J}})
\end{aligned}$$

Q4-3: Hierarchical Log Linear Models

a) Identify if the following models are hierarchical, and for the ones that are, use the square bracket [] notation (e.g. $[X, Y]$ for the two-way independence model) presented throughout Sections 3 and 4 to represent them, and discuss what assumptions about the associations between the variables they convey:

$$\begin{aligned}
\text{i. } \log E_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} \qquad \forall i, j, k \\
\text{ii. } \log E_{ijkl} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{il}^{XW} + \lambda_{kl}^{ZW} + \lambda_{ikl}^{XZW} \qquad \forall i, j, k, l \\
\text{iii. } \log E_{ijkl} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ik}^{XZ} + \lambda_{il}^{XW} + \lambda_{kl}^{ZW} + \lambda_{ikl}^{XZW} \qquad \forall i, j, k, l
\end{aligned}$$

Answer:

i. $[X, YZ]$ - X is jointly independent with both Y and Z (but Y and Z (could[10]) interact with each other).

ii. This is not a hierarchical log-linear model due to the presence of a λ_{ikl}^{XZW} term but the absence of a λ_{ik}^{XZ} term.

[10] *could* in the sense that the model permits the modelling of interaction of these two variables with each other, but if the λ_{jk}^{YZ} term was very small, it may indicate a lack of interaction between these variables as well. In contrast, the model directly assumes the joint independence of X with both Y and Z through its structure.

iii. $[Z, XYW]$ - Z is jointly independent with X, Y and W, but these three variables (could) exhibit full (three-way) interaction with each other.

b) Write down the hierarchical log-linear models corresponding to

(i) $[XY, XZ]$

(ii) $[XY, XZ, XW]$

(iii) $[X_1X_3, X_2X_3X_4, X_4X_5, X_5X_6]$

and discuss what assumptions about the associations between the variables they convey.

Answer:

$$\text{i. } \log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \qquad \forall i, j, k \tag{11.185}$$

which models Y and Z as being conditionally independent given X.

$$\text{ii. } \log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{il}^{XW} \qquad \forall i, j, k \tag{11.186}$$

which models Y, Z and W as being conditionally (given X) mutually independent. Note that Y, Z and W may not be marginally mutually independent given X.

$$\begin{aligned} \text{iii. } \log E_{i_1 \ldots i_6} \quad = \quad & \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3} + \lambda_{i_4}^{X_4} + \lambda_{i_5}^{X_5} + \lambda_{i_6}^{X_6} \\ & + \lambda_{i_1 i_3}^{X_1 X_3} + \lambda_{i_2 i_3}^{X_2 X_3} + \lambda_{i_2 i_4}^{X_2 X_4} + \lambda_{i_3 i_4}^{X_3 X_4} + \lambda_{i_4 i_5}^{X_4 X_5} + \lambda_{i_5 i_6}^{X_5 X_6} \\ & + \lambda_{i_2 i_3 i_4}^{X_2 X_3 X_4} \qquad \forall i_1, i_2, i_3, i_4, i_5, i_6, i_7 \end{aligned} \tag{11.187}$$

which models a three-way interaction between variables X_2, X_3 and X_4, with additional two-way interactions between X_1 and X_3, X_4 and X_5, and X_5 and X_6.

c) For the model discussed in part b-iii above, are variables X_1 and X_6 necessarily independent? Explain your answer.

Answer: X_1 and X_6 are not in general (i.e. marginally over X_5) independent. X_1 and X_6 are conditionally independent given X_5. In other words, X_1 and X_6 (may) interact with each other, but this interaction can be captured and explained through the value of X_5.

Q4-4: 1988 General Social Survey

The 1988 General Social Survey compiled by the National Opinion Research Center asked:

- "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms."

The table in Figure 10.2 summarizes opinions about health care costs (H) and the information program (I), classified also by the respondent's gender (G).

Regarding the questions below, any inference based on hypothesis tests should be performed at the 5% level of significance.

a) By using appropriate statistical tools, compare the following associations of the classification variables.

- mutual independence of X, Y and Z, against
- conditional independence of X and Y on Z.

Justify the way you address the problem.

Answer: This is a comparison between nested models, so I'll perform a LRT of the following hypotheses:

$$\mathcal{H}_0 : \mathcal{M}_0 : [X, Y, Z] \tag{11.188}$$
$$\mathcal{H}_1 : \mathcal{M}_1 : [XZ, YZ] \tag{11.189}$$

The likelihood ratio test statistic is

$$G^2(\mathcal{M}_0, \mathcal{M}_1) = 2 \sum_{i,j,k} n_{ijk} \log\Big(\frac{\hat{E}^{(1)}_{ijk}}{\hat{E}^{(0)}_{ijk}}\Big) \sim \chi^2_{df} \tag{11.190}$$

where

$$\hat{E}^{(0)}_{ijk} = \frac{n_{i++} n_{+j+} n_{++k}}{n^2_{+++}} \tag{11.191}$$
$$\hat{E}^{(1)}_{ijk} = \frac{n_{i+k} n_{+jk}}{n_{+++}} \tag{11.192}$$
$$df = (I-1)(K-1) + (I-1)(J-1) = 2 \tag{11.193}$$

since the two models differ only in terms of the parameters λ^{XZ}_{ik} and λ^{XY}_{ij}.

Let

$$G_{ijk} = 2 n_{ijk} \log\Big(\frac{\hat{E}^{(1)}_{ijk}}{\hat{E}^{(0)}_{ijk}}\Big) \tag{11.194}$$

In order to calculate the G_{ijk}, I calculate the quantities in the table given in Figure 11.10.

i	j	k	n_{ijk}	$\hat{E}^{(0)}_{ijk}$	$\hat{E}^{(1)}_{ijk}$	G_{ijk}
1	1	1	76	76.1	72.4	-7.4
2	1	1	6	12.8	9.5	-3.6
1	2	1	160	152.2	163.5	22.95
2	2	1	25	25.7	21.4	-9.15
1	1	2	114	100.8	104.1	7.2
2	1	2	11	17.1	20.8	4.34
1	2	2	181	201.7	190.8	-20.22
2	2	2	48	34.2	38.1	10.52

Figure 11.10: Quantities required for Question 4-3a.

Therefore $G^2_{obs} = \sum_{i,j,k} G_{ijk} = 4.652218$.

The critical value is $\chi^2_{2,0.95} = 5.991$, therefore the test does not reject the null hypothesis model of mutual independence in favour of model $\mathcal{M}_1$ at the 5% level of significance.

b) By using appropriate statistical tools, compare the following associations of the classification variables

- joint independence of X and Z on Y, against
- joint independence of X and Y on Z.

Justify the way you address the problem.

Answer: This is a comparison between non-nested models, so I will use information criteria for comparison purposes. For the purposes of this question, let's compare using both AIC and BIC.

$$\begin{aligned} \text{AIC}(\mathcal{M}) &= -2l_{\mathcal{M}}(\hat{\lambda}) + 2d_{\mathcal{M}} && (11.195) \\ \text{BIC}(\mathcal{M}) &= -2l_{\mathcal{M}}(\hat{\lambda}) + \log(n_{+++})d_{\mathcal{M}} && (11.196) \end{aligned}$$

Joint independence of X and Z on Y is given by model $\mathcal{M}_A : [XZ, Y]$. Assuming Poisson sampling, we have that

$$l_{\mathcal{M}_A}(\hat{\lambda}) = \sum_{ijk}(-\hat{E}^A_{ijk} + n_{ijk}\log\hat{E}^A_{ijk} - \log n_{ijk}!) \propto -n_{+++} + \sum_{ijk} n_{ijk}\log\hat{E}^A_{ijk} \quad (11.197)$$

where

$$\hat{E}^A_{ijk} = \frac{n_{i+k}n_{+j+}}{n_{+++}} \quad (11.198)$$

and $d_{\mathcal{M}_\mathcal{A}} = 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) = 5$.

Joint independence of X and Y on Z is given by model $\mathcal{M}_B : [XY, Z]$. Assuming Poisson sampling, we have that

$$l_{\mathcal{M}_B}(\hat{\lambda}) = \sum_{ijk}(-\hat{E}^B_{ijk} + n_{ijk}\log\hat{E}^B_{ijk} - \log n_{ijk}!) \propto -n_{+++} + \sum_{ijk} n_{ijk}\log\hat{E}^B_{ijk} \quad (11.199)$$

where

$$\hat{E}^B_{ijk} = \frac{n_{ij+}n_{++k}}{n_{+++}} \quad (11.200)$$

and $d_{\mathcal{M}_\mathcal{B}} = 1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) = 5$.

Using the quantities calculated in the tables presented in Figures 11.11 and 11.12, we have that

$$\begin{aligned} AIC(\mathcal{M}_A) &\propto -4584.997 && (11.201) \\ AIC(\mathcal{M}_B) &\propto -4592.519 && (11.202) \\ BIC(\mathcal{M}_A) &\propto -4562.844 && (11.203) \\ BIC(\mathcal{M}_B) &\propto -4570.365 && (11.204) \end{aligned}$$

Both AIC and BIC indicate that model $\mathcal{M}_B : [XY, Z]$ is preferable.

i	j	k	n_{ijk}	$\hat{E}^A_{ijk}$	$n_{ijk} \log \hat{E}^A_{ijk}$
1	1	1	76	78.6	331.7
2	1	1	6	10.3	14.01
1	2	1	160	157.3	809.3
2	2	1	25	20.6	75.71
1	1	2	114	98.3	523.07
2	1	2	11	19.6	32.76
1	2	2	181	196.6	955.95
2	2	2	48	39.3	176.25

Figure 11.11: Quantities required for Question 4-3b.

i	j	k	n_{ijk}	$\hat{E}^B_{ijk}$	$n_{ijk} \log \hat{E}^B_{ijk}$
1	1	1	76	81.6	334.7
2	1	1	6	7.3	11.93
1	2	1	160	146.6	798.04
2	2	1	25	31.3	86.15
1	1	2	114	108.3	534.08
2	1	2	11	9.6	24.98
1	2	2	181	194.3	953.84
2	2	2	48	41.6	178.96

Figure 11.12: Quantities required for Question 4-3b.

Q4-5: $2 \times 2 \times 2$ Contingency Table

Consider a $2 \times 2 \times 2$ Contingency Table with classification variables X, Y and Z.

a) State the equation of the Log-linear model describing the dependency type $[XY, XZ]$.

Answer: This model describes the dependency type Z and Y being conditionally independent on X. It is given by the following LLM:

$$\log E_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \qquad \forall (i,j,k) \in \{1,2\} \tag{11.205}$$

b) Apply the two types of the non-identifiability constraints (corner points and sum-to-zero).

Answer: The corner point constraints, with reference levels $i = 2, j = 2, k = 2$, are

$$\begin{aligned}
\lambda_2^X = \lambda_2^Y = \lambda_2^Z &= 0 && (11.206)\\
\lambda_{2j}^{XY} = \lambda_{i2}^{XY} &= 0 \quad \forall i,j && (11.207)\\
\lambda_{2k}^{XZ} = \lambda_{i2}^{XZ} &= 0 \quad \forall i,k && (11.208)
\end{aligned}$$

The sum-to-zero constraints are given by

$$\begin{aligned}
\lambda_1^X &= -\lambda_2^X && (11.209)\\
\lambda_1^Y &= -\lambda_2^Y && (11.210)\\
\lambda_1^Z &= -\lambda_2^Z && (11.211)\\
\lambda_{i1}^{XY} &= -\lambda_{i2}^{XY} \quad \forall i && (11.212)\\
\lambda_{1j}^{XY} &= -\lambda_{2j}^{XY} \quad \forall j && (11.213)\\
\lambda_{i1}^{XZ} &= -\lambda_{i2}^{XZ} \quad \forall i && (11.214)\\
\lambda_{1k}^{XZ} &= -\lambda_{2k}^{XZ} \quad \forall k && (11.215)\\
& && (11.216)
\end{aligned}$$

c) Write down the number of free parameters, and explain how you calculated them.

Answer: The number of free parameters is given by

$$\begin{aligned}
d &= 1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) + (I-1)(K-1) && (11.217)\\
&= 1 + (2-1) + (2-1) + (2-1) + (2-1)(2-1) + (2-1)(2-1) = 6 && (11.218)
\end{aligned}$$

because

- from λ I have 1 free parameters
- from λ_i^X I have $I = 2$ parameters, but because of the constraints one of them is set to zero or can be specified from the others. Same goes for λ_j^Y and λ_k^Z.
- from λ_{ik}^{XZ} I have $I \times K = 2 \times 2 = 4$ parameters, but because of the constraints $I + K - 1 = 3$ of them are set to zero or can be specified from the others. Same goes for λ_{ij}^{XY}.

For parts (d) and (e), consider the corner point constraints only.

d) Express $\log(\pi_{1|jk}/\pi_{2|jk})$ as a function of the linear model λ coefficients.

Answer: We have that

$$\log(\pi_{1|jk}/\pi_{2|jk}) = \log(E_{1jk}/E_{2jk}) = \log(E_{1jk}) - \log(E_{2jk}) = \begin{cases} \lambda_1^X + \lambda_{11}^{XY} + \lambda_{11}^{XZ}, & j=1, k=1 \\ \lambda_1^X + \lambda_{11}^{XY}, & j=1, k=2 \\ \lambda_1^X + \lambda_{11}^{XZ}, & j=2, k=1 \\ \lambda_1^X, & j=2, k=2 \end{cases} \tag{11.219}$$

e) Express $\log(r_{ij(k)}^{XY})$, that is, the log conditional (on Z) odds ratio, as a function of the linear model λ coefficients. Give a short interpretation of λ_{11}^{XY} based on this.

Answer: We have that

$$\log(r_{ij(k)}^{XY}) = \log(\frac{E_{11k}/E_{21k}}{E_{12k}/E_{22k}}) = \lambda_{11}^{XY} \quad \forall k \tag{11.220}$$

hence λ_{11}^{XY} is equal to the log conditional (on Z) odds ratio of X to Y, for any level of Z.

Q4-6: Exam-Style Question

All possible hierarchical log-linear models were fitted to a contingency table, based on a sample of 312 individuals, for three factors: X (2 levels), Y (3 levels) and Z (5 levels). The sampling scheme that was implemented was the Poisson sampling scheme. The results are shown in the table in Figure 10.5.

a) Calculate the number of free parameters resulting after applying corner point non-identifiability constraints for each model in the table.

Answer: The number of the free parameters resulting after applying corner point non-identifiability constraints for each model is in the 3rd column of the table in Figure 11.13. These are for the Poisson sampling scheme.

Model	Log-likelihood	free parameters	AIC	BIC
$[X, Y, Z]$	-71.75	8	159.49	198.44
$[Z, XY]$	-71.69	10	163.37	200.80
$[Y, XZ]$	-70.37	12	164.73	209.65
$[X, YZ]$	-48.13	16	128.26	188.15
$[XY, XZ]$	-70.30	14	168.6	221.00
$[XY, YZ]$	-47.83	18	131.66	199.03
$[XZ, YZ]$	-39.83	20	119.67	194.53
$[XY, XZ, YZ]$	-39.30	22	122.6	205.00
$[XYZ]$	-38.00	30	136.0	248.30

Figure 11.13: Table for Q4-5.

b) Define the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC), then compute both AIC and BIC for each model in the table.

For model $\mathcal{M}$, the Akaike Information Criterion (AIC), is given by

$$AIC(\mathcal{M}) = -2l_{\mathcal{M}}(\hat{\boldsymbol{\lambda}}) + 2d_{\mathcal{M}} \tag{11.221}$$

where $l_{\mathcal{M}}(\hat{\boldsymbol{\lambda}})$ is the value of the log-likelihood function at the MLE for $\boldsymbol{\lambda}$ for the appropriate set of λ parameters corresponding to model $\mathcal{M}$, and $d_{\mathcal{M}}$ is the number of free parameters in model $\mathcal{M}$. The Bayesian Information Criterion (BIC), is given by

$$BIC(\mathcal{M}) = -2l_{\mathcal{M}}(\hat{\boldsymbol{\lambda}}) + \log(n_{+\ldots+})d_{\mathcal{M}} \tag{11.222}$$

where additionally here we define $n_{+\ldots+}$ to be the total number of observations.

The Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) are calculated and presented in the 4th and 5th columns respectively of the table in Figure 11.13.

c) Identify which model is selected by each criterion and give a short explanation of what each criterion is intended to achieve.

Using criterion AIC would lead to selecting model $[XZ, YZ]$, while using criterion BIC would lead to selecting model $[X, YZ]$.

Both AIC and BIC seek to maximise the likelihood subject to a penalty on the complexity of the model in terms of the number of parameters. If $AIC(\mathcal{M}_1) > AIC(\mathcal{M}_2)$ then model $\mathcal{M}_2$ is preferable to model $\mathcal{M}_1$ according to criterion AIC. If $BIC(\mathcal{M}_1) > BIC(\mathcal{M}_2)$ then model $\mathcal{M}_2$ is preferable to model $\mathcal{M}_1$ according to criterion BIC.

d) Explain precisely what model $[XY, YZ]$ says about the dependence of factor Y on the other two factors.

Model $[XY, YZ]$ means that X and Z are conditionally independent given Y.

Chapter 5

Q5-1: Tokyo Rainfall

We consider rainfall data recorded in Tokyo in the first 11 days of the year 1983, as given by the table in Figure 10.6. For each day, we know whether it rained ($z_i = 1$) or did not rain ($z_i = 0$) on that day.

The goal is to find a model which predicts the probability of rain on day $i + 1$ based on the rainfall on day i, which is given by z_i. Therefore, we define the response $y_i = z_{i+1}$, $i = 1, \ldots, 10$, so that we can write the data in the form given by the table in Figure 10.7

a) Formulate a logistic regression model for $\pi(z)$ (consider the elements stated in Section 5.3.2), using the linear predictor $\eta = \beta_1 + \beta_2 z$, that is, take the predictors to be $x_1 \equiv 1$ and $x_2 = z$.

Answer: The elements of the logistic regression model are as follows:

- Linear predictor: $\eta(z) = \beta_1 + \beta_2 z$, for $z \in \{0, 1\}$.
- Response function $h(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$.

- Distributional assumption: $y|\boldsymbol{\beta}, z \sim \text{Bern}(\pi(z))$, where $\pi(z) = h(\eta(z))$. One could also add that, given z_i (and β), each y_i is independent of the other y_j and z_j, $j \neq i$.

Since z only takes on two values, 0 and 1, there are only two values of $\pi(z)$, which we call $\pi_0 = \pi(0)$ and $\pi_1 = \pi(1)$.

b) Formulate the score function $S(\boldsymbol{\beta})$.

Answer: The general formula for the score function for the logistic model is:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \pi(\boldsymbol{x}_i))\boldsymbol{x}_i \tag{11.223}$$

Here, $\boldsymbol{x}_i = (1, z_i)$, where each of the $z_i \in \{0, 1\}$. So we have

$$S(\boldsymbol{\beta}) = \sum_{i:\, z_i=0} (y_i - \pi_0) \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \sum_{i:\, z_i=1} (y_i - \pi_1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{11.224}$$

c) Solve the estimating equation $S(\hat{\boldsymbol{\beta}}) = 0$.

Answer: Setting $S(\hat{\boldsymbol{\beta}}) = 0$ gives two equations:

$$0 = \sum_{i:\, z_i=0} (y_i - \hat{\pi}_0) + \sum_{i:\, z_i=1} (y_i - \hat{\pi}_1) \tag{11.225}$$

$$0 = \sum_{i:\, z_i=1} (y_i - \hat{\pi}_1) \tag{11.226}$$

Let $n_b = \sum_{i:\, z_i=b} 1$, and $\bar{y}_b = \frac{1}{n_b} \sum_{i:\, z_i=b} y_i$, for $b \in \{0, 1\}$. Then the above equations give:

$$0 = n_0(\bar{y}_0 - \hat{\pi}_0) + n_1(\bar{y}_1 - \hat{\pi}_1) \tag{11.227}$$

$$0 = n_1(\bar{y}_1 - \hat{\pi}_1) \tag{11.228}$$

This gives immediately that

$$\hat{\pi}_b = \bar{y}_b \tag{11.229}$$

for $b \in \{0, 1\}$. Note that this is just the proportion of rainy days following a non-rainy ($b = 0$) or a rainy ($b = 1$) day, as you might expect.

Now we have to solve for $\hat{\boldsymbol{\beta}}$. We have:

$$\hat{\pi}_0 = \frac{e^{\hat{\beta}_1}}{1 + e^{\hat{\beta}_1}} \tag{11.230}$$

$$\hat{\pi}_1 = \frac{e^{\hat{\beta}_1 + \hat{\beta}_2}}{1 + e^{\hat{\beta}_1 + \hat{\beta}_2}} \tag{11.231}$$

This gives:

$$\hat{\beta}_1 = \log \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \tag{11.232}$$

$$\hat{\beta}_1 + \hat{\beta}_2 = \log \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} \tag{11.233}$$

whence it is easy to find $\hat{\beta}_2$.

In this particular case, we have $\bar{y}_0 = \frac{2}{6} = \frac{1}{3}$ and $\bar{y}_1 = \frac{2}{4} = \frac{1}{2}$, which then gives that

$$\hat{\beta}_1 = -0.6931472 \tag{11.234}$$

$$\hat{\beta}_2 = 0.6931472 \tag{11.235}$$

An important point must be made here. We can only solve the equations for the $\hat{\pi}_b$ first because there are as many $\hat{\pi}$ probabilities as there are $\hat{\beta}$ parameters (two). In a more complex model, there will generally be an infinite set of $\hat{\pi}_i$ (or more generally means $\hat{\mu}_i$) satisfying the equations

$$\sum_i (y_i - \hat{\pi}_i)\boldsymbol{x}_i = 0 \tag{11.236}$$

simply because there are p equations and n unknowns. However, only one of this infinite set can be written in the form $h(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$, for which there are only p unknowns, corresponding to the $\hat{\beta}$ parameters.

d) According to BBC Weather, it didn't rain in Tokyo on Thursday November 17, 2022. Use your model to predict the rainfall probability in Tokyo on Friday November 18, 2022.

Answer: On 17/11/2022, we observed the covariate value $z = 0$. Hence the predicted probability of rain on 18/11/2022 is

$$\hat{\pi}_0 = \bar{y}_0 = \frac{e^{\hat{\beta}_1}}{1 + e^{\hat{\beta}_1}} = \frac{1}{3} \tag{11.237}$$

Chapter 6

Q6-1: Properties of the Exponential Dispersion Family

Let the distribution for Y be an exponential dispersion family:

$$P(y|\theta(\mu), \phi) = \exp\Big[\frac{y\theta(\mu) - b(\theta(\mu))}{\phi} + c(y, \phi)\Big] \tag{11.238}$$

where $\mu(\theta) = \mathrm{E}[Y|\theta, \phi]$.

a) Use the fact that $\mu(\theta)$ and $\theta(\mu)$ are inverses to show that $\theta'(\mu) = 1/\mathcal{V}(\mu)$, where $\mathcal{V}(\mu)$ is the variance function.

Answer: The fact that μ and θ are inverses means that

$$\mu = \mu(\theta(\mu)) \tag{11.239}$$

Differentiating both sides with respect to μ gives

$$1 = \mu'(\theta(\mu))\, \theta'(\mu) \tag{11.240}$$

and hence that

$$\theta'(\mu) = \frac{1}{\mu'(\theta(\mu))} \tag{11.241}$$

Since $\mu(\theta) = b'(\theta)$, this gives that

$$\theta'(\mu) = \frac{1}{b''((b')^{-1})(\mu)} = \frac{1}{\mathcal{V}(\mu)} \tag{11.242}$$

b) Show that with one data point $(\boldsymbol{x}, y)$, the maximum likelihood estimator for μ is y.

Answer: The log likelihood is

$$\ell(y|\mu) = \frac{1}{\phi}(y\ \theta(\mu) - b(\theta(\mu))) + c(y, \phi) \tag{11.243}$$

Differentiating with respect to μ, setting the result equal to zero, and multiplying by ϕ, then gives

$$y\ \theta'(\hat{\mu}) - b'(\theta(\hat{\mu}))\ \theta'(\hat{\mu}) = 0 \tag{11.244}$$

and hence, if $\theta'(\hat{\mu}) \neq 0$, that

$$y = b'(\theta(\hat{\mu})) = \hat{\mu} \tag{11.245}$$

Chapter 7

Q7-1: The Gamma Distribution

We consider the exponential dispersion family (EDF)

$$P(y|\theta, \phi) = \exp\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\} \qquad y > 0 \tag{11.246}$$

a) Show that the expectation of a distribution belonging to a EDF is given by $b'(\theta)$.

Answer: We showed this in Sections 6.2 and 6.2.1 of the lecture notes. Refer to these sections for the solution.

b) Show that the Gamma distribution with density

$$P(y|\nu, \alpha) = \frac{\alpha^\nu}{\Gamma(\nu)} y^{\nu-1} e^{-\alpha y} \tag{11.247}$$

(with shape parameter ν and rate parameter α) is a member of the EDF.

Note: The Gamma-function is defined by $\Gamma(\nu) = \int_0^\infty e^{-t} t^{\nu-1}\, dt \quad (\nu > 0)$.

Answer: Taking the logarithm of the gamma density gives:

$$\log P(y|\nu, \alpha) = -\alpha y + \nu \log \alpha - \log \Gamma(\nu) + (\nu - 1) \log y \tag{11.248}$$

$$= \frac{-\frac{\alpha}{\nu} y + \log(\alpha)}{\frac{1}{\nu}} - \log \Gamma(\nu) + (\nu - 1) \log y \tag{11.249}$$

$$= \frac{-\frac{\alpha}{\nu} y + \log(\frac{\alpha}{\nu})}{\frac{1}{\nu}} + \nu \log \nu - \log \Gamma(\nu) + (\nu - 1) \log y \tag{11.250}$$

So we identify:

$$\theta = -\frac{\alpha}{\nu} \tag{11.251}$$

$$\phi = \frac{1}{\nu} \tag{11.252}$$

$$b(\theta) = -\log(-\theta) \tag{11.253}$$

$$c(y, \phi) = -\frac{1}{\phi}\log\phi - \log\Gamma\Big(\frac{1}{\phi}\Big) + \Big(\frac{1}{\phi} - 1\Big)\log y \tag{11.254}$$

c) Exploiting properties of the EDF, find the mean μ and the variance of the Gamma distribution.

Answer: Since $b(\theta) = -\log(-\theta)$, we have that

$$\mu \equiv E[y|\nu, \alpha] = b'(\theta) = -\frac{1}{\theta} = \frac{\nu}{\alpha} \tag{11.255}$$

and

$$\mathrm{Var}[y|\nu, \alpha] = \phi\, b''(\theta) = \frac{1}{\nu}\frac{1}{\theta^2} = \frac{\nu}{\alpha^2} \tag{11.256}$$

d) Reparametrize the density function so that it depends on the parameters ν and μ.

Answer: We have that $\alpha = \frac{\nu}{\mu}$. The density becomes:

$$\rho(y|\nu, \mu) = \frac{\big(\frac{\nu}{\mu}\big)^{\nu}}{\Gamma(\nu)}\, y^{\nu-1}\, e^{-\frac{\nu}{\mu}y} \tag{11.257}$$

e) For gamma-distributed response, identify the natural link for use in a generalized linear model. Why is this link function often unsuitable in practice? Suggest an alternative link function.

Answer: The natural response function is $h = b'$, so here:

$$\mu = h(\eta) = -\frac{1}{\eta} \tag{11.258}$$

This means that the natural link function is

$$\eta = g(\mu) = -\frac{1}{\mu} \tag{11.259}$$

This will be unsuitable in practice because the mean of a non-negative variable can never be negative, whereas $\mu = -\frac{1}{\boldsymbol{\beta}^T\boldsymbol{x}}$ will certainly be negative for some $\boldsymbol{x}$ unless the range of $\boldsymbol{x}$ is restricted. A possible alternative link is the logarithmic link, or in other words the exponential response function: $\mu = e^{\boldsymbol{\beta}^T\boldsymbol{x}}$, which is positive.

Q7-2: Coronary Heart Disease

We are given data collected in the framework of a study of coronary heart disease in the table presented in Figure 10.8. It shows 1329 patients cross-classified by the level of their serum cholesterol (below or above 260) and the presence or absence of heart disease.

We can consider this as a data set which is grouped with respect to a binary covariate, with values (say) $z_1 = 0$ (if < 260) and $z_2 = 1$ (if ≥ 260), and response defined through group-wise heart disease presence rates, that is $y_1 = 51/1043$ and $y_2 = 41/286$. We model this data through a Binomial logit model, that is

$$\pi(z_i) = \frac{\exp(\beta_1 + \beta_2 z_i)}{1 + \exp(\beta_1 + \beta_2 z_i)} \tag{11.260}$$

where $y_i|z_i \sim B(n_i, \pi(z_i))/n_i$ (*rescaled* Binomial distribution).

a) Verify through differentiation of the log-likelihood that the score-function is given by

$$S(\beta_1, \beta_2) = \sum_{i=1}^{2} n_i \begin{pmatrix} 1 \\ z_i \end{pmatrix} \Big(y_i - \frac{\exp(\beta_1 + \beta_2 z_i)}{1 + \exp(\beta_1 + \beta_2 z_i)} \Big) \tag{11.261}$$

Answer: In the stated model, we have $y_i|z_i \sim \text{Bin}(n_i, \pi_i)/n_i$, with $\pi_i \equiv \pi(z_i) = \frac{e^{\beta_1+\beta_2 z_i}}{1+e^{\beta_1+\beta_2 z_i}}$.

Likelihood:

$$L(\beta_1, \beta_2) = \prod_{i=1}^{2} \begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} \pi_i^{n_i y_i} (1 - \pi_i)^{n_i - n_i y_i} \tag{11.262}$$

Log-likelihood:

$$\begin{aligned}
\ell(\beta_1, \beta_2) &= \log L(\beta_1, \beta_2) && (11.263)\\
&= \sum_{i=1}^{2} n_i y_i \log \pi_i + (n_i - n_i y_i) \log(1 - \pi_i) + \log \begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} && (11.264)\\
&= \sum_{i=1}^{2} n_i y_i \log \frac{\pi_i}{1 - \pi_i} + n_i \log(1 - \pi_i) + \log \begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} && (11.265)\\
&= \sum_{i=1}^{2} n_i y_i (\beta_1 + \beta_2 z_i) - n_i \log(1 + e^{\beta_1 + \beta_2 z_i}) + \text{const}(n_i, y_i) && (11.266)
\end{aligned}$$

Score-function:

$$\begin{aligned}
S(\beta_1, \beta_2) &= \begin{pmatrix} \partial/\partial\beta_1 \\ \partial/\partial\beta_2 \end{pmatrix} \ell(\beta_1, \beta_2) && (11.267)\\
&= \sum_{i=1}^{2} n_i y_i \begin{pmatrix} 1 \\ z_i \end{pmatrix} - n_i \frac{1}{1 + e^{\beta_1 + \beta_2 z_i}} e^{\beta_1 + \beta_2 z_i} \begin{pmatrix} 1 \\ z_i \end{pmatrix} && (11.268)\\
&= \sum_{i=1}^{2} n_i \begin{pmatrix} 1 \\ z_i \end{pmatrix} (y_i - \pi_i) && (11.269)
\end{aligned}$$

b) Solve the score equation by hand (i.e. without R).

Serum Cholesterol	i	z_i	Heart Disease present	Heart Disease absent	n_i	y_i
< 260	1	0	51	992	$n_1 = 1043$	$y_1 = 51/1043$
≥ 260	2	1	41	245	$n_2 = 286$	$y_2 = 41/286$
	$n = 2$				$M = 1329$	

Figure 11.14: Table of data cross-classifying cholesterol (treated as a binary covariate) and presence or absence of heart disease.

Answer: We have an expanded version of the table given in Figure 10.8 presented in Figure 11.14.

We then have that:

$$S(\beta_1, \beta_2) = \begin{pmatrix} 1043\left(\frac{51}{1043} - \frac{e^{\beta_1}}{1+e^{\beta_1}}\right) & + & 286\left(\frac{41}{286} - \frac{e^{\beta_1+\beta_2}}{1+e^{\beta_1+\beta_2}}\right) \\ & & 286\left(\frac{41}{286} - \frac{e^{\beta_1+\beta_2}}{1+e^{\beta_1+\beta_2}}\right) \end{pmatrix} \tag{11.270}$$

Equating this to zero and subtracting the second row from the first, one has

$$\begin{aligned} 51 - 1043\frac{e^{\hat{\beta}_1}}{1+e^{\hat{\beta}_1}} &= 0 && (11.271) \\ \hat{\beta}_1 &= \text{logit}\frac{51}{1043} = -2.9679 && (11.272) \end{aligned}$$

Plugging this into the second row, one has

$$\begin{aligned} 41 - 286\frac{e^{\hat{\beta}_1+\hat{\beta}_2}}{1+e^{\hat{\beta}_1+\hat{\beta}_2}} &= 0 && (11.273) \\ \hat{\beta}_1 + \hat{\beta}_2 &= \text{logit}\frac{41}{286} && (11.274) \\ \hat{\beta}_2 &= \text{logit}\frac{41}{286} - \text{logit}\frac{51}{1043} = 1.1802 && (11.275) \end{aligned}$$

So, summarising, we have:

$$\begin{aligned} \hat{\beta}_1 &= -2.9679 \\ \hat{\beta}_2 &= 1.1802 \end{aligned}$$

c) Comment on the relationship between this and the solution to Q5-1.

The score function in this case is the same as the score function in the case of Q5-1, but is arrived at in a different way. In the context of Q5-1, we could have treated z as taking on only two different values (it rained today or did not), and instead of using the explicit binary values of y, we could have used the proportion of *'tomorrows'* on which there was rain for each of the two values of z. This shows the equivalence of the binary or grouped approaches to modelling such data.

Q7-3: Exam-Style Question

Consider the following one-parameter family of probability densities:

$$P(y|a) = \frac{\cos(a)}{e^{-(a-\frac{\pi}{2})y} + e^{-(a+\frac{\pi}{2})y}} \tag{11.276}$$

where $y \in \mathbb{R}$ and $a \in (-\frac{\pi}{2}, \frac{\pi}{2})$.

a) Show that the above family of distributions forms an exponential dispersion family of distributions. Be careful to define all the elements of an exponential dispersion family.

Answer:

$$\begin{aligned}
\log(P(y|a)) &= \log\cos a - \log\Big(e^{-(a-\frac{\pi}{2})y} + e^{-(a+\frac{\pi}{2})y}\Big) && (11.277)\\
&= \log\cos a - \log(e^{-ay}) - \log\Big(e^{\frac{\pi}{2}y} + e^{-\frac{\pi}{2}y}\Big) && (11.278)\\
&= ay - (-\log\cos a) - \log\Big(2\cosh\Big(\frac{\pi}{2}y\Big)\Big) && (11.279)
\end{aligned}$$

so that

- natural parameter $\theta = a$
- log-normaliser $b(\theta) = -\log\cos\theta$
- dispersion parameter $\phi = 1$
- $c(y, \phi) = \log\Big(2\cosh\Big(\frac{\pi}{2}y\Big)\Big)$

b) Derive the mean and the variance as a function of the natural parameter, and express the variance in terms of the mean. What happens as a reaches the limits of its range?

Answer: We have that

$$\mathrm{E}[Y] = b'(\theta) = -\frac{1}{\cos\theta}(-\sin\theta) = \tan\theta \tag{11.280}$$

$$\mathrm{Var}[Y] = b''(\theta) = \sec^2(\theta) \tag{11.281}$$

and thus

$$\mathrm{Var}[Y] = 1 + \mathrm{E}[Y]^2 \tag{11.282}$$

As a reaches the limits of its range, $\mathrm{E}[Y] \to \pm\infty$, while $\mathrm{Var}[Y] \to \infty$.

c) If you were building a GLM using this distribution, what would be the natural link function?

Answer: The natural response function would be

$$\mu = h(\theta) = b'(\theta) = \tan\theta \tag{11.283}$$

and hence the natural link function would thus be

$$g(\mu) = \tan^{-1}(\mu) \tag{11.284}$$

Bibliography

A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley, New York, 3rd edition, 2019.

C. R. Bilder and T. M. Loughin. *Analysis of Categorical Data with R*. CRC press, Boca Raton, 2015.

L. S. Covey, A. H. Glassman, and F. Stetner. Depression and depressive symptoms in smoking cessation. *Comprehensive Psychiatry*, 31(4):350–354, 1990.

R. J. M. Dawson. The unusual episode data revisited. *Journal of Statistical Education*, 3(3), 1995.

R. Doll and R. Peto. Mortality in relation to smoking - 20 years' observations on male british doctos. *British Medical Journal*, 2(6051):1525–1536, 1976.

J. J. Faraway. *Extending the Linear Model with R*. CRC press, London, 2nd edition, 2016.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1937.

S. J. Haberman. The analysis of residuals in cross-classified tables. *Biometrics*, 29:205–220, 1973.

Kateri. *Contingency Table Analysis - Methods and Implementation using R*. Birkhauser, New York, 2014.

N. Mantel. Chi-square tests with one degree of freedom- extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58:690–700, 1963.

G. Tutz. *Regression for Categorical Data*. Cambridge University Press, Cambridge, 2012.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9(1):60–62, 1938.

F. Yates. Contingency tables involving small numbers and the chi square test. *Journal of the Royal Statistical Society Supplement*, 1:217–235, 1934.