SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

Post Graduate Program

TITLE:

Bayesian Model Comparison and Hypothesis Testing for Contingency Tables Using R

Author : Nikos Matsavelas

Supervisor : Dr.Ioannis Ntzoufras

A thesis Submitted to the Department of Statistics of the Athens University of Economics and Business in partial fulfilment of the requirements for the degree of Master of Science in Statistics

Athens,Greece

December,2019

Contents

Abstract

Primary goal of this thesis is to gain the knowledge of probabilistic perspective in statistical analysis and model comparison for a much scientific and practical understanding and handling of statistical data analysis. With this thesis we explore the Bayesian framework in Statistics for model comparison, we implement variable selection, prior distribution specification and hypothesis testing in contingency tables using R and Stan which is an imperative modelling programming language for Bayesian Statistics using the sophisticated gradient-based MCMC method. In the first chapter we introduce to the reader the Bayesian methodology and the process of Bayes Factor calculation given the historically research that has been done and we present examples produced using Stan in R. In the second chapter we present the probability structure in contingency tables, the distributional sampling (design) of them and how we implement independence test and some of the most important statistical measurements like odds ratio, risk ratio and difference in proportions using conjugate priors.In the third chapter we dive in the Generalized Linear Models world, through the conjugate prior analysis of logistic regression, log-linear models, multinomial and ordinal models.In chapter 4, we introduce to the reader the idea of penalised likelihood for model comparison. AIC,BIC and DIC are the main criteria of interest along with the Leave-One-Out cross validation criterion, that has been created as a model comparison process, examining the expected predictive accuracy of the models produced in Stan. In addition in chapter 5 we examining the dependent observations for Matched Pair models. Symmetry, marginal homogeneity,Kappa coefficient, McNemar test are subjects that are concerning scientists in many fields of science and in this chapter we present the probabilistic approach of them with Bayesian analysis.

Dedication

To my mum and my sister

Acknowledgements

I want to thank my professor Dr. Ioannis Ntzoufras (supervisor of this thesis), for his precious and substantial scientific aid and guidance through out this thesis.

Part I

Introduction to Bayesian Model Comparison

Chapter 1

Introduction to Bayesian Model Comparison

This chapter is an introduction to Bayesian framework, how we compute the posterior distribution, the prior distribution selection, the Bayes Factor as a tool of model comparison, hypothesis testing, the different approaches for computing the marginal likelihoods and at the end we will provide to the reader an example of this methodology.

1.1 Introduction to Bayes Theorem

Considering that we have a parametric model: $p(y \mid \theta)$ for the data y given the parameters θ. We propose this model given the structure of the data, information about how the data was collected and knowledge about the context from which it arise. But there is an uncertainty about the parameters θ which we hope the data will reduce but usually there is also uncertainty about the model itself.

Instead of regarding the probability distribution $p(y \mid \theta)$ as a function of y , we might view it as a function of θ. This **Likelihood** $L(\theta)$ contains the uncertainty about θ. With this frequentist approach we aim to estimation of the parameters θ by computing their most likely values with maximizing

$L(\theta)$.

In the frequentist approach the parameters θ are fixed and unknown. In the Bayesian point of view, the parameters are random variables. The big difference between the frequentist and the bayesian approach is that in the latter before we observe the data, our uncertainty about the parameters is expressed through a prior density $\pi(\theta)$. So there is a update about the uncertainty regarding the parameters using the data and the model to calculate a posterior density $p(\theta \mid y)$.

The equation of **Bayes Theorem** is :

$$p(\theta \mid y) = \frac{p(y \mid \theta) \times \pi(\theta)}{p(y)}$$

In words this equation can be written as :

$$Posterior \propto Likelihood \times Prior$$

which makes it very obvious and simple how the posterior is derived from combining the likelihood and the prior. Problems usually appear in the computation of the denominator of Bayes Theorem which contains the normalising constant $p(y)$ which equals to:

$$p(y) = \int_{\theta} p(y \mid \theta)\pi(\theta)d\theta$$

This is called the marginal likelihood of the data under the model. If the number of parameters is large then we have to compute multiple integrals which makes the problem even harder.

Computing integrals such as these is the reason why the simplicity of Bayes Theorem requires rather more work than would first appear.

This was the reason why this probabilistic approach was abandoned for many years until the exponentially growth of computer based calculations gave birth to the MCMC algorithm implementation.

1.2 Prior Specification

The use of prior distribution for calculating the posterior distribution is the strongest and simultaneously the weakest part of Bayesian methodology which the frequentist approach doubts about the proper selection of prior among other priors.

There are many approaches of which is the proper selection of prior distribution. Some of them are:

1. **Informative Priors**. These kind of priors incorporate the information about the parameters θ according past experience and previous surveys. Sometimes these kind of prior distributions are posterior distributions derived from past experiments.

2. **Conjugate Priors**. A more convenient choice of prior distributions are so called conjugate priors which are distributions that they belong in the same family of distributions with the likelihood of the data. Those types of priors will be our main concern on this thesis.

   (a) **Example 1.1 : Binomial Distribution**

   $$X|\theta \sim Binomial(N,\, \theta)$$

   $$f(x|\theta) = \binom{N}{\theta}\theta^x(1-\theta)^{N-x}$$

   with N known and for the parameter θ we assume that is following the Beta Distribution with parameters α and β:

   $$\pi_\theta \sim Beta(\alpha, \beta)$$

   $$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

   The part of the type that contains the value θ of interest is called the kernel of the distribution. The part that does not contain the parameter θ is called the normalising constant and can be excluded from the calculation of the posterior density for posterior inference.

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)\pi(\theta) \\ &\propto \left[\theta^X(1-\theta)^{N-X}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \\ &= \theta^{\alpha+X-1}(1-\theta)^{\beta+N-X-1} \end{aligned}$$

Concluding to the posterior density:

$$p(\theta|x) \sim Beta(\alpha + X, \beta + N - X)$$

(b) **Example 1.2 : Poisson Distribution**

Here the appropriate conjugate prior selection when the data are Poisson distributed is the Gamma Distribution . Let us assume:

$$X|\theta \sim Poisson(\lambda)$$

The Poisson Likelihood is given by:

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$$

Thus the gamma distribution likelihood is given by:

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma_a \beta^\alpha}$$

We calculate again the posterior density $p(\theta|x)$

$$p(\theta|x) \propto f(x|\theta)\pi(\theta)$$

$$\propto \left[e^{-\theta}\theta^x\right]\left[\theta^{\alpha-1}e^{-\theta/\beta}\right]$$

$$= \theta^{X+\alpha-1}e^{-\theta(1+1/\beta)}$$

Concluding to a gamma the posterior $p(\theta|x) \sim Gamma(\alpha+X, \beta+N)$.

(c) **Example 1.3 : Normal Distribution (with known variance)**

When the data y are following $y \sim Normal(\mu, \sigma)$ distribution then a prior $\sim Normal(\mu, \sigma)$ is a very reasonable selection. E.g for X observations we have $X|\mu \sim Normal(\mu, \sigma^2)$ and σ^2 is known. Suppose we have an unknown parameter μ for which the prior beliefs can be express in terms of a normal distribution, so that:

$$\mu \sim Normal(\mu_0, \sigma_0^2)$$

Prior : $f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(\mu-\mu_0)^2/2\sigma_0^2}$

Likelihood : $f(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_0)^2/2\sigma^2}$

$$\begin{aligned}
f(\mu \mid x) &= \frac{f(\mu) f(x \mid \mu)}{\int_{-\infty}^{\infty} f(\mu) f(x \mid \mu) d\mu} \\
&= \frac{f(\mu) f(x|\mu)}{f(x)} \propto f(\mu) f(x|\mu) \\
&= \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left(\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}\right) \times \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{2\pi\sqrt{\sigma_0^2\sigma^2}} exp\left(\frac{-\mu^2 + \mu\mu_0 - \mu_0^2}{2\sigma_0^2} - \frac{x^2 - 2\mu x + \mu^2}{2\sigma^2}\right) \\
&\propto exp\left(\frac{-\mu^2\sigma^2 + 2\mu\mu_0\sigma^2 - \mu_0^2\sigma^2 - \sigma_)^2 x^2 + 2\mu\sigma_0^2 x - \mu^2\sigma_0^2}{2\sigma_0\sigma^2}\right) \\
&\propto exp\left(\frac{-\mu^2(\sigma^2 + \sigma_0^2) + 2\mu(\mu_0\sigma^2 + \sigma_0^2 x) - (\mu_0^2\sigma^2 + \sigma_0^2 x^2)}{2\sigma_0\sigma^2}\right) \\
&\propto exp\left[\frac{-\mu^2 + 2\mu(\frac{\mu_0\sigma^2 + \sigma_0^2 x}{\sigma^2 + \sigma_0^2}) - (\frac{\mu\sigma^2 + \sigma_0^2 x}{\sigma^2 + \sigma_0^2})^2}{\frac{2\sigma_0^2\sigma^2}{\sigma^2 + \sigma_0^2}}\right] \times exp\left(-\frac{\mu_0\sigma^2 + \sigma_0 x^2}{2\sigma_0\sigma^2}\right) \\
&\propto exp\left[\frac{-(\mu - \frac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma^2 + \sigma_0^2})^2}{2 \times \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}}\right]
\end{aligned}$$

So the posterior inference becomes:

$$f(\mu|x) \sim Normal\left(\frac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma^2 + \sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}\right)$$

3. **Jeffreys Priors**.The Jeffreys prior is:

$$p(\theta) \propto \left[I(\theta)^{1/2}\right],$$

where

$$\left[I(\theta)^{1/2}\right] = E_\theta\left[(\frac{\partial}{\partial\theta} log f(X|\theta))^2\right] = -E_{x|\theta}\left[(\frac{\partial^2}{\partial\theta^2} log f(X|\theta))\right]$$

This prior is invariant to reparameterisation. Unfortunately, the Jeffreys prior produces outcomes under some models so it cannot be regarded as a universal solution to the prior choice problem.

4. **Non-Informative Priors**. Priors that does not favor one value of θ over another. All values of discrete θ or all intervals of equal length of the continuous θ,will have the same probability a-priori. An example of that kind of priors is (from Panagiotis Tsiamyrtzis lecture notes): For any $\theta \in \Theta = 1, 2, \ldots, K$ we have :

$$p(\theta = i) = \frac{1}{K}$$

$$\forall i = 1, 2, \ldots, K$$

1.3 Bayes Factor

Starting with data y, assuming to have arisen under one of the two hypothesis H_0 and H_1 according to a probability density $p(y|H_0)$ and $p(y|H_1)$ we can assign to them a priori probabilities $p(H_0)$ and $p(H_1) = 1-p(H_0)$. Multiplied with the likelihood of the data y we get the posterior probabilities $p(H_0|y)$ and $p(H_1|y) = 1 - p(H_0|y)$.

Because any prior view gets transformed to a posterior view given the data from the likelihood, the transformation itself represents the evidence provided by the data. If we convert to the odds scale

$$odds = \frac{P}{(1-P)},$$

the transformation takes a simple form.

From Bayes Theorem we take :

$$p(H_k|y) = \frac{p(y|H_k)p(H_k)}{p(y|H_0)p(H_0) + p(y|H_1)p(H_1)}.$$

Becomes:

$$\frac{p(H_0|y)}{p(H_1|y)} = \frac{p(y|H_0)}{p(y|H_1)} \times \frac{p(H_0)}{p(H_1)}.$$

So the transformation gives us the Bayes Factor:

$$B_{01} = \frac{p(y|H_0)}{p(y|H_1)}.$$

In other words the Bayes Factor can be written or expressed as :

$$PosteriorOdds = BayesFactor \times PriorOdds$$

Note that the null Bayesian evidence of the hypothesis is placed in the numerator and the corresponding measure of the alternative H_1 in the denominator. These placement is interpreted as evidence in support of the null hypothesis H_0. If the null hypothesis was placed in the denominator the interpretation would be in reverse, meaning that there is evidence against the H_0.

The frequentist approach of the null hypothesis testing is based on the rule p.value $> a = 0.05$ for not rejecting the H_0 and p.value $< a = 0.05$ for rejecting the H_0. In the Bayesian framework this does not exist because the Bayes Factor is a summary of the evidence provided by the data in favour of one scientific theory, presented by a statistical model, as opposed to another.

Jeffreys (1961) suggested interpreting B_{10} in half units on the $\log_{10}$ scale. Kass and Raftery (1995) proposed t0 consider twice the natural logarithm of the Bayes Factor, which is on the same scale as the familiar deviance and likelihood ratio test statistic G^2. Those two scales are presented in the table 1.

Table 1.1: Jeffreys Scale of Bayes Factor

$\log_{10}$	B_{10}	Evidence against Ho
0 to 1/2	1 to 3.2	Not worth mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Table 1.2: Kass & Raftery Scale of Bayes Factor

$2 \log$	B_{10}	Evidence against Ho
0 to 2	1 to 3	Not worth mention
2 to 6	3 to 20	Substantial
6 to 10	20 to 150	Strong
>10	>150	Very Strong

The Bayes Factor, B_{01}, apart from testing the null hypothesis against the alternative, has also application in model comparison. Thus for comparing model M_0 against model M_1 for observed data, y, the Bayes Factor becomes the ratio of the posterior odds for M_0 against M_1 of the prior odds:

$$B_{01} = \frac{f(y|M_0)}{f(y|M_1)}$$

From the above we can conclude that the Bayes Factor is the ratio of the marginal likelihoods under the two models being compared.So the whole idea nails down to the calculation of the following integral:

$$f(y|M_m) = \int f(y|\theta_m, M_m) f(\theta_m|M_m)\, d\theta_m, m = 0, 1$$

where θ_m is the vector of parameters in model M_m and $f(\theta_m|M_m)$ is its prior density.Dropping the notational dependence on the model,this can be written as:

$$f(y) = \int f(y|\theta) f(\theta)\, d\theta.$$

1.4 Computing the marginal likelihood

Bayes Factors are the ratio of the marginal likelihoods. Another expression of the marginal likelihood is the integration likelihood. Historically the integration required for calculating marginal likelihoods has been done by taking advantage of conjugacy or by assuming approximate posterior normality.

In other cases the requisite integrals have been approximated using such methods as Gaussian quadrature, the Laplace approximation or Monte Carlo methods. With the availability of increasing computer power, Markov chain Monte Carlo (MCMC) has become a reasonable alternative.

All begun back in the year 1953. A Greek physicist Nickolas Metropolis (1953) published and introduced the very famous method of Markov Chain Monte Carlo which followed by a mathematician from Canada Dr.Hastings (1970) who finished his work and has open up, until now, the road for calculating these computationally complicated integrals through their method.

Until then a lot of statisticians, mathematicians and computer developers have proposed their opinion and view for overcoming this obstacle. In this subsection, of this introduction chapter, we will focus on two main methods that have been published in 1995 through scientific papers.

The first is the paper of Siddhartha Chid (1995) that proposed the calculation of marginal likelihood from Gibbs sampler output. The other paper is from Kass & Raftery (1995) who they suggested that the Laplace-Metropolis method is the most proper approach of calculating the marginal likelihood.

The approach of Chid (in the general case) is that we have a model with $\theta_r, \ldots, \theta_s$ with $s \neq r$ parameters. From Gibbs Sampler we take $\theta^\star$ simulated parameters. For using the Gibbs Sampler we must first calculate the full conditional densities of each parameter given all the remaining parameters and so B are the number of vectors of complete conditional densities. From the conditional Gibbs output the marginal likelihood can be calculated from:

$$\hat{\pi}(\theta_r^\star | y, \theta_s^\star(s < r)) = G^{-1} \sum_{j=1}^{G} \pi(\theta_r^\star | y, \theta_1^\star, \ldots, \theta_{r-1}^\star, \theta_l^{(j)}(l > r), z^{(j)}),$$

whereas an estimate of the joint density is:

$$\prod_{r=1}^{B} \hat{\pi}(\theta_r^{\star}|y, \theta_s^{\star}(s < r)).$$

The log of the marginal likelihood is :

$$\log(\hat{m}(y)) = log(f(y|\theta^{\star})) + log(\pi(\theta^{\star})) - \sum_{r=1}^{B} log(\hat{\pi}(\theta^{\star}|y, \theta_s^{\star}(s < r))).$$

Raftery (1995) originally proposed the Laplace method estimator. This method is an approximation of the marginal density of the data given from the integral:

$$I = \int p(y|\theta, H)\pi(\theta|H)\ d\theta$$

Assuming that the posterior density, which is proportional to $p(y|\theta, H)\pi(\theta|H)$, is highly peaked about its maximum $\hat{\theta}$, which is the posterior mode.

This will usually be the case if the likelihood function $p(y|\theta, H)$ is highly peaked near its maximum $\hat{\theta}$, which will be the case for large samples. We let:

$$I(\theta) = log(p(y|\theta, H)\pi(\theta|H))$$

Expanding $\hat{I}(\theta)$ as a quadratic about $\hat{\theta}$ and then exponentiating yields an approximation to $p(y|\theta, H)\pi(\theta|H)$ that has the form of a normal density with with mean $\hat{\theta}$ and covariance matrix $\hat{\Sigma} = (-y^2\hat{I}(\theta))^{-1}$,where $-y^2\hat{I}(\theta)$ is a Hessian matrix of second derivatives. Integrating this approximation yields:

$$I = (2\pi)^{d/2}|\hat{\Sigma}|^{1/2}p(y|\hat{\theta}, H)\pi(\hat{\theta}|H),$$

where d is the dimensions (or the number) of parameters. This is the Laplace's method of approximation. For calculation issues is more convenient to take log scale of Laplace's Method:

$$\log(\hat{I}) \approx (d/2)log(2\pi) + log(p(y|\hat{\theta}, H)\pi(\hat{\theta}|H) + (1/2)log|-\hat{\Sigma}|$$

For many problems in which the sample size n is moderate, it produces answers well within the accuracy required for drawing conclusions according

to the tables of Bayes Factor scaled by Jeffreys which have been represented in the previous chapter. Formally as $n \to \infty$, $I = \hat{I}(1+O(n^{-1}))$. The relative error is $O(n^{-1})$.

The weakness of Laplace method is that it's not accurate when n grows large. Raftery suggested what he called "Laplace -Metropolis" estimator of $p(y)$,obtained by using the posterior simulation output to estimate the quantities needed to compute the Laplace approximation as described before.,namely the posterior mode $\hat{\theta}$, and minus the inverse Hessian at he posterior mode $\hat{\Sigma}$.

The posterior mode can be estimated as the θ^i simulated that maximises the unormalized $p(y|\theta^i\pi(\theta^i)$. This requires computing the likelihood for each simulated θ^i.If this takes too much computer time, then an alternative is to use the multivariate or component-wise posterior median or to estimate the posterior mode by nonparametric density estimation. The matrix $\hat{\Sigma}$ can be estimated by the posterior covariance matrix.

Within the example chapter we will use the Laplace method which calculates the normalising constant of the marginal likelihood and we will compare different models and one way hypothesis testing using this method.

Many packages in R now are using both Chid95 and Laplace method for calculating the marginal posterior. Due to the small and moderate size datasets we find attractive to use Laplace Method in R.

Another method of calculating the marginal likelihood is via the Bridge-sampling method which was proposed by Meng and Wong (1996). Bridge sampling can be thought of as a generalisation of simpler methods for estimating normalising constants such as the naive Monte Carlo estimator, the generalized harmonic mean estimator, and importance sampling.

These simpler methods typically use samples from a single distribution, whereas bridge sampling combines samples from two distributions. For instance, in its original formulation (Meng and Wong 1996), bridge sampling was used to estimate a ratio of two normalising constants such as the Bayes factor. In this scenario, the two distributions for the bridge sampler are the posteriors for each of the two models involved.

However, the accuracy of the estimator depends crucially on the overlap

between the two involved distributions; consequently, the accuracy can be increased by estimating a single normalising constant at a time, using as a second distribution a convenient normalised proposal distribution that closely matches the distribution of interest. The bridge sampling estimator of the marginal likelihood is then given by:

$$\begin{aligned} p(y) &= \frac{E_{g(\theta)}\left[h(\theta)p(y \mid \theta)p(\theta)\right]}{E_{p(\theta|y)}\left[h(\theta)g(\theta)\right]} \\ &\approx \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} h(\hat{\theta}_j)p(y \mid \hat{\theta}_j)p(\hat{\theta}_j)}{\frac{1}{n_2}\sum_{i=1}^{n_1} h(\theta_i^*)g(\theta_i^*)}, \end{aligned}$$

where $h(\theta)$ is called the bridge function and $g(\theta)$ denotes the proposal distribution $\{\theta_1^\star, \theta_2^\star, \ldots, \theta_{n_1}^\star\}$ denote n_1 samples from the posterior distribution $p(\theta \mid y)$ and $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{n_2},\}$ denote n_2 samples from the proposal distribution $g(\theta)$.

To use bridge sampling in practice, one has to specify the bridge function $h(\theta)$ and the proposal distribution $g(\theta)$. For the bridge function $h(\theta)$, the bridge-sampling method minimises the relative mean-squared error of the estimator.

Using this particular bridge function, the bridge sampling estimate of the marginal likelihood is obtained via an iterative scheme that updates an initial guess of the marginal likelihood $\hat{p}(y)$ until convergence. The estimate at iteration t + 1 is obtained as follows:

$$\hat{p}(y)^{(t+1)} = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} \frac{l_{2,j}}{s_1 l_{2,j} + s_2 \hat{p}(y)^{(t)}}}{\frac{1}{n_1}\sum_{i=1}^{n_2} \frac{1}{s_1 l_{1,i} + s_2 \hat{p}(y)^{t}}},$$

where $l_{1,i} = \frac{p(y|\theta_i^\star)p(\theta_i^\star)}{g(\theta_i^\star)}$ and $l_{2,j} = \frac{p(y|\theta_j^\star)p(\theta_j^\star)}{g(\theta_j^\star)}$.

After having specified the bridge function, one needs to choose the proposal distribution $g(\theta)$. The bridge-sampling method implements two different choices: (a) a multivariate normal proposal distribution with mean vector and covariance matrix that match the respective posterior samples quantities and (b) a standard multivariate normal distribution combined with a warped posterior distribution.

Both choices increase the efficiency of the estimator by making the proposal and the posterior distribution as similar as possible. Note that under the optimal bridge function, the bridge sampling estimator is robust to the relative tail behaviour of the posterior and the proposal distribution. This stands in sharp contrast to the importance and the generalized harmonic mean estimator for which unwanted tail behaviour produces estimators with very large or even infinite variances.

1.5 Importance Sampling

Let us return to the basic problem of computing an integral in Bayesian inference. In many situation, the normalizing constant of the posterior density $p(\theta|y)$ will be unknown. So the posterior mean of the function $h(\theta)$ will be be given by the ratio of integrals:

$$E(h(\theta)|y) = \frac{\int h(\theta)\pi(\theta)f(y|\theta)\, d\theta}{\int \pi(\theta)f(y|\theta)|\, d\theta},$$

where $\pi(\theta)$ is the prior and $f(y|\theta)$ is the likelihood function.

If we were able to simulate a sample $\{\theta^j\}$ directly from the posterior density π , then we could approximate this expectation by a Monte Carlo estimate. In the case where we are not able to generate a sample directly p that we can simulate and that approximates the posterior density $p(\theta|y)$.

We estimate the posterior mean as:

$$\begin{aligned} E(h(\theta)|y) &= \frac{\int h(\theta)\frac{p(\theta)f(y|\theta)}{\pi(\theta)}\, d\theta}{\int \frac{p(\theta)f(y|\theta)}{\pi(\theta)}|\, d\theta} \\ &= \frac{\int h(\theta)w(\theta)p(\theta)\, d\theta}{\int w(\theta)p(\theta)|\, d\theta}, \end{aligned}$$

where $w(\theta) = p(\theta)f(y|\theta)/\pi(\theta)$ is the weight function. If $\theta^1, \theta^2, \ldots, \theta^m$ are a simulated sample from approximation density π, the importance sampling estimate of the posterior mean is:

$$\hat{h}_{IS} = \frac{\sum_{j=1}^m h(\theta^j)w(\theta^j)}{\sum_{j=1}^m w(\theta^j)}$$

.

This is an importance sampling estimate because we are sampling values of θ that are important in computing the integrals in the numerator and denominator. The simulation standard error of an importance sampling estimate is estimated by :

$$se_{\hat{h}_{is}} = \frac{\sqrt{\sum_{j=1}^{m}\left[(h(\theta^j) - \hat{h}_{IS}w(\theta^j)\right]^2}}{\sum_{j=2}^{m} w(\theta^j)}$$

1.6 Sensitivity analysis

Sensitivity analysis concerns distributional forms for models $p(y|\theta_k, H_k)$ as well as priors. When alternatives are introduced (e.g Student's t distribution in place of the normal), Bayes Factors may be used to determine which best fits the data.

One may also assess the influence of individual data values by computing the Bayes Factor after omitting each observation in turn. Asymptotic approximation makes the "leave-one-out" procedure diagnostic approach easy that we will present in chapter 4 .

Because Bayes Factor is sensitive to the prior it is important to evaluate the Bayes Factor over a range of possibilities. This involves specifying classes of priors to use under H_0 and H_1, and it also makes the issue of computation more urgent, because many multidimensional integrals must be calculated.

To be more accurate in this specific subject of Bayesian Analysis,it is important to assess the sensitivity of any inferences with respect to changes in the model assumptions, including assumptions about the sampling density $p(y|\theta)$ and the prior density $\pi(\theta)$.

To express it generally: sensitivity is the exploration of our posterior inferences with respect to the choice of parameters in the prior distribution. Someone can tell that this is a tool for investigating the model uncertainty.

Notice that importance sampling may be used for sensitivity analysis even if the original posterior sample $\theta^1, \theta^2, \ldots, \theta^m$ was obtained using some other

method (e.g the Gibbs Sampler). The new posterior estimates are not likely to be as accurate as the original, but are probably sufficient for the purpose of a sensitivity analysis.

1.7 Example of Bayes Factor For Model Comparison (Euroleague)

In the first example we will demonstrate a hypothesis test for Kostas Sloukas average points in Euroleague Regular Season 2018-19.The real data taken from the `https://www.euroleague.net/competition/players/showplayer?pcode=001926&seasoncode=E2018#!E2018_RS` .

Here we assume the measurements are normally distributed with mean μ and standard deviation σ. Kostas Sloukas scored the following points in each 27 games $y_i = 8, 13, 20, \ldots, 11, 8, 22$ (see Appendix)

We are ready to test two different Bayesian models using the Bayes Factor for prior selection. Suppose we have prior knowledge about his true point average coming from the data from 2009-8 season up until the previous season 2017-18 and we assign the $\mu = 9$.

The Bayes Factor for support of M_0 against the M_1 as shown previously is:

$$B_{01} = \frac{M_0}{M_1}$$

,where M_0 and M_1 are the two models that have been calculated by the gradient-based MCMC in Stan.

So μ for M_0 is distributed as

$$\mu \sim Normal(9, 5)$$

and

$$\sigma \sim Cauchy(0, 1)$$

(Note here that Stan is using standard deviations and not variances). The log marginal likelihood of this H_0 is: -61.56 as reported by Stan.

For the second model in the denominator we gave prior density on the mean:

$$\mu \sim Normal(11, 1)$$

and prior on standard deviation :

$$\sigma \sim Cauchy(0, 1)$$

and we obtained log marginal likelihood -59.60875.

Continuing our process of Bayes Factor calculation we load the BridgeSampling package in R that works with Stanfit models and calculates the Bayes factor as Kass and Raftery (1995) presented.

Putting in the numerator the marginal likelihood of the first model and the denominator the log marginal likelihood of the second model the Bayes Factor becomes :

$$BF_{01} = \frac{M_0}{M_1} = 0.14133$$

where there is evidence against M_0.

We proceed by reversing the hypotheses in the Bayes Factor :

$$BF_{10} = \frac{M_1}{M_0} = 7.07547$$

where we have substantial evidence against the H_0. So we can conclude that Kostas Sloukas is scoring on average more than 10 points on each Euroleague game.

1.8 Example of Bayes Factor For Model Comparison (PremierLeague)

To explain in practice the usage of Bayes Factor, the website `https://www.premierleague.com/clubs/10/Liverpool/results?co=1&se=210` gives the number of Liverpool goals for each game in Premier League for the season 2018-19 (see Appendix).

Suppose we observe **Liverpool's** goals $y_1, y_2, \ldots, y_N$ for N consecutive games. The general model for these data are: $y_1, \ldots, y_N$ are independent $f(y|\theta)$. Since goals are relatively rare events, it is reasonable to assume that the $y_i s$ are Poisson distributed with mean λ. The values of y are :

$$y_i = 4, 2, 1, \ldots, 5, 3, 2$$

Now let's suppose we have some priors beliefs about the mean Liverpool's goals $E(y)$. Let's say that the average mean of goals are 2 according to the last two previous Premier League seasons and standard deviation 1. Also we should think about different choices for the prior density.

For the prior selection, there are many possible choices. In this example we will compare 3 priors on the Liverpool's likelihood mean and standard deviation from different distributions. The reason for that is that we want to give significant knowledge of prior and some weakly informative priors.

After some careful thinking about the prior choices and sampling density, we will compare those models with Bayes Factors.To do this, we compute the prior predictive density of the real data for each model. Then we will compute the log marginal likelihoods of the models with Stan modelling language in R (see Appendix).

Continuing our example , we will use a conjugate prior to express our belief about the mean Liverpool's goals θ which is a $Gamma(4.5, 2)$.

Our data constist of: $Y = 89$ the total goal counts, $N = 38$ the total games played. Expressing our prior belief in $Gamma$ for the average goals θ, then from the expected mean of $Gamma$

$$E(y) = \frac{a}{b} = 2.25$$

and variance

$$V(y) = \frac{a}{b^2} = 1.125$$

solving this system of equations with two unknowns we take: $\alpha = 4.5, \beta = 2$.

Taking the kernel of $Gamma$ and $Poisson$ $\theta^{X+\alpha-1}e^{-\theta(1+1/\beta)}$, then the joint posterior distribution becomes

$$\theta|y \sim Gamma(\alpha + Y, \beta + N)$$

and we obtain:

$$Gamma(93.5, 40).$$

But without the normalising constant computed in the denominator. This will be conducted with the calculation of marginal likelihood with the Laplace method.

But this conjugate prior selection with $Gamma$ prior (model M_1) is probably unsure how well will fit the data and for this reason we would like to compare it with an other or other models. So we can fit a second sampling model (model M_2) which the prior will be normal in log scale with log λ having mean 1 and standard deviation 0.5.

A third model (model M_3) will be the same for log λ with mean 2 and standard deviation 0.5 and finally a fourth model (model M_4) again log λ with mean 1 and standard deviation 2.

Now we display the posterior modes, posterior standard deviations and log marginal densities for the four models corresponding to the four models.

Table 1.3: Marginal Likelihood results

Summary		
Posterior Mode	Posterior SD	log Marginal
0.84	0.10	1.06
0.85	0.10	-1.14
0.89	0.10	-3.65
0.85	0.10	-2.46

Assessing the results and plugging them in the Bayes Factor.Now for comparing in support of the model M_2 over each else after we exponentiate, we take the following results:

$$BF_{21} = \frac{M_2}{M_1} = exp(-1.145 + 1.065) = 0.976.$$

First for the comparison the support of model M_2 over the model M_1 the outcome is 0.97 which means that actually there is no difference between

these models and the support of M_2 over M_1 is not worth mention it.

$$BF_{23} = \frac{M_2}{M_3} = exp(-1.145 + 3.651) = 12.255.$$

For the comparison of model M_2 over M_3 we find that there is a strong evidence in support of model M_2 against M_3

$$BF_{24} = \frac{M_2}{M_4} = exp(-1.145 + 2.468) = 3.755.$$

Finally comparing model M_2 and M_4 there is a substantial evidence in support of M_2 over M_4 but not as strong as the previous comparison.

Part II

Bayesian Hypothesis Testing for Two-Way Contingency Tables

Chapter 2

Bayesian Hypothesis Testing for Two-Way Contingency Tables

2.1 Probability Structure for Contingency Tables

Let X and Y denote two categorical variables. X has I categories and Y has J categories as well. I categories of X apart from categories denotes the number of rows that will be placed in a contingency table. In the other hand J expect of categories of the variable Y denotes also the number of columns.

A possible subject in a randomised trial has IJ possible combinations of classification in a contingency table. Usually the response variable Y is being placed into columns and the explanatory X into rows. But sometimes both Y and X are response variables. In that case we focus on their joint distribution, which also determines the marginal and conditional distributions.

When Y is a response variable and X is an explanatory variable, we focus on the conditional distribution of Y and how it changes as the category of X changes.

Now let us examine the case that both X and Y are response variable: Suppose we conduct a randomised trial from a chosen population, such as in a sample survey employing simple random sampling. Then the X and Y are treated as both response variables of a randomly chosen subject which these variables have a probability distribution.

Let π_{ij} denote the probability that (X, Y) occurs in a cell in row i and column j. The probability distribution $\{\pi_{ij}\}$ is the joint probability of X and Y. The marginal distributions are the row and column totals that result from summing the joint probabilities. We denote these by $\{\pi_{i+}\}$ for the row variable and $\{\pi_{+j}\}$ for the column variable. The subscript " $+$ " denotes the sum over that index, which is:

$$\pi_{i+} = \sum_j \pi_{ij}$$

and

$$\pi_{+j} = \sum_i \pi_{ij}$$

The above equations can give us the total marginal probability :

$$\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij}$$

In the case that Y is the response variable and X explanatory variable we no longer have joint probability. In this case we are examining the probability that given a subject is in a row i of X to be classified in one column j of Y.

This probability is called the conditional probability and is denotes as $\pi_{j|i}$. So $\sum_j \pi_{j|i} = 1$. The conditional distribution of Y given X relates to the joint distribution by: $\pi_{j|i} = \pi_{ij}/\pi_{i+}$ for all i and j.

The cell frequencies are denoted by $\{n_{ij}\}$, and $n = \sum_i \sum_j n_{ij}$ is the total sample size. So:

$$p_{ij} = \frac{n_{ij}}{n}$$

The sample proportion of times that subjects in row i made response j is

$$p_{j|i} = \frac{p_{ij}}{p_{i+}} = \frac{n_{ij}}{n_{i+}},$$

Table 2.1: Probability Structure Table in 2 × 2 table

	Y		
X	1	2	Total
1	$\pi_{1\|1}$	$\pi_{2\|1}$	π_{1+}
2	$\pi_{1\|2}$	$\pi_{2\|2}$	π_{2+}
Total	π_{+1}	π_{+2}	1

where

$$n_{i+} = np_{i+} = \sum_j n_{ij}.$$

2.2 Distributional Sampling

In this section we will examine the distributional assumption in a contingency table under various sampling plans.The key point here is to represent different sampling methods before a survey is being conducted.

With the phrase different distributional sampling methods we mean how the variables X and Y are designed before the trial and the subjects were selected according to their cell frequencies or their conditional probabilities.

1. **Joint Multinomial Sampling**. In this kind of distributional sampling the total sample size n is fixed, but the row and column totals are not.The probability mass function of the cell counts has the multinomial form:
$$\left[\frac{n!}{(n_{11}1\ldots n_{IJ}!)}\right]\prod_i\prod_j \pi_{ij}^{n_ij}$$
Here the joint probabilities can be calculated because rows and columns are not fixed. We are interested of observing how much randomly the sample has been split into rows and columns.

2. **Independent Multinomial Sampling**. In this scheme there are two restrictions, either on the row totals or on the column totals. In other words, either all row margins or all column margins are fixed.

Consequently, the cell counts are multinomially distributed within each row or column. In experimental trial design in medicine or in social sciences, this is the most common sampling scheme. Suppose that the n_i observations on Y at using i of X are independent, each with probability distribution $\{\pi_{1|i}, \ldots, \pi_{j|i}\}$. The sampling scheme is called independent multinomial sampling and is given by:

$$\prod_i \left[\frac{n_i!}{\prod_j n_{ij}} \prod_j \pi_{j|i}^{n_{ij}} \right]$$

. The special case of the multinomial independent sampling when the $J = 2$ is the independent binomial sampling.

3. **Poisson Sampling**. Each cell count is random, and so is the total N size is not fixed. Each of the cell counts is Poisson distributed. This design often occurs in purely observational work.

   Poisson sampling model treats cell counts $\{Y_{ij}\}$ as independent Poisson random variables with parameters $\{\mu_{ij}\}$. The joint probability mass function for IJ cells is:

$$\prod_i \prod_j \frac{exp(-\mu_{ij})\mu_{ij}^{n_{ij}}}{n_{ij}!}$$

4. **Hypergeometric Sampling**. The scheme of hypergeometric distribution occurs when both row and column totals are fixed.The cell counts are said to be hypergeometrically distributed. Practical application of the hypergeometric with sampling scheme is rare.

   For the 2×2 table, an infinite number of examples can be constructed by classifying participants according to a median split on two continuous variables. For example, suppose we have 100 participants, with income and altruism as variables of interest.

   The first median split creates a group of 50 rich participants and 50 poor participants; the second median split creates a group of 50 altruistic participants and 50 arrogant participants. Hence, all row and column margins are fixed, and a single cell count suffices to uniquely identify the remaining three.

2.3 Testing Independence in Two-Way Contingency Tables

In this section we will represent the method that helps the frequentist approach to calculate the independence of Y and X. We are interested on examine whether there is an independence between the response variable Y and the explanatory X. Alternatively is there is an association between those two variables.

So we have to design a hypothesis test. The "tools" for this test are joint probabilities $\{\pi_{ij}\}$ in an $I \times J$ contingency table. Here we will first examine the independence test in a 2×2 contingency table.

The H_0 null hypothesis test of independence is $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. The alternative H_1 is $\pi_{ij} \neq \pi_{i+}\pi_{+j}$.

In the independence test problem the $\{n_{ij}\}$ is substituted by $\{n_i\}$ and with $\mu_{ij} = n\pi_{i+}\pi_{+j}$ in place of μ_i. Here $\mu_{ij} = E(n_{ij})$ under H_0. Usually,$\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown. The test statistic for independence test is:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

This function will give one value. This value follows asymptotically the chi-squared distribution with degrees of freedom $(I-1)(J-1)$. From the distribution table of chi-squared distribution this statistic will give us one probability.

The probability value (p-value). If this value is greater than $a = 0.05$ when a is the confidence level the H_0 is not rejected. If the value is less than $a = 0.05$ there is evidence against the null hypothesis H_0 i.e the variables X and Y are independent .

2.4 Bayesian Approach for Testing Independence in Two-Way Contingency Tables

For the analysis of 2×2 contingency tables, we will borrow the idea that Gunel and Dickey (1974) introduced, which is a generalization of $I \times J$ contingency tables Bayes analysis. Gunel and Dickey (1974) proposed the use of Bayes Factor for calculating the independence or I rows and J columns.

Below we describe, separately for each of the four sampling schemes that we have represented above. Bayes factors as we described them in Chapter 1 they test the H_0 in support of H_1 if H_0 is placed in the denominator and vice versa. The idea is the same for the row-column independence model H_0 over the row-column dependence model H_1.

Bayes factors are often difficult to calculate as we have showed, because they are obtained by integrating over the entire parameter space, a process that is non-trivial when the integrals are high-dimensional and intractable.

In order to describe how the Bayes factor are being calculated, we must first introduce the idea of a "conditional" Bayes factor. Consider that we want to test a simple normal mean and variance with two participants. Here we don't care for the hypothesis design for now. In other hand, we focus on the information in the data (likelihood) coming from participant to participant.

If we were sampling sequentially, we might compute the Bayes factor for our hypothesis after the first participant, and then after the second participant. The second Bayes factor takes into account all the data both coming from the first participant and the data information from the second participant. We can also look at the Bayes factor due to having observed second's participant data, already taking into account the data from participant 1.

This Bayes factor represents the "extra" information about the hypothesis offered by participant 2 over and above that offered by participant 1. We can call it the Bayes factor for participant 2 given, or conditional on, participant 1. However, we can partition the data in other ways besides participants.

Since the sample mean and variance jointly capture all the information in the data, we can also describe the Bayes factor for the sample mean

$$Y_{**} = \left(\begin{array}{c|cccc} & Y_1 & Y_2 & \dots & Y_J \\ \hline y_1 & y_{11} & y_{12} & \cdots & y_{1J} \\ y_2 & y_{21} & y_{22} & \cdots & y_{2J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_I & y_{I1} & y_{I2} & \cdots & y_{IJ} \end{array}\right)$$

conditioned on knowing the sample variance.

In the context of contingency tables, there are logical ways of partitioning the data. To begin, we partition the data into a part that contains the information about the overall quantity of observations, and a part that contains the information about how cells differ from one another.

To compute the evidence assuming that the total number of observations is fixed, we look at the change from the Bayes factor using only the first part of the data (the total number of observations) to the Bayes factor conditioned on the whole data set.

Due to the way of parameterization of models ,model parameters corresponding to the components of the partition this successive conditionalization produces Bayes factors that are easy to compute.

2.5 Bayes Factors according to distributional sampling

First we will clarify the notation that we are going to use for explaining the Bayes Factor for $I \times J$ contingency table.

Let y_{**} be a data matrix of I rows and J columns. and let α_{**} be a matrix of prior parameters with the same dimension as the data matrix y_{**}:

In vector form , $\vec{y} = (y_{11}, y_{12}, \dots, y_{IJ})$ and $\vec{\alpha} = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{IJ})$. For notation reasons dot is the summation across a single dimension (row or column) and star is the entire vector of that dimension. Which are clarified

$$\alpha_{**} = \left(\begin{array}{c|cccc} & \alpha_1 & \alpha_2 & \dots & \alpha_J \\ \hline \alpha_1 & \alpha_{11} & \alpha_{12} & \dots & \alpha_{1J} \\ \alpha_2 & \alpha_{21} & \alpha_{22} & \dots & \alpha_{2J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_I & \alpha_{I1} & \alpha_{I2} & \dots & \alpha_{IJ} \end{array}\right)$$

by the equations below:

$$y_{**} = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{IJ} = y_{11} + y_{12} + \cdots + y_{IJ}$$

$$y_{*.} = \sum_{i=1}^{I} y_{IJ} = (y_1, \dots, y_I)$$

$$y_{.*} = \sum_{j=1}^{J} y_{IJ} = (y_1, \dots, y_J)$$

$$\alpha_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} \alpha_{IJ}$$

$$\alpha_{*.} = \sum_{i=1}^{I} \alpha_{IJ} = (\alpha_{1.}, \dots, \alpha_{I.})$$

$$\alpha_{.*} = \sum_{j=1}^{J} \alpha_{IJ} = (\alpha_{.1}, \dots, \alpha_{.J})$$

$$\xi_{*.} = \alpha_{*.} - (J-1)$$

$$\xi_{.*} = \alpha_{.*} - (I-1)$$

$$\xi_{..} = \alpha_{..} - (I-1)(J-1$$

$$D(\alpha_{**}) = \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{\Gamma(\alpha_{IJ})}{\Gamma(\alpha_{..})}$$

To give a notion to the reader of this thesis, α_{**} is the matrix of the prior parameters. For example α is the scale parameter in Γ distribution for Poisson sampling models. For the multinomial distributional sampling that case is $\alpha = 1$, which says, that every combination of parameter values is equally likely a priori.

$\xi_{*.}$ is the number of rows (vector length) and $\xi_{.*}$ is the number of columns (vector length). Finally $D(\alpha_{**})$ is the Dirichlet distribution that we shortly explain below.

1. **Bayes factor under the Joint Multinomial sampling**. Under this sampling scheme, the total N or $y_{..}$ is fixed. Cell counts are assumed to be jointly multinomially distributed:

   $$y_{11}, \ldots, y_{IJ} \sim Multinomial(y_{..}, \pi_{**})$$

   and the conjugate prior selection for Multinomial parameters is the multidimensional version of Beta distribution, the Dirichlet distribution:

   $$\pi_{**} \sim Dirichlet(\alpha_{**})$$

   where π_{**} are the parameters to be estimated. The Bayes factor for independence under the joint multinomial sampling scheme is:

   $$BF_{01}^{M} = \frac{D(y_{*.} + \xi_{*.})}{D(\xi_{*.})} \frac{D(y_{.*} + \xi_{.*})}{D(\xi_{.*})} \frac{D(\alpha_{**})}{D(y_{**} + \alpha_{**})}.$$

   For the 2×2 table with $\alpha = 1$, the Bayes factor becomes:

   $$BF_{10}^{M} = \frac{6(y_{**} + 1)(y_{1.} + 1)}{(y_{**} + 3)(y_{**} + 2} \left[\frac{y_{11}! y_{12}! y_{21}! y_{22}!}{(y_{1.} + 1)! y_{2.}! y_{.1}! y_{.2}!} \right].$$

2. **Bayes factor under the Independent Multinomial Sampling**. Here we have to remind to the reader of this thesis that independent multinomial sampling can have either fixed number of rows or fixed number of columns. The Bayes factor that evaluates independence in support of the null hypothesis under this sampling scheme is :

   $$BF_{01}^{I} = \frac{D(y_{*.} + \xi_{*.})}{D(\xi_{.*})} \frac{D(y_{*.} + \alpha_{*.})}{D(\alpha_{*.})} \frac{D(\alpha_{**})}{D(y_{**} + \alpha_{**})}.$$

   When the row margins are fixed in a $I \times J$ table the Bayes factor is:

   $$BF_{01}^{I} = \frac{D(y_{*.} + \xi_{*.})}{D(\xi_{*.})} \frac{D(y_{.*} + \alpha_{.*})}{D(\alpha_{.*})} \frac{D(\alpha_{**})}{D(y_{**} + \alpha_{**})}.$$

   For the 2×2 contingency table, the Bayes factor for the independent multinomial sampling plan reduces to a test for the equality of two

proportions, $\theta 1$ and $\theta 2$. Under the default setting $\alpha = 1$ for prior Dirichlet Distribution, the Bayes Factor becomes:

$$BF_{01}^{I} = \left[\frac{\binom{y_{.1}}{y_{11}}\binom{y_{.2}}{y_{12}}}{\binom{y_{.1}+y_{y.2}}{y_{11}+y_{12}}}\right]\left[\frac{y_{11}!y_{12}!y_{21}!y_{22}!}{(y_{1.}+1)!y_{2.}!y_{.1}!y_{.2}!}\right].$$

where in the left parentheses lie the binomial coefficients.

3. **Bayes factor under the Poisson sampling**. Under this sampling scheme, none of the cell counts are fixed. Each cell count is assumed to be Poisson distributed: $y_{IJ} \sim Poisson(\lambda_{IJ})$. Each of the rate parameters λ_{IJ} is assigned a conjugate gamma prior with shape parameter α and scale parameter β:$\lambda_{IJ} \sim \Gamma(\alpha_{IJ}, \beta)$.Here,

$$\Gamma(\alpha_{IJ}, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \lambda > 0, \alpha > 0, \beta > 0$$

and $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = (\alpha - 1)$ The Bayes Factor for Poisson scheme is:

$$BF_{01}^{P} = (1+\frac{1}{\beta})^{(I-1)(J-1)}\frac{\Gamma(y_{..}+\xi_{..})}{\Gamma_{\xi_{..}}}\prod_{IJ}\frac{D(y_{*.}+\xi_{*.})}{D(\xi_{*.})}\frac{D(y_{.*}+\xi_{.*})}{D(\xi_{.*})}\frac{D(\alpha_{**})}{D(y_{**}+\alpha_{**})}$$

and for the special case for 2×2 table becomes:

$$BF_{10}^{P} = \frac{8(y_{**}+1)(y_{1.}+1)}{(y_{**}+4)(y_{**}+2}\left[\frac{y_{11}!y_{12}!y_{21}!y_{22}!}{(y_{1.}+1)!y_{2.}!y_{.1}!y_{.2}!}\right].$$

4. **Bayes factor under the Hypergeometric**. Hypergeometrical distributional scheme are not so often as the independent multinomial or poisson schemes because rows and column margins are fixed. Although the Bayes Factor for the hypergeometric scheme for 2×2 table is:

$$BF_{10}^{H} = \frac{y_{11}!y_{12}!y_{21}!y_{22}!y_{..}!}{(y_{1.}+1)!y_{2.}!y_{.1}!y_{.2}!}.$$

These results of each distributional scheme given their prior distributions are substantial for implementing Bayes Factors in contingency tables.

Usually the schemes that scientists adapt for their statistical designs are the joint multinomial, independent multinomial and poisson.

2.6 Bayes Factor example in Independence Test

Ending this subsection of Bayesian approach of independence of two categorical variables, we demonstrate an example of both the frequentist and bayesian framework.

Here, we provide an example analysis of Hraba and Grant's (1970) data, included as part of the BayesFactor package in R as the raceDolls data set.

71 white children and 89 black children from Lincoln, Nebraska were offered two dolls, one of whose "race" was the same as the child's and one that was different (either white or black). The children were then asked to select one of the dolls, with prompts such as "Give me the doll that is a nice doll." 50 of the 71 white children (70%) selected the white doll, while 48 of the 89 black children (54%) selected the black doll (see Appendix).

Table 2.2: Race Dolls Data

	Child		
Dolls	White Child	Black Child	Total
Same-Race Doll	50	48	98
Different-race doll	21	41	62
Total	71	89	160

The χ^2 statistic of independence gave us the value of 3.8566 and with degrees of freedom $df = 1$,the p.value is 0.04955 which a boundary evidence of association between the race of children and the colour of their doll selection.

The Bayesian approach of this independence test conducted with the BayesFactor package in R and the command $contingencyTableBF()$,which take as an input a $I \times J$ matrix and test the independence of columns and rows.

An additional argument on this command is that we must specify which distributional scheme this experiment was held.In our example the $sampleType =''$ $indepMulti''$ and the $fixedMargin =''$ $cols''$ to specify if the rows or the columns margins were fixed.

This argument executes the Bayes Factor depending on the distributional

sampling and the prior selection of Gunel & Dickey (1974) as we described above.

The Bayes factor in favor of the alternative that the categorical factors are not independent is 1.815, which is not providing strong enough evidence against the null hypothesis. For the full R code see at the Appendix.

2.7 Bayesian Estimation of Odds Ratio and Difference of Two Binomial Proportions

In term of probabilities in a 2×2 contingency table, the Relative Risk is defined as:

$$RR = \frac{\pi_1}{\pi_2}$$

,which is the ratio of the two proportions of two different groups. Usually in Biostatistics the main purpose of calculating the RR is to examine two groups of exposure and no exposure in a disease in a clinical trial.

Many sociologists and psychologists use the RR in their surveys for a quick calculations in the difference in two examining groups. RR is by default a nonnegative real number.

A relative risk of 1.00 occurs when $\pi_1 = \pi_2$, that is, when the response variable is independent of the group. The ratio of failure probabilities, $(1 - \pi_1/(1 - \pi_2)) = \pi_2 - \pi_1$ takes a different value than the ratio of the success probabilities.

From the other hand Odds Ratio is the another measure of association for a 2×2 contingency table for a probability of success π, the OR are defined as:

$$OR = \frac{\pi}{(1 - \pi)}.$$

Also the success probability π is a function of the odds,

$$\pi = \frac{odds}{(odds + 1)}$$

In 2×2 tables, within row 1 the odds of success are $odds_1 = \pi_1/(1-\pi_1)$,and within row 2 the odds of success equal $odds_2 = \pi_2/(1-\pi_2)$. So the ratio of the odds from a 2×2 table is:

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

A more typical form of Odds Ratio calculation type according to notation that we have already used in this thesis and is very common mathematically in practice is :

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

The relationship between Relative Risk and Odds Ratio is:

$$OR = RR\left(\frac{1-\pi_2}{1-\pi_1}\right).$$

Bayesian approach in a 2×2 contingency table for calculating the odds ratio we will assume that the cell counts are poisson distributed with:

$$Y_{11} \sim Poisson(\lambda_{11})$$

$$Y_{12} \sim Poisson(\lambda_{12})$$

$$Y_{21} \sim Poisson(\lambda_{21})$$

$$Y_{22} \sim Poisson(\lambda_{22})$$

These λ's parameters have prior distribution $\lambda \sim Gamma(\alpha, \beta)$. Finding the posterior distribution for the conjugate analysis we will end as we have showed in chapter 1:

$$\lambda_{11} \mid y_{11} \sim Gamma(y_{11} + \alpha, N + \beta)$$

$$\lambda_{12} \mid y_{12} \sim Gamma(y_{12} + \alpha, N + \beta)$$

$$\lambda_{21} \mid y_{21} \sim Gamma(y_{21} + \alpha, N + \beta)$$

$$\lambda_{22} \mid y_{22} \sim Gamma(y_{22} + \alpha, N + \beta)$$

where α and β are common scale and rate Gamma parameters and N is the grand total count of all cells. Generating $\theta^\star$ random numbers from a Gamma

distribution for λ's we are able to estimate a new simulated odds ratio, which is denoted:

$$OR^{\star} = \frac{\lambda_{11}^{\star} \times \lambda_{22}^{\star}}{\lambda_{12}^{\star} \times \lambda_{21}^{\star}}$$

Taking advantage of the Monte Carlo mean estimator:

$$\frac{1}{N_{\theta^{\star}}} \sum OR^{\star}$$

we can have the new estimated odds ratio from the posterior simulation.

2.8 Odds Ratio Simulation example

In the electronic Appendix the example from Hennekens (1987) study is provided. From the data of this study we obtain a 2×2 contingency table with 104 myocardial infractions (fatal and non-fatal) among 11.037 patients in the treatment group and 189 myocardial infractions incidents among 11.034 patients in the placebo group and is presented below:

Table 2.3: Hennekens Data

	Drug		
Status	Treatment	Placebo	Total
Exposed	104	189	293
Non-Exposed	11037	11034	22071
Total	11141	11223	22364

Calculating the odds ratio for this table we take:

$$OR = \frac{104 \times 11.034}{189 \times 11037} = 0.55.$$

The simulated $OR^{\star}$ with Gamma parameters $\alpha = 1$ and $\beta = 1$, we obtained:

$$OR^{\star} = 0.55,$$

with standard error equal to 0.067.

2.9 Difference in Proportions (A/B Testing) example

A/B Testing is comparison procedure of two different method of marketing. Companies are very interested knowing which of the two marketing strategies is better or more highly likely to occur. A simple mathematic division contains the outcome of a method. This outcome is called Conversion Rate and is actually a proportion of the total number customers e.g divided by the total number of visitors in a web application.

$$ConversionRate = \frac{\#Customers}{\#Visitors}$$

For example, let us assume that you own a web application which provides to the user taxi services and calling a taxi cab by the single click in your mobile phone. The marketing department of this company have done a study and tried two different versions of the app and wanted to know which one version is better according to the rides made from each version:

The data analysts received the numbers from the first version and saw that 72 users from 600 that they saw the saw the version A the same day took a taxi ride, and 120 from 750 that saw the version B took a ride the same day or some minutes after they saw the competition.

So now we have 72 conversions from method A and 120 from from method B. The next matrix shows the No Conversions in the first column and Conversions in second column.

$$A = \begin{bmatrix} 600 & 72 \\ 750 & 120 \end{bmatrix}.$$

There is no need to write that the conversion rate is actually statistically speaking the p in the Binomial Distribution.Binomial Distribution has two parameters N which represents the total number of trials and p success probability for each trial. The distribution is :

$$Y_i \sim Binomial(N, p).$$

In our example we have:

$$Y_1 \sim Binomial(600, p_1)$$

$$Y_2 \sim Binomial(750, p_2)$$

The conversion

$$\hat{p}_1 = \frac{72}{600} = 0.12$$

and

$$\hat{p}_2 = \frac{120}{750} = 0.16$$

In our web taxi app problem we have two two binomial samples with size $N1$ and $N2$ with parameters $\hat{\theta}_1$ and $\hat{\theta}_2$. Those two parameters now due to conjugate prior analysis they follow another distribution called Beta Distribution which has two scale parameters α andβ. So our random parameters that need to be estimated are

$$Y_1 \sim Binomial(N_1, \theta_1)$$

$$Y_2 \sim Binomial(N_1, \theta_2)$$

for Difference of Proportions we used prior:

$$\hat{\theta}_1 \sim Beta(3, 25)$$

$$\hat{\theta}_2 \sim Beta(3, 25)$$

Assuming beta priors for the success probabilities θ_1, θ_2 with parameters (α_1, β_1) and (α_2, β_2) , respectively, we end up with posteriors given by:

$$\theta_1 \mid y_1 \sim Beta(y_1 + \alpha_1, N_1 + \beta_1)$$

and

$$\theta_2 \mid y_2 \sim Beta(y_2 + \alpha_2, N_2 + \beta_2)$$

Substituting the numbers we obtain:

$$\theta_1 \mid y_1 \sim Beta(72 + 3, 600 + 25)$$

and

$$\theta_2 \mid y_2 \sim Beta(120 + 3, 750 + 25)$$

Next we generate 1000 numbers from a beta distribution with the given parameters as above:

Generate

$$\theta_1^{n*} \sim beta(75, 625)$$

and

Generate

$$\theta_2^{n*} \sim beta(123, 775)$$

where $n*$ are the number of simulation pseudo random numbers.

The final step is to subtract θ_2 from θ_1 to calculate the difference in proportions.

For the Difference of Proportions, R reported 0.03 with standard error 0.01 and according to the posterior distribution with around probability $p = 0.96$, the $p(\beta)$ is higher than $p(\alpha)$, so version B is more successive than version A.

Part III

Bayesian Analysis for Generalized Linear Regression Models

Chapter 3

Bayesian Analysis for Generalized Linear Regression Models

3.1 Logistic Regression

Generalized linear modelling is a framework for statistical analysis that includes linear, log linear and logistic regression (as well as other cases from the same exponential family distributions) as special cases as have been proposed from Nelder & Wedderburn (1972).

Linear regression directly predicts continuous data y from a linear predictor $X\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$. Logistic regression predicts $P(y = 1)$ for binary data from a linear predictor with an inverse logit transformation.

A generalized linear model involves:

1. **Random Part**: Data vector $y = (y_1, \ldots, y_n$ that are identically independent distributed (iid) from the same exponential family.
2. **Systematic Part** : Here we will find the linear predictor $\eta_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$,where $X_{i1}, \ldots, X_{iK}$ are the explanatory variables and $\beta_0 + \beta_1 + \cdots + \beta_K$ are the estimated coefficients.This predictor η_i is linear

in respect to the parameters and not to the (in)dependent variable X. The assumptions about X variable, of zero correlation or independence or lack of interaction or additivity, cannot be taken for granted in GLM and must be checked. The coefficients $\hat{\beta}$ are Normally distributed and due to that,the statistical inference is for the population average.

$$\hat{\beta}_0 \sim N(\hat{\beta}, \sigma^2_{\beta_0}),$$

$$\hat{\beta}_1 \sim N(\hat{\beta}, \sigma^2_{\beta_1}),$$

$$\dots$$

$$\hat{\beta}_k \sim N(\hat{\beta}, \sigma^2_{\beta_k}).$$

3. **Link Function** : A link function $\mathbb{G}$, yielding a vector of transformed data $\hat{y} = \mathbb{G}^{-1}(\beta X)$ that are used to model the data.

Logistic regression is the standard way to model binary outcomes (that is, data y_i that take on the values 0 or 1). It would not make sense to fit the continuous linear regression model, $X\beta + error$, to data y, that take on, the values 0 and 1. Instead, we model the probability that $y = 1$ to a non-linear transformation:

$$P(y_i = 1|x) = \mathbb{G}(X_i\beta) = \frac{\exp(x)}{1 + \exp(x)}.$$

This is called the inverse logit, $logit^{-1}(\eta_i) = \mathbb{G}^{-1}(\eta_i)$. Alternatively we can write:

$$\mathbb{G}^{-1} = \frac{\exp(x)}{1 + \exp(x)}$$

This function transforms continuous values to the probability range (0, 1),where

$$\mathbb{G}(\pi) = log\left[\frac{\pi}{1-\pi}\right]$$

is a function mapping the range (0, 1) to the range $-\infty$ to ∞ in the whole real line. We prefer to work with $\mathbb{G}^{-1}$ because it is natural to focus on the mapping from the linear predictor to the probabilities, rather than the reverse. For

instance, if a linear logistic model has been used with two covariates χ_1 and χ_2,we have the model:

$$log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2$$

for the log odds of a positive response. Equivalently, the model may be written in terms of the odds of a positive response, giving:

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1\chi_1 + \beta_2\chi_2).$$

Finally the probability of a positive response yields the inverse function which is:

$$\pi = \frac{\exp(\beta_0 + \beta_1\chi_1 + \beta_2\chi_2)}{1 + \exp(\beta_0 + \beta_1\chi_1 + \beta_2\chi_2)}$$

The curve of the linear regression is a straight line.In logistic regression (generalized linear models) the regression curve through this non-linearity is S-curved. Inverse-logistic function is curved, and so the expected difference in y corresponding to a fixed difference in x is not a constant. Exponentiating logistic regression coefficients can be interpreted as odds ratios. For simplicity, we illustrate with a model with one predictor X_1, so that:

$$log\left[\frac{P(y=1|x=1)}{P(y=0|x=1)}\right] = \beta_0 + \beta_1 X_1.$$

In logistic regression there are two different data inputs:

1. **Ungrouped Data**: Are data that y_i have binary form of 0,1 and are Bernoulli Distributed,with $n = 1$.

$$y_i \sim Bernoulli(\theta_i)$$

2. **Grouped Data** : Are data that y_i have the form of successes and failures.These y_i are Binomial Distributed.

$$y_i \sim Binomial(N_i, \theta_i)$$

Deviance is a goodness-of-fit statistic for a GLM model. It is often used for statistical hypothesis testing. It is a generalization of the idea of using the sum of squares of residuals in ordinary least squares to cases where model fitting is achieved by maximum likelihood as it is presented in this section. It plays an important role in exponential dispersion models and generalized linear models. Unbiased estimation of ϕ is the:

$$\phi = \frac{deviance}{df}.$$

The priors that we must give are on interest of parameters β_i in the linear predictor are usually normal priors for these parameters, given the dispersion parameter ϕ.

$$\hat{\beta}_i|\phi \sim Normal(\mu_{\beta_i}, \sigma^2_{\beta_i}|\phi)$$

The variance of the normal prior depends on the dispersion parameter ϕ in order to achieve an appropriate scaling of the prior distribution. When no prior information is available, the prior mean is set equal to zero, while the corresponding variance is set large to express prior ignorance. Alternatively, a prior independent to the dispersion parameter can be considered.

When the variance is set large to express prior ignorance, then no differences in the resulting posterior distribution will be observed. Another alternative is for diffuse priors is the student t distribution as prior due to its fat tails. Another alternative for prior selection when there is no information to the analyst is the *Cauchy* distribution which has even fatter tails than Student t.

3.2 Bernoulli & Binomial Regression examples via Hamiltonian Monte Carlo

To illustrate an example we will use the data (Bernoulli) of NASA's Challenger Disaster from the spaceship's rings as presented in Appendix. The cause of the Challenger lift off disaster was traced to an O-ring, a circular gasket that sealed the right rocket booster. NASA before Challenger's disaster in 1986 had 23 previous shuttle missions that tested the O-rings and their

resistance to temperature. The independent variable y_i are binary, which 1 indicated failure and 0 otherwiset. The X variable is the temperature at launch in degrees Fahrenheit.

The implementation of frequentist approach of GLM model,R reported the $\eta = 15.0429 - 0.2322$. So the probability of failure on Orings boosters in the mean temperature (69.56) is :

$$\pi = \frac{\exp(15.0429 - 0.2322 * 69.56)}{1 + \exp(15.0429 - 0.2322 * 69.56)} = 0.24.$$

For the Bayesian approach we are using the Stan modelling language which uses Euclidean Hamiltonian Monte Carlo Simulation. The coefficients β_0 β_1 were calculated both with prior $Normal(\mu = 0, \sigma = 100)$.Stan reported $\beta_0 = 18.678$ and $\beta_1 = -0.286$.

The model is

$$\eta = 18.678 - 0.286 \times \chi_{thermal}.$$

The probability of failure in the mean temperature is :

$$\pi = \frac{\exp(18.678 - 0.286 * 69.56)}{1 + \exp(18.678 - 0.286 * 69.56)} = 0.22,$$

but the probability of failure of the O'Rings in the temperature of 31 which was the temperature the time that Challenger exploded is :

$$\pi = \frac{\exp(18.678 - 0.286 * 31)}{1 + \exp(18.678 - 0.286 * 31)} = 0.99$$

Table 3.1: NASA Challenger Analysis

RStan Regression Analysis							
Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
b_0	1.0	951	18.7	8.4	4.9	17.6	38.0
b_1	1.0	960	-0.3	0.1	-0.6	-0.3	-0.1
log-posterior	1.0	1526	-11.3	1.1	-14.0	-10.9	-10.2

For Binomial regression (grouped data) we present data from World Cup of Basketball in China 2019. We focus on Facundo Campazzo's 3 point shots

Table 3.2: Facundo Campazzo's Attempts-Made

Facundo Campazzo 3 point shots		
Versus	Attempts	Made
South Korea	3	4
Nigeria	3	8
Russia	2	7
Venezuela	2	3
Poland	1	4
Serbia	3	6
France	3	9
Spain	1	5

attempts through out the tournament in 8 games (including the final). The data are the following: Attempts:$4, 8, 7, 3, 4, 6, 9, 5$ and made:$3, 3, 2, 2, 1, 3, 3, 1$.

We have selected prior Normal for regression's intercept,$\beta_0 \sim Normal(0, 1)$. Stan reported $\beta_0 = -0.89$ with $sd = 0.26$. We can conclude that the probability of success in Facundo Campazzo's 3 point shot is:

$$\pi = \frac{\exp(-0.89)}{1 + \exp(-0.89)} = 0.29$$

Table 3.3: Facundo Campazzo's Logistic Regression Analysis

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
α	1.0	1549	-0.89	0.26	-1.4	-0.9	-0.4
log-posterior	1.0	1176	-38.9	0.7	-40.9	-38.7	-38.4

3.3 Logistic Regression via Laplace-Metropolis algorithm

As we have seen in Chapter 1, Kass and Raftery (1995) proposed a method for estimating the marginal likelihood by combining the Laplace approximation with MCMC for bayes factor calculation.

They suggested the Metropolis algorithm as a means for estimating the quantities required for the Laplace approximation. By applying the Laplace method we can derive the following approximation for the marginal likelihood:

$$f(Y) = (2\pi)^{P/2} \mid H \mid^{1/2} f(\theta^{\star}) f(Y \mid \theta^{\star}),$$

where $\theta^{\star}$ is the value of θ at which $h(\theta) = \log\{f(\theta)f(Y \mid \theta)\}$ attains its maximum $H^{\star}$ is the minus the inverse Hessian matrix of second derivative of h evaluated at $\theta^{\star}$ and P is the dimension of the parameter space.

For numerical reasons and since it is customary to work with log likelihoods, it is better to work with this approximation on the logarithmic scale. Taking logarithms, we can rewrite the Laplace approximation as:

$$\log(f(Y)) = \frac{P}{2}\log(2\pi) + \frac{1}{2}\log(\mid H^{\star} \mid) + \log(f(\theta^{\star})) + \log(f(Y \mid \theta^{\star}))$$

We refer to this estimator as the Laplace estimator.

In the Laplace-Metropolis estimator there are two key quantities that we need to derive from the posterior simulation output, namely $\theta^{\star}$ and $H^{\star}$. The conceptually simplest way to estimate $\theta^{\star}$ would be to compute $h(\theta)$ for each draw from the posterior simulation output and use that θ for which $h(\theta)$ is largest for $\theta^{\star}$. This can take a lot of computing effort and resources.

As an alternative we consider the multivariate median, or L_1 center, which is defined as the value of θ which minimises:

$$d(\theta^{(j)}) = \sum_{l=1}^{J} \mid \theta^{(l)} - \theta^{(j)} \mid,$$

where $\mid \theta^{(l)} - \theta^{(j)} \mid$, denotes L_1 distance. We use this as an estimator of the posterior mode.

The other Laplace quantity needed for the the Laplace-Metropolis estimator is $H^{\star}$. This is asymptotically equal to the posterior variance matrix. We could use the sample covariance matrix of the simulation output for $H^{\star}$.

The package in R that implements the Laplace-Metropolis algorithm for the calculation of the marginal likelihood is the MCMCpack. With the command *MCMClogit* we have implemented the NASA's Challenger example with

prior for intercept $\beta_0 \sim Normal(15, 1)$ and $\beta_1 \sim Normal(15, 1)$. We obtained the mean of the intercept $\beta_0 = 14.7799$, with standard deviation equal to 0.2528 and for $\beta_1 = -0.2158$ with standard error equal to 0.007995.

3.4 Log-Linear Models

The Poisson distribution is used to model variation in count data (that is, data that can equal $0, 1, 2, \dots$). In the Poisson model, each unit i corresponds to a setting (typically a spatial location or a time interval) in which y_i events are observed.

The Poisson distribution:

$$P(y) = \frac{e^{-\mu}\mu^y}{y!}$$

, has only one adjustable parameter, namely the mean μ,which must be positive. Thus the mean alone determines the distribution entirely. Since the Poisson mean is required to be positive, an additive model for μ is normally considered to be unsatisfactory.

Hence, although $\mu = \sum \beta_i x_i$ it might become, it may be found over the range of the data, it is often scientifically dubious and logically unsatisfactory for extrapolation. In the model with multiplicative effects, we set $\mu = \exp^{\eta}$ and η rather than μ obeys the linear model. This construction ensures that μ remains positive for all η and hence positive for all parameter and covariate combinations.

As with linear and logistic regression, the variation in y can be explained with linear predictors X.For example, in a log linear model predictors could include: a constant term, a measure of the mean for variable $\beta_i x_i$. The basic Poisson regression model (systematic component)has the form:

$$\eta = \beta_0 + \beta_1 x_1, \dots, \beta_k x_k$$

, The canonical link of Poisson distribution for generalized linear models is $\eta = log\mu$.The cumulant function is $exp(\theta)$. The problem with poisson models is that the the parameter λ (here presented as μ)is the mean and also the

variance, hence the standard deviation equals the square root of the mean, as a result and by default we expect to have overdispersion. Note here that the ϕ variant parameter is also 1 but in practice is rarely achieved.

In Poisson log-linear models, the effect of each X , is linear to the log-mean of Y , resulting in an exponential effect of X, on the mean of Y. Then the mean of Y can be expressed as:

$$\begin{aligned} \log \lambda_i &= \beta_0 + \beta_1 X_i \\ \lambda_i &= e^{\beta_0} e^{\beta_1 x_i} \\ \lambda_i &= B_0 B_i^{x_i}, \end{aligned}$$

where $B_j = e^{\beta_j}$ for $j = 0, 1$, where B_0 denotes the expected counts when the covariate is equal to zero ($X = 0$). Interpretation of β_1 is slightly different from the corresponding one in normal models since relative mean differences are considered in the Poisson case. Let us denote by $\lambda(x) = E(Y \mid X = x)$ the expected counts for covariate with $X = 2$.Then:

$$\log [\lambda(x + 1)] - log(\lambda(x)) = \beta_1,$$

resulting in:

$$\lambda(x + 1) = B_1 \lambda(x) = e^{\beta_1} \lambda(x).$$

From that, we conclude, that when the variable X is increasing by one unit the expected counts (of Y) equals to B_1 times the corresponding value of Y for $X = x$. An other alternative interpretation of B_1 coefficient is the percentage.We take $(B_1 - 1) \times 100$ when X increases by one unit of measurement.

When the variable X is categorical with K levels, then the linear predictor is expressed as a linear function of $K - 1$ dummy variables denoted by D_j, for $j = 2, \ldots, K$. For the simpler case of (j = 1) as the reference baseline category we set $\beta = 0$. Then we express the model by:

$$\log \lambda_i = \beta_0 + \sum_{j=2}^{K} \beta_j D_{ij} \Leftrightarrow$$
$$\lambda_i = e^{\beta_0} exp\left(\sum_{j=2}^{K} \beta_j D_{ij}\right)$$
$$\lambda_i = B_0 \prod_{j=2}^{k} B_j^{D_{ij}}$$

where $B_j = e^{\beta_j}$ for $j \in (0, 2, 3, \ldots, K)$.

If the individual i belongs in the first category of X we have $(X_i = 1)$, then $\lambda_i = B_0$, while the individual i belongs in the κth category $(\kappa > 1)$ of X, $(X_i = j)$, then:

$$\lambda(X = \kappa) = B_0 B_\kappa = B_\kappa \lambda(X = 1).$$

Therefore, quantity $B_\kappa = e^{\beta_k}$ can be now interpreted as the relative change of the Poisson expectation λ when an individual belongs in κ category of X compared to the baseline-reference category. The interpretation for the sum to zero parameterization is similar, but all coefficients express the relative change of the current level compared with an overall "average" level instead of the baseline category used in corner parameterization.

This is similar to the interpretation of the parameters in the multiple Poisson regression case. The difference here is that in every change of a single explanatory variable (say, X_j), other covariates need to remain constant since:

$$\lambda_i = B_0 \prod_{j=1}^{p} B_j^{x_{ij}},$$

where $B_j = e^{\beta_j}$ for $j = (0, 1, 2, \ldots, p)$.

In Bayesian inference, a usual point estimate for the model parameters is provided by the posterior means (or medians). Alternatively, the exponent of the posterior mean or median of β_j, can be used as estimates for inference based on the bayesian probabilistic framework modelling.

For the Bayesian perspective the priors in this case of log-linear models can be expressed for β_j coefficients. According to generalized linear models theory the coefficients follow the Normal distribution $\beta_j \sim Normal(\hat{\beta}_j, \hat{\sigma}_{\beta_j})$ similar to the logistic regression as we saw in the previous section. The prior distribution for log-linear models coefficients in Stan can be normal, student t or $Cauchy$, including the intercept coefficient β_0.

Due to simulation algorithm (Euclidean Hamiltonian Dynamics) used by STAN, the calculation of the target posterior distribution is given by the $-log$ likelihood, the inverse transformation is included but for the posterior, not for the exponentiation of the parameters. This calculation can be held inside the model code or separately.

Next we demonstrate both cases in the same example for two different R packages that implement Hamiltonian Monte Carlo from Stan in R. One is from RStan that is the syntax Stan modelling and the second is Rstanarm that has the same form command as the $glm()$ command in base R.

For example purposes we take the data from eba1977 in the ISwR package in R which is data frame with 24 observations on 4 variables. The y variable is the number of lung cancer cases. The x_i predictor varaibles are:1) x_1 which is the age, a factor with levels 40-54, 55-59, 60-64, 65-69, 70-74, and 75+. 2) is the x_2 which is the city variable, a factor with levels Fredericia, Horsens, Kolding, and Vejle,and 3) x_3 is the population which is the number of inhabitants of each 4 cities.

Putting the logarithm of the population into the model as an offset, is equivalent to including it as a regression predictor, but with its coefficient fixed to the value 1. Another option is to include it as a predictor and let its coefficient be estimated from the data. In some settings, this makes sense in that it can allow the data to be fit better. In other settings, it is simpler to just keep it as an offset so that the estimated rate θ has a more direct interpretation.

The exponentiated coefficients gave us for $e^{\beta_1} = e^{1.10} = 3.004$ for the first category of age factor. So the r of the expected lung cancer for the age category of 55-59 is 3 times higher from the other age categories with the variable city constant.

Table 3.4: Log Linear Stan results

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β_0	1.0	5554	-5.6	0.2	-6.0	-5.6	-5.3
β_1	1.0	7152	1.1	0.3	0.6	1.1	1.6
β_2	1.0	6970	1.5	0.2	1.1	1.5	2.0
β_3	1.0	6971	1.8	0.2	1.3	1.8	2.2
β_4	1.0	6865	1.9	0.2	1.4	1.8	2.3
β_5	1.0	7437	1.4	0.3	0.9	1.4	1.9
β_6	1.0	9331	-0.3	0.2	-0.7	-0.3	0.0
β_7	1.0	9796	-0.4	0.2	-0.7	-0.4	-0.0
β_8	1.0	10014	-0.3	0.2	-0.7	-0.3	0.1

For the age 60-64 , $e^{\beta_2} = e^{1.52} = 4.57$ and for the age 65-69, $e^{\beta_3} = e^{1.77} = 5.87$, we conclude that as the factor age grows the estimated ratio of lung cancers is getting higher. For the city variable Fredericia is the reference level. According to the output the chance that the inhabitants of city Horsens having a lung cancer is $1 - 0.71 = 0.29$ lower from inhabitants of city Fredericia.

3.5 Log linear Models for counts in contingency tables

First we consider an $R \times C$ contingency table that cross classifies n subjects on two categorical response variables, a row variable X and a column variable Y. When X and Y are statistically independent, the joint cell probabilities $\{\pi_{ij} = P(X = i, Y = j)\}$ are determined by the row and column marginal probabilities,

$$\pi_{ij} = P(X = i)P(Y = j) = \pi_{i+} + \pi_{+j}, i = 1, ..., r, j = 1, ..., c.$$

The cell probabilities π_{ij} are parameters for a multinomial distribution. Loglinear model formulas use expected frequencies $\{\mu_{ij} = n\pi_{ij}\}$ rather than π_{ij}. Then they apply also to the Poisson distribution for cell counts with expected values $\{\mu_{ij}\}$. Under independence, $\mu_{ij} = n\pi_{i+} + \pi_{+j}$ for all i and j.

The independence condition, $\mu_{ij} = n\pi_{i+} + \pi_{+j}$,is multiplicative. Taking the log of both sides of the equation yields an additive relation. That is, independence has the form:

$$log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y,$$

with an intercept coefficient based on the sample size of n, a row effect coefficient λ_i^X based on the probability in row i, and a column effect coefficient λ_j^Y based on the probability in column j. This model is called the loglinear model of independence. The larger the value of λ_i^X, the larger each expected frequency is in row i. The larger the value of λ_j^Y , the larger each expected frequency is in column j (here the X and Y superscripts are labels for the variables, not power exponents).

Loglinear models for contingency tables are generalized linear models that treat the cell counts as independent observations from Poisson distributions and use the log link function. As the log linear model of independence suggests, loglinear models do not separate response and explanatory classification variables.

This formula specifies how the expected cell counts vary according to the categories of X and Y . The model regards the observations to be the cell counts rather than the classifications of individual subjects.

For $R \times 2$ contingency tables, for instance, the logit in row i is:

$$logit\left[Pr(Y=1)\right] =$$
$$\log\left[\frac{Pr(Y=1)}{1-Pr(Y=1)}\right] =$$
$$\log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) = \log(\mu_{i1}) - \log(\mu_{i2})$$
$$(\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) =$$
$$\lambda_1^Y - \lambda_2^Y,$$

Interpretation is carried out in terms of the odds. For given column category j, under log linear independence model , the odds of being in row $i1$ instead

of row $i_2 (i_1 \neq i_2), i_1, i_2 = 1, \ldots, I$, is:

$$\frac{\mu_{i1j}}{\mu_{i2j}} = \frac{exp(\lambda + \lambda_{i1}^X) + \lambda_j^Y)}{exp(\lambda + \lambda_{i2} + \lambda_j^Y)} = exp(\lambda_{j1}^Y - \lambda_{j2}^Y), i = 1, \ldots, I,$$

i.e., the odds of being in row j_1 instead of j_2 is determined only by the distance of the corresponding column main effect values and is independent of i. The conditional columns j_1 and j_2 column probabilities (within row i) we have:

$$\frac{Pr(Y = j_1 \mid X = i)}{Pr(Y = j_2 \mid X = i)} = exp(\lambda_{j_1^Y} - \lambda_{j_2^Y}), i = 1, \ldots, I,$$

relate the same for all rows and this is true for any pair of columns j_1 and j_2. Thus, the conditional column distribution is the same for all rows, as should be for independent X and Y .

In case the classification variables X and Y are not independent, their interaction is significant and the corresponding XY interaction term has to be added in the log linear model expression, leading to the saturated model and we defined them:

$$log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, i = 1, \ldots, I, j = 1, \ldots, J.$$

The $\{\lambda_{ij}^{XY}\}$ parameters are association terms. The parameters represent interactions between X and Y , whereby the effect of either variable on the expected cell count depends on the category of the other variable. Direct relationships exist between log odds ratios and the $\{\lambda_{ij}^{XY}\}$. For example, this ij model for 2×2 contingency tables has the log odds ratio:

$$log\theta = log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) = log\mu_{11} + log\mu_{22} - log\mu_{12} - log\mu_{21} =$$

$$(\lambda+\lambda_1^X+\lambda_1^Y+\lambda_{11}^{XY})+(\lambda+\lambda_2^X+\lambda_2^Y+\lambda_{22}^{XY})-(\lambda+\lambda_1^X+\lambda_2^Y+\lambda_{12}^{XY})-(\lambda+\lambda_2^X+\lambda_1^Y+\lambda_{21}^{XY}) =$$
$$\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

With three-way contingency tables, loglinear models can represent various independence and association patterns. Two-factor association terms

describe conditional odds ratios between variables. For cell expected frequencies $\{\mu_{ijk}\}$, consider the loglinear model:

$$log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

Since it contains an XZ term, it permits association between X and Z, at each category for Y . It also permits a YZ association, at each category for X. It does not contain an XY term, so this loglinear model specifies independence between X and Y , at each category for Z, that is, conditional independence, so given Z, X does not depend on Y.

This model holds when an association between two variables (X and Y) disappears after we adjust for a third variable (Z). We symbolise the model by (XZ, YZ). The symbol lists the highest-order terms in the model for each variable.

Models that delete additional association terms are too simple to fit most data sets well. For instance, the model that contains only single-factor terms, denoted by (X,Y,Z), is called the mutual independence model.

It treats each pair of variables as independent, both conditionally and marginally. When variables are chosen wisely for a study, this model is rarely appropriate. A model that permits all three pairs of variables to have conditional associations is:

$$log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}\lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

For it, we will see that conditional odds ratios between any two variables are the same at each category of the third variable. This is the property of homogeneous association. We symbolise this loglinear model, called the homogeneous association model, by (XY, XZ, YZ). The most general loglinear model for three-way tables adds a three-factor interaction term,λ_{ijk}^{XYZ}, to the homogeneous association model. Denoted by (XYZ), it is the saturated model. It provides a perfect fit.

Table 3.5: Marijuana Log-Linear Count Analysis

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
(Intercept)	1.0	3628	6.8	0.0	6.7	6.8	6.9
a	1.0	1894	-5.4	0.5	-6.4	-5.4	-4.6
c	1.0	1919	-3.0	0.2	-3.3	-3.0	-2.7
m	1.0	3135	-0.5	0.1	-0.6	-0.5	-0.4
a:c	1.0	1499	1.9	0.8	0.1	1.9	3.4
a:m	1.0	1979	2.8	0.5	1.9	2.8	3.9
c:m	1.0	1945	2.8	0.2	2.5	2.8	3.2

3.6 Multinomial Regression

Multinomial regression is a generalization of logistic regression which applies to categorical responses that have more than two categories. Models for nominal-scale response variables treat the categories as unordered. Explanatory variables can again be quantitative, categorical (using indicator variables), or both. We let K denote the number of categories of the response variable Y .

The response probabilities $(\pi_1, \ldots, \pi_K)$ at any setting for the explanatory variables satisfy in the previous subsections. The analysis of this section apply when the sample consists of independent observations.

When all explanatory variables are discrete, the data file can be ungrouped or can have the grouped-data form of counts in the K categories of Y at each setting of the explanatory variables. The models assume that those counts have a multinomial distribution the multicategory generalization of the binomial distribution.

The multicategory logit model for nominal response variables simultaneously uses all pairs of categories by specifying the odds of outcome in one category instead of another. The order of listing the categories is irrelevant. The basic model formula pairs each category with a baseline category. Software (R) usually sets the first category (K) as the baseline, in which case

the baseline-category logits are:

$$log\left(\frac{\pi_j}{\pi_K}\right), j = 1, \ldots, K-1,$$

but this is adjustable (reference category can be changed manually).

For $K = 3$, for instance, the model uses $log\left(\frac{\pi_1}{\pi_3}\right)$ and $log\left(\frac{\pi_2}{\pi_3}\right)$ uses the last category as reference. Conditional on the response falling in category j or in category K, $log\left(\frac{\pi_j}{\pi_K}\right)$ is the log odds that the response is j. The baseline-category logit model with an explanatory variable x is:

$$log\left(\frac{\pi_1}{\pi_3}\right) = \alpha_j + \beta_j x, j = 1, \ldots, K-1.$$

The model has $\kappa - 1$ equations, with separate parameters for each $\kappa - 1$ categories. The effects vary according to the category paired with the baseline. These equations determine equations for all pairs of categories. When $K = 3$, for example we have:

$$\begin{aligned} log\left(\frac{\pi_1}{\pi_2}\right) = log\left(\frac{\pi_1/\pi_3}{\pi_2/\pi_3}\right) &= log\left(\frac{\pi_1}{\pi_3}\right) - log\left(\frac{\pi_2}{\pi_3}\right) \\ &= (\alpha_1 + \beta_1 x) - (\alpha_2 + \beta_2 x) \\ &= (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x \end{aligned}$$

This equation has the form $\alpha + \beta x$ with intercept parameter $\alpha = (\alpha_1 - \alpha_2)$ and with slope parameter $\beta = (\beta_1 - \beta_2)$. With p explanatory variables the model extends to :

$$log\left(\frac{\pi_j}{\pi_K}\right) = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jp}x_p, j = 1, \ldots, K-1$$

The Bayesian multinomial specification requires prior distributions to be placed on all of the model parameters. Priors can be non-informative (i.e, diffuse) or informative. The weight of the informative provided by a prior depends on its certainty relative to the certainty in the likelihood.

For example, a Normal prior regression parameter of $\beta_1 \sim Normal(\mu = 6, \sigma^2 = 10)$ will have less influence than the prior $\beta_1 \sim Normal(\mu = 6, \sigma^2 = 2)$.Similarly, truncated distributions add increased influence, such as $\beta_1 \sim Normal(\mu = 6, \sigma^2 = 2), 0 < \beta_1 < 8$.

When non-iformative or ignorance priors are used,Bayesian models (especially for large data sets) result in parameter estimates that converge to those of maximum-likelihood estimates.Informative priors may be obtained from prior research results.For example, an analyst may know that a parameter θ is always positive and between 2 and 10 of the 95% confidence interval.

A Bayesian approach argues that this information is useful and should be included, and reflects the modelling procedure on updating the prior knowledge with the current information (new data - likelihood).

For example we took a dataset from UCLA (free in web) which contains variables on 200 students. The response variable is prog, which is the program type of its student. The predictor variables are social economic status (ses),a three-level categorical variable and writing score (write), which is a continuous variable. We assign the reference level of Y variable *prog* to be the "academic" category.

We fitted a multinomial regression in Rstan and the HMC output, gave us (see Appendix for Rstan code) for the "sesmiddle" of the *ses* explanatory variable $\beta = 0.54$ with $sd = 0.5$, in general category compared to the academic.

These results are suggesting that for one unit increase in social economic status level, the logit coefficient for 'general' relative to 'academic' will decrease by the amount of -0.54. In other words, if your social economic status level increases one unit (level), your chances of being in the academic category are higher compared to staying in general program type category.

Exponentiating the coefficients the interpretation changes to estimated odds scale. For example in the vocation row, exponentiating out the sesmiddle gave us 1.37, which means that keeping all other variables constant, if your social economic status level increases one unit, you are 1.37 times more likely to stay in the vocation program type category as compared to the

Table 3.6: Multinomial Regression Analysis

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
$\beta_{1,1}$	1.0	6263	2.9	1.2	0.5	2.9	5.3
$\beta_{2,1}$	1.0	5983	5.3	1.2	3.1	5.3	7.7
$\beta_{1,2}$	1.0	8646	-0.5	0.5	-1.4	-0.5	0.4
$\beta_{2,2}$	1.0	8316	0.3	0.5	-0.6	0.3	1.3
$\beta_{1,3}$	1.0	8750	-1.2	0.5	-2.2	-1.2	-0.1
$\beta_{2,3}$	1.0	8031	-1.0	0.6	-2.2	-1.0	0.2
$\beta_{1,4}$	1.0	6451	-0.1	0.0	-0.1	-0.1	-0.0
$\beta_{2,4}$	1.0	6332	-0.1	0.0	-0.2	-0.1	-0.1

academic program category (the risk is 37% higher).

3.7 Ordinal Models

In the previous section we have discussed the nominal scale of the response variable y with no ordinality. On this section we will discuss the form of modelling the variable y that its categories have ordinal responses.

We utilizing the category ordering by forming logits of cumulative probabilities,

$$P(Y \leq j \mid x) = \pi_1(x) + \cdots + \pi_j(x), j = 1, \ldots, J.$$

The cumulative logits are defined as:

$$logit\left(P(Y \leq j \mid x)\right) = log\left(\frac{P(Y \leq j \mid x)}{1 - P(Y \leq j \mid x)}\right)$$

$$= \log\left(\frac{\pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \cdots + \pi_J(x)}\right), j = 1. \ldots, J - 1.$$

Each cumulative logit uses all J response categories.

For a multinomial response variable Y with possible J ordered categorical outcomes and the associated p-dimensional vector of covariates x, the cumulative probability for Y on x is given by:

$$P(Y \leq j \mid x) = \frac{\exp(\alpha_j + x'\beta)}{1 + \exp(\alpha_j + x'\beta)}, j = 1, 2, \ldots, J - 1,$$

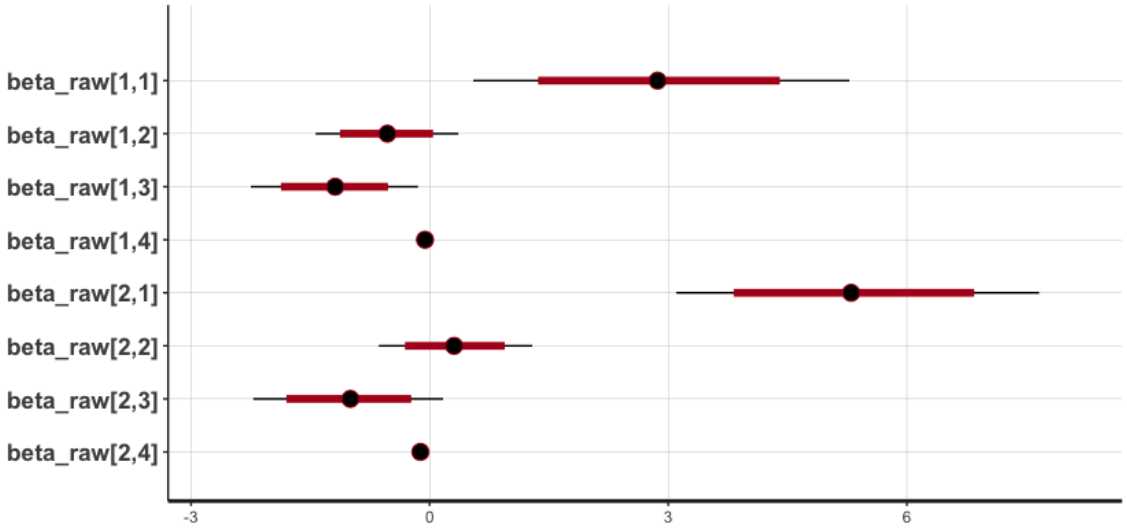

Figure 3.1: Plot of beta coefficients of Multinomial Regression

or the cumulative logit form as:

$$\log\left(\frac{P(Y \leq j \mid x)}{P(Y > j \mid x)}\right), j = 1, 2, \ldots, J-1,$$

where α_j is an unknown intercept parameter associated with the j th category and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the common vector of effect coefficients across the categories. The Proportional Odds Logistic Regression (POLR) models have the cumulative probabilities $P(Y \leq j)$ as described above,rather than the specific category probabilities $P(Y = j)$ as in the nominal logistic regression.

A model with a single $logit[P(Y \leq j)]$ is alone an ordinary logistic model model for binary response in which categories 1 to j form one outcome and categories $j + 1$ to J form the second. A model that simultaneously uses all $(J - 1)$ cumulative logits in a single parsimonious model is:

$$logit[P(Y \leq j \mid x)] = \alpha_j + \beta^T x, j = 1, \ldots, J-1.$$

Each logit has its own intercept α_j.The $\{\alpha_j\}$ intercepts are increasing in j,because $P(Y \leq j \mid x)$ increases in j for fixed x and the logit is an increasing function of $P(Y \leq j \mid x)$.

For fixed j,the response curve is a logistic regression curve for a binary response with outcomes $(Y \leq j)$ and $(Y > j)$. The curves for example of $J = 4$ the $j = 1, 2$ and 3 will have the same logistic S-curve and that because they share the same rate of increase or decrease but are horizontally displaced from each other.

The cumulative logit for two binary responses is:

$$logit[P(Y \leq j \mid x_1)] - logit[P(Y \leq j \mid x_2)]$$

$$= \log \left[\frac{P(Y \leq j \mid x_1)/P(Y > j \mid x_1)}{P(Y \leq j \mid x_2)/P(Y > j \mid x_2)} \right] = \beta^T (x_1 - x_2).$$

An odds ratio of cumulative probabilities is called a cumulative odds ratio. The odds of making response $Y \leq j$ at $x = x_1$ are $\exp^{[\beta^T (x_1 - x_2)]}$ times the odds at $x = x_2$. The log cumulative odds ration is proportional to the distance between x_1 and x_2. The same proportionality constant applies to each logit. Because of this property, the above cumulative logit for two binary responses, is called proportional odds model (McCullagh 1980).

Now consider a categorical outcome y that can take on the values (categories) $1, 2, \ldots, J$. The ordered logistic model can be written in two equivalent ways. First we express it as a series of logistic regressions:

$$\begin{aligned} P(Y > 1) &= logit^{-1}(X\beta) \\ P(Y > 2) &= logit^{-1}(X\beta - \alpha_2) \\ P(Y > 3) &= logit^{-1}(X\beta - \alpha_3) \\ &\cdots \\ P(Y > K - 1) &= logit^{-1}(X\beta - \alpha_{J-1}) \end{aligned}$$

The parameters α_j (which are called thresholds or cutpoints, for reasons which we shall explain shortly) are constrained to increase:

$$0 = \alpha_1 < \alpha_1 < \cdots < \alpha_{J-1},$$

because the probabilities are strictly decreasing (assuming that all J outcomes have nonzero probabilities of occurring). Since α_1 is defined to be 0, the model with J categories has $J-2$ free parameters α_j in addition to β.

This makes sense since $J = 2$ for the usual logistic regression, for which only β needs to be estimated. The cutpoints $\alpha_2, \ldots, \alpha_{J-1}$ can be estimated using maximum likelihood, simultaneously with the coefficients β. For some datasets, however, the parameters can be non identified, as with logistic regression for binary data.

Along with the already theory for cumulative logits there is a another representation of cumulative logits that motivates a continuous latent variable for proportional odds structure. These models have a latent continuous variable assumed to underlie Y that actuates the common effect β for different j categories in the proportional odds form of the model.

Let $Y^\star$ denote this underlying latent variable. Now let:

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$$

denote cutpoints of the continuous scale for $Y^\star$ such that the ordinal variable Y satisfies $y = j$ if $\alpha_{j-1} < y^\star \leq \alpha_j$.

Its cumulative distribution function is $G(y^\star - \eta)$,where values of $y^\star$ vary around a location parameter η (such as a mean) that depends on x through $\eta(x) = \beta^T x$. We observe Y in category j when the latent variable falls in the jth interval of values. Now, suppose the latent variable satisfies an ordinary linear model relating it to the explanatory variables,

$$Y^\star = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

where ϵ has some probability distribution with mean 0 and the same variance at all values of the explanatory variables. Then, we can tell that the observed ordinal categorical variable satisfies the model:

$$logit\left[P(Y \leq j)\right] = \alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p, j = 1, \ldots, J-1,$$

for a link function that depends on the distribution of ϵ

With logit link, this is the cumulative logit model of proportional odds form, except that the signs of the effects change. Because of this latent

variable model connection, some software fits the model with this parameterization.

One of the assumptions underlying ordinal logistic (and ordinal probit) regression is that the relationship between each pair of outcome groups is the same. In other words, ordinal logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc.

This is called the proportional odds assumption or the parallel regression assumption. Because the relationship between all pairs of groups is the same, there is only one set of coefficients. If this was not the case, we would need different sets of coefficients in the model to describe the relationship between each pair of outcome groups.

Thus, in order to assess the appropriateness of our model, we need to evaluate whether the proportional odds assumption is tenable. Statistical tests to do this are available in R packages.

A Bayesian method for modelling ordinal data for the cumulative logit models we ha the coefficients with Dirichlet distribution,

$$\theta \sim Dirichlet(\alpha_1, \ldots, \alpha_j)$$

$$p(\theta) = Dirichlet(\theta \mid \alpha_1, \ldots, \alpha_k), \alpha_j > 0; \alpha_0 = \sum_{j=1}^{K} \alpha_j$$

$$p(\theta \mid \alpha) \propto \prod_{j=1}^{K} \theta_j^{\alpha_j - 1},$$

where the distribution is restricted to nonnegative θ_j's with $\sum_{j=1}^{K} \theta_j = 1$. The Dirichlet distribution is a multivariate generalization of the beta distribution. Apart from conjugate, it is perhaps the easiest prior distribution to specify because the concentration parameters can be interpreted as prior counts (although they need not to be integers in Stan) of a multinomial random variable.

The Dirichlet distribution is used in Stan for an implicit prior on the cutpoints $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ in an ordinal regression model. More

specifically, the Dirichlet prior pertains to the prior probability of observing each category of the ordinal outcome when the predictors are at their sample means. Given these prior probabilities, it is straightforward to add them to form cumulative probabilities and then use an inverse cumulative distribution function transformation of the cumulative probabilities to define the cutpoints.

If a scalar is passed to the scale parameter α of the dirichlet function, then it is replicated to the appropriate length and the Dirichlet distribution is symmetric. If dirichlet's α is a vector and all elements are 1 , then the Dirichlet distribution is jointly uniform. If all concentration parameters are equal but greater than 1 then the prior mode is that the categories are equiprobable, and the larger the value of the identical dirichlet parameters, the more sharply peaked the distribution is at the mode. The elements in dirichlet parameter can also be given different values to represent that not all outcome categories are a priori equiprobable.

Implementing cumulative logit and proportional odds model in Stan is straightforward with the command *stan_polr* from *rstanarm* package and the priors on the coefficients are dirichlet as described above. Along with that with have to specify the R^2 prior which is the prior in the variance of the continuous variable $y^\star$.

We used the *housing* dataset from $MASS$ package in R which is a frequency table from a Copenhagen Housing conditions Survey. The dataset is a data frame of 72 rows and 5 variables.

1. Sat: (response variable) which is an ordinal categorical variable named and its levels are satisfaction of householders with their present housing circumstances, (High, Medium or Low).
2. $Infl$: Perceived degree of influence householders have on the management of the property (High, Medium, Low).
3. $Type$: Type of rental accommodation, (Tower, Atrium, Apartment, Terrace).
4. $Cont$: Contact residents are afforded with other residents, (Low, High).
5. $Freq$: Frequencies: the numbers of residents in each class.

The response variable is an ordinal categorical variable named Sat which is a satisfaction of householders with their present housing circumstances, (High, Medium or Low). The $Infl$ Perceived degree of influence householders have on the management of the property (High, Medium, Low).$Type$: type of rental accommodation, (Tower, Atrium, Apartment, Terrace).$Cont$: Contact residents are afforded with other residents, (Low, High).$Freq$:Frequencies: the numbers of residents in each class.

We fitted a proportional odds logistic regression model with all the explanatory variables included and additionally with weights the $Freq$ variable.The main effects cumulative logit model of proportional odds form estimates are β_1 for inflMedium 0.6 with standard deviation $sd = 0.1$ and for β_2 inflHigh 1.3 with standard deviation $sd = 0.1$,suggesting that the cumulative probability starting at the low satisfaction level is increasing as the influence degree of householders in the management of the property is increasing from medium to high. This can also be seen from the estimated odds of the exponentiated coefficients.

Table 3.7: Ordinal Regression Analysis

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
InflMedium	1.0	3989	0.6	0.1	0.4	0.6	0.8
InflHigh	1.0	3065	1.3	0.1	1.0	1.3	1.5
TypeAtrium	1.0	3632	-0.4	0.2	-0.7	-0.4	-0.1
ContHigh	1.0	4596	0.4	0.1	0.2	0.4	0.5
Low|Medium	1.0	3862	-0.5	0.1	-0.7	-0.5	-0.3
TypeApartment	1.0	4063	-0.6	0.1	-0.8	-0.6	-0.3
TypeTerrace	1.0	3808	-1.1	0.1	-1.4	-1.1	-0.8
log-posterior	1.0	896	-1754.0	2.6	-1760.2	-1753.6	-1749.7

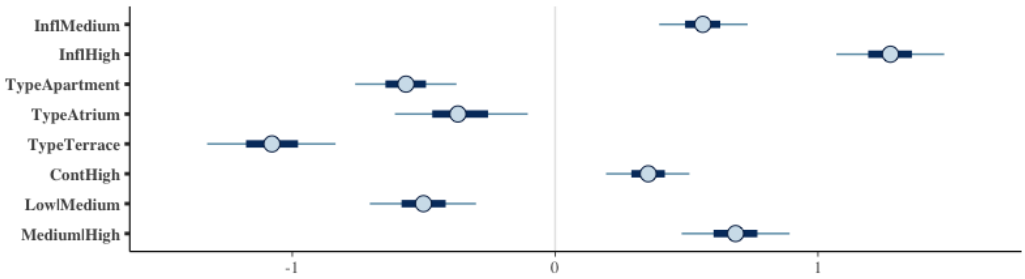

Figure 3.2: Plot of Beta Ordinal coefficients

Part IV

Penalised Likelihood Criteria

Chapter 4

Penalised Likelihood Criteria

4.1 Penalised Likelihood Criteria

In this section of chapter 4 we will introduce to the reader the methods of model comparison between models in the Bayesian framework. After fitting some models of interest we wish to choose the best fitting model that separates from the other.
The criteria of doing that in Bayesian framework is BIC, AIC, DIC and LOO. Each of them will be presented below and we will give their pathologies along with their capabilities. Finally we will compare different models according to the purpose of this comparison.

4.2 Bayes Information Criterion (BIC)

The Bayes information criterion (BIC) is based on the criterion originally introduced by Schwarz (1978), and is denoted by:

$$S_{01} = \log(f(y \mid \hat{\theta}_{m1}, m_1)) - \log(f(y \mid \hat{\theta}_{m0}, m_0)) - \frac{1}{2}(d_{m1} - d_{m0}) \log(n),$$

where n is the sample size.$\hat{\theta}_m$ are the maximum likelihood estimate of parameters θ_m of each model m and d_m is the dimension of θ_m or the number

of parameters of model m.
The main property of the Schwarz criterion S it that:

$$\frac{S_{01} - \log(B_{10})}{\log(B_{10})} \to 0,$$

when $n \to \infty$
and consequently can be an approximation of the log-Bayes factor as has been proposed by Kass and Raftery (1995). The BIC criterion is widely known by the expression:

$$BIC_{(m)} = D(\hat{\theta}_m, m) + d_m \log(n)$$

, where $D(\hat{\theta}_m, m)$ is the deviance measure of model m and equals to :

$$D(\hat{\theta}_m, m) = -2 \log f(y \mid \theta_m, m)$$

, where $\hat{\theta}_m$ is the posterior mean of the parameters involved in the model m. As can be seen BIC is a penalised deviance (or log-likelihood) measure with penalty equal to the $\log(n)$ for each parameter estimated by the model. If the BIC difference between two models is lower than 2, then we cannot discriminate between the two compared models.
BIC differences from 2-6,6-10 and higher than 10 express positive, strong, and very strong evidence in favour of the model with the lower BIC value.

4.3 Akaike Information Criterion (AIC)

The Akaike information criterion is named after the statistician Hirotugu Akaike (1973) is another statistical criterion for model validation and comparison with other fitted models and is given by:

$$AIC_m = D(\hat{\theta}_m, m) + 2d_m,$$

where $D(\hat{\theta}_m, m)$ is the same penalised log likelihood as described before in BIC subsection plus this time with penalty equal to 2 for each estimated parameter of model m.

As we saw BIC is an approximation of the log-Bayes factor proposed from Kass and Raftery (1995), AIC is an approximately unbiased estimator of the

expected Kullback-Leibler (KL) distance between true and estimated models and supports models that have predictive performance equivalent to the true performance.

Since AIC is one of the approximately unbiased estimators the KL distance, a wide variety of other estimators have been proposed in the literature, see Kuha (2003) for AIC and related methods.

4.4 Deviance Information Criterion (DIC)

Deviance Information Criterion (DIC) is a method of model comparison in Bayesian Statistics.Models that have been calculated via the Bayesian framework can be compared to each other with this criterion. DIC was first introduced by Spigelhalter et al.(2002).

Spiegelhalter proposed DIC as a method, or tool of model comparison which is a generalation of AIC that is based on the posterior distribution of the deviance statistic and is given by:

$$D(\theta) = -2\log f(y \mid \theta) + 2\log h(y)$$

where $f(y \mid \theta)$ is the likelihood function for the data vector y given the parameter vector θ , and $h(y)$ is some standardizing function of the data alone (which thus has no impact on model selection). In this approach, the fit of a model is summarized by the posterior expectation of the deviance,

$$\overline{D} = \mathbb{E}_{(\theta|y)}[D],$$

The complexity of the each individual model is given by the effective number of parameters which is denoted p_m, and can be given that:

$$p_m = \mathbb{E}_{\theta|y}[D] - D(\mathbb{E}_{\theta|y}[\theta])$$

or equivalently:

$$p_m = \overline{D}(\theta_m, m) - D(\overline{\theta}_m, m)$$

Finally the Deviance Information Criterion can be written as:

$$\begin{aligned} DIC(m) &= 2\overline{D}(\theta_m, m) - D(\overline{\theta}_m, m) \\ &= D(\overline{\theta}_m, m) + 2p_m. \end{aligned}$$

Smaller values of DIC indicating a better-fitting model.

Note that DIC is scale-free the choice of standardising function $h(y)$ is arbitrary. Thus values of DIC have no intrinsic meaning; as with AIC, only differences in DIC across models are meaningful, with differences of 3 to 5 normally being thought of as the smallest that are interesting.

Besides its generality, another attractive aspect of DIC is that it may be readily calculated during an MCMC run by monitoring both θ and $D(\theta)$, and at the end of the run simply taking the sample mean of the simulated values of D, minus the plug-in estimate of the deviance using the sample means of the simulated values of θ.

4.5 Other Information Criterion

A wide variety of penalized likelihood or deviance criteria is available in the statistical literature. Generally, most information criteria minimise the quantity:

$$IC_{(m)} \approx D(\hat{\theta}_m, m) + d_m F,$$

where F is the penalty that takes the deviance measure for each additional parameter added in the model m. For example the difference between AIC and BIC is the F type of penalty that differs from the one criterion to the other. AIC's $F = 2$,while BIC's $F = \log(n)$.

Comparing two models,for example, m_1,m_2 we select the model with the smallest value of IC, even if this criterion is the AIC or BIC. Consequently we can use the difference $DIFF - IC_{01}$ between the IC values of these models m_0 and m_1:

$$DIFF - IC_{01} = D(\hat{\theta}_{m0}, m_0) - D(\hat{\theta}_{m1}, m_1) - d_{m1} - d(m_0) \times F.$$

Here we can denote : $\psi = (d_{m1} - d(m_0)) \times F$, as a more complicated penalty function and without loss of generality, we assume that $d_{m0} < d_{m1}$, and therefore the $DIFF - IC_{01} < 0$, we model m_0 and id $DIFF - IC_{01} > 0$, we select model m_1.

Shao (1993) showed that the two information criteria can be splitted into two categories:

1. criteria that are asymptotically valid under the assumption that a true model exists and
2. criteria that are asymptotically valid under the assumption that a true model does exists.

Generally, information criteria with the penalty F fixed as $n \to \infty$ (such as AIC) and criteria with $F \to \infty$ as $n \to \infty$ (such as BIC) are two different categories of penalised likelihood criteria, usually referred as AIC-like and BIC-like criteria.

For example purposes we have created a dataset with 1000 random numbers from Gamma distribution $y \sim Gamma(2, 4)$ and we have fitted them to two different models:

1. a gamma model and
2. a lognormal model.

We have extracted the posterior samples to calculate AIC and DIC. The results are given in table 4.1:

Table 4.1: Criteria Comparison for the Gamma & Lognormal Models

Information Criteria			
Models	AIC	BIC	DIC
Gamma	377.48	387.30	187.7
Lognormal	1363.01	1372.82	679.9

Continuing our model comparison we generated again 100 random numbers from gamma distribution and calculated the AIC, BIC and DIC (out of sample data) and the results are given in the table 4.2:

Table 4.2: Criteria Comparison for Out of Sample data for Gamma & Log-normal Models

Information Criteria			
Models	AIC	BIC	DIC
Gamma	51.48	52.87	13.28
Lognormal	124.70	143.63	69.82

So again the Gamma model has a better (lower) AIC , BIC and DIC than Log-Normal.

For example purposes we will use, for now, the *LOO* package in R, that implements the (Leave One Out Criterion) and we will use it for extracting the maximum a posteriori likelihood for the calculations of AIC and BIC.

We illustrate an example of model comparison with *LOO* using the *birthwt* dataset from the MASS package. *Birthwt* dataset is a data frame that has 189 rows and 10 variables (columns). The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

The model in logistic regression form has the following variables:

1. *low*: (response variable) which is an indicator of birth weight less than 2.5 kg and if a baby is born under 2.5 kg is 1 and 0 otherwise,
2. *race*: of the mother (1 = white, 2 = black, 3 = other),
3. *lwt*: which is a covariate is the mother's weight in pounds at last menstrual period,
4. *smoke*: which is if mother was smoking during her pregnancy,
5. *ptl*: is the number of previous premature labours,
6. *ht*: if the mother had a history of hypertension,
7. *ui*: which is the presence of uterine irritability,
8. ftv: number of physician visits during the first trimester.

We fitted the following 9 models:

$$
\begin{aligned}
m_0 &= low \sim 1, \\
m_1 &= low \sim race, \\
m_2 &= low \sim race + lwt, \\
m_3 &= low \sim race + lwt + age, \\
m_4 &= low \sim race + lwt + age + smoke, \\
m_5 &= low \sim race + lwt + age + smoke + ptl, \\
m_6 &= low \sim race + lwt + age + smoke + ptl + ht, \\
m_7 &= low \sim race + lwt + age + smoke + ptl + ht + ui, \\
m_8 &= low \sim race + lwt + age + smoke + ptl + ht + ui + ftc
\end{aligned}
$$

Comparing the above models in the minimum, mean and median from the point-wise log-likelihood from the posterior and we obtained the table below.

Table 4.3: Information Criteria comparison for the 8 different models

Models	dm	minAIC	minBIC	meanAIC	meanBIC	medianAIC	medianBIC
m_0	2	961.00	967.48	946.68	953.16	946.16	952.65
m_1	2	956.24	962.72	937.16	943.65	936.40	942.89
m_2	3	949.01	958.73	922.23	931.95	921.38	931.11
m_3	4	942.50	955.47	924.28	937.25	923.77	936.73
m_4	5	922.60	938.80	895.97	912.18	895.52	911.73
m_5	6	917.07	936.52	885.92	905.37	885.41	904.86
m_6	7	898.82	932.00	866.80	899.97	866.12	899.30
m_7	8	907.96	933.89	863.33	889.26	862.54	888.48
m_8	9	904.27	933.45	869.13	898.31	868.49	897.67

As we can clearly identify from this comparison that model m_8 has the lower IC_m criteria than all models apart from model m_7 which is slightly close to m_8.

4.6 Widely Applicable Information Criterion (WAIC - LOO)

In the previous subsections of information criteria we introduced criteria that hold in any type of bayesian model and are well known and used through out the years of MCMC development, apart from WAIC which is a product of Columbia's University of Stan's language HMC calculation.

After fitting a Stan model for purposes of model comparison, selection, or averaging (Geisser and Eddy, 1979, Hoeting et al., 1999, Vehtari and Lampinen, 2002, Ando and Tsay, 2010, Vehtari and Ojanen, 2012) proposed the WAIC criterion which is asymptotically the same as LOO-CV (leave one out cross validation) criterion-method.

Cross-validation and information criteria are two approaches of estimating out-of-sample predictive accuracy using within-sample fits (Akaike, 1973, Stone, 1977).

In this thesis as have presented by Gelman, Vehtari and Gabry (2017) we consider computations using the log-likelihood evaluated at the usual posterior simulations of the parameters. Computation time for the predictive accuracy measures should be negligible compared to the cost of fitting the model and obtaining posterior draws in the first place.

Widely applicable information criterion (WAIC) and Leave-one-out cross-validation (LOO) are methods for estimating point-wise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values.

LOO and WAIC have various advantages over simpler estimates of predictive error such as AIC and DIC but are less used in practice because they involve additional computational steps.

Here we present the LOO and WAIC information criteria that can be performed using existing simulation draws which have been presented in Gelman, Vehtari and Gabry (2017). They introduced an efficient computation of LOO using Pareto-smoothed importance sampling (PSIS), a new procedure for regularising importance weights. Although WAIC is asymptotically equal

to LOO.

LOO and WAIC were developed for Stan's modelling calculations and are capable of obtain approximate standard errors for estimated predictive errors and for comparing of predictive errors between two models. We implement the computations in an R package called loo and demonstrate using models fit with the Bayesian inference package RStan.

Consider data $y_1, \ldots, y_n$, modeled as independent given parameters θ, thus:

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta).$$

This formulation also encompasses latent variable models with

$$p(y_i \mid f_i, \theta),$$

where f_i are latent variables.

Also suppose we have a prior distribution $p(\theta)$, thus yielding a posterior distribution $p(\theta \mid y)$ and a posterior predictive distribution

$$p(\hat{y} \mid y) = \int p(\hat{y}_i \mid \theta) p(\theta \mid y) d\theta.$$

To maintain comparability with the given dataset and to get easier interpretation of the differences in scale of effective number of parameters, we define a measure of predictive accuracy for the n data points taken one at a time: (epld = expected log pointwise predictive for a new dataset)

$$elpd = \sum_{i=1}^{n} \int p_t(\hat{y}_i) logp(\hat{y}_i \mid y) d\hat{y}_i,$$

where $p_t(\hat{y}_i)$ is the distribution representing the true data-generating process for $\hat{y}_i$.The $p_t(\hat{y}_i)$'s are unknown and we will use cross-validation or WAIC to approximate the *elpd*.

The Bayesian LOO (Leave-one-out) estimate of out-of-sample predictive fit is:

$$elpd_{loo} = \sum_{i=1}^{n} logp(y_i \mid y_{-i}),$$

where

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta) p(\theta \mid y_{-i}) d\theta$$

is the leave-one-out predictive density given the data without the ith data point.

As noted by Gelfand, Dey, and Chang (1992), if the n points are conditionally independent in the data model we can then evaluate the previous equation with draws θ^s from the full posterior $p(\theta \mid y)$ using importance ratios:

$$r_i^s = \frac{1}{p(y_i \mid \theta^\star)} \propto \frac{p(\theta^\star \mid y_{-i})}{p(\theta^s \mid y)}$$

to get the importance sampling leave-one-out (IS-LOO) predictive distribution,

$$p(\hat{y}_i \mid y_{-i}) \approx \frac{\sum_{s=1}^S r_i^s p(\hat{y}_i \mid \theta^s)}{\sum_{s=1}^S r_i^s}.$$

Evaluating the LOO log predictive density at the held-out data point y, we get

$$p(y_i \mid y_{-i}) \approx \frac{1}{\frac{1}{S}\sum_{s=1}^S \frac{1}{p(y_i|\theta^s)}}.$$

However, the posterior $p(\theta \mid y)$ is likely to have a smaller variance and thinner tails than the leave-one out distributions $p(\theta \mid y_{-i})$, and thus a direct use of the latter equation induces instability because the importance ratios can have high or infinite variance.

As noted above, the distribution of the importance weights used in LOO may have a long right tail. We use the empirical Bayes estimate of Zhang and Stephens (2009) to fit a generalized Pareto distribution to the tail (20% largest importance ratios).

By examining the shape parameter K of the fitted Pareto distribution, we are able to obtain sample based estimates of the existence of the moments. When the tail of the weight distribution is long, a direct use of importance sampling is sensitive to one or few largest values. By fitting a generalized Pareto distribution to the upper tail of the importance weights, we smooth these values.

WAIC (Watanabe, 2010) is an alternative approach to estimate the expected log point-wise predictive density and is defined as:

$$\widehat{elpd_{waic}} = \widehat{lpd} - \hat{p}_{waic},$$

where $\hat{p}_{waic}$ is the estimated effective number of parameters and computed based on:

$$p_{waic} = \sum_{i=1}^{n} Var_{post}(logp(y_i \mid \theta)),$$

which we can calculate using the posterior variance of the log predictive density for each data point y_i, that is,

$$V_{s=1}^{S} logp(y_i \mid \theta^s),$$

where $V_{s=1}^{S}$ represents the sample variance,

$$V_{s=1}^{S}\alpha_s = \frac{1}{S-1}\sum_{s=1}^{S}(\alpha_s - \hat{\alpha})^2.$$

Summing over all the data points y_i gives a simulation-estimated effective number of parameters,

$$\hat{p}_{waic} = \sum_{i=1}^{n} V_{s=1}^{S}(logp(y_i \mid \theta^s)).$$

For DIC, there is a similar variance-based computation of the number of parameters that is unreliable, but the WAIC version is more stable because it computes the variance separately for each data point and then takes the sum (the summing yields stability).

The effective number of parameters $\hat{p_{waic}}$ can be used as measure of complexity of the model, but it should not be overinterpreted, as the original goal is to estimate the difference *lpd* and *elpd*.

When comparing two fitted models, we can estimate the difference in their expected predictive accuracy by the difference in $elpd_{loo}$ or $elpd_{waic}$. When using LOO model comparison, software will return a matrix that will have one row per model and several columns of estimates.

The values in the difference in *elpd* and its standard error columns of the returned matrix are computed by making pairwise comparisons between each model and the model with the largest *elpd* (the model in the first row).

To compute the standard error of the difference in *elpds* which should not be expected to equal the difference of the standard errors. LOO uses a paired estimate to take advantage of the fact that the same set of N data points was used to fit both models.

These calculations should be most useful when N is large, because then non-normality of the distribution is not such an issue when estimating the uncertainty in these sums.

These standard errors, for all their flaws, should give a better sense of uncertainty than what is obtained using the current standard approach of comparing differences of deviances to a chi-squared distribution, a practice derived for Gaussian linear models or asymptotically, and which only applies to nested models in any case.

For example purpose we will revisit the birth weight dataset of the previous chapter and we compare model

$$m_1 = low \sim race + lwt$$

with a second model

$$m_2 = low \sim race + \log(lwt).$$

Finally we compare the expected predictive accuracy of those two models.

The first row of the *LOO* output will be always 0 because the model in the first row is compared with itself. The row of interest is the second row. When the difference, $elpd_{diff}$, is positive then the expected predictive accuracy for the second model is higher. A negative $elpd_{diff}$ favours the first model.

In our case the model 2 with the $log(lwt)$ variable has slightly smaller

expected predictive accuracy than the first model because the model 2 in the second row has $elpd_{diff} = -0.7$.

Table 4.4: LOO comparison for the 2 different models

Models	$ELPD_{diff}$	SE_{diff}
$m_1 = low \sim race + lwt$	0.0	0.0
$m_2 = low \sim race + \log(lwt)$	-0.7	1.6

Additionally we fitted 9 models the same as for variable selection in section 4.5 and we are examining their predictive accuracies in LOO-CV methodology. R reported us the following table for leave one out cross validation:

Table 4.5: LOO comparison for the 9 different models

Models	$ELPD_{diff}$	SE_{diff}
m_7	0.0	0.0
m_6	-0.1	1.4
m_8	-1.1	0.2
m_5	-2.3	2.3
m_4	-3.1	3.2
m_2	-6.1	4.1
m_3	-6.5	4.1
m_1	-7.8	4.5
m_0	-8.5	4.7

The models with the better predictive accuracy are models m_7 and models m_6. Recall that from the AIC, BIC methodology of stepwise forward variable selection the best model was model m_8 but here its predictive accuracy is ranked third compared to the 7 other models.

Additionally we revisit the Liverpool's example from chapter 1. Remember that the data y where Poisson and we checked three models. Here in the model comparison with LOO-CV method we compare the log Poisson-Gamma versus the log Poisson-Normal.

The LOO comparison in our case model2 with poisson-normal has a better expected predictive accuracy and it's better that the first with poisson-gamma model which its $elpd = -0.5$.

Table 4.6: LOO & min(AIC) & min(BIC) comparison for the 2 different models

Models	$ELPD_{diff}$	SE_{diff}	minAIC	minBIC
$m_1(NormalPrior)$	0.0	0.0	-206.58	203.31
$m_2(GammaPrior)$	-0.5	0.0	-213.63	-210.36

Additionally even the AIC and BIC agree with the LOO-CV criteria. Here as we have already presented AIC and BIC have been calculated from the point-wise log-likehihood maximum a posteriori and took the minimum value.

Part V

Matched Pairs Models

Chapter 5

Matched Pairs Models

5.1 Bayesian McNemar Test

The McNemar test is a test on a 2×2 classification table when the two classification factors are dependent, or when you want to test the difference between paired proportions, e.g. in studies in which patients serve as their own control, or in studies with "before and after" design.

Correlated proportions are usually expressed in the form of a 2×2 contingency table and their standard treatment consists of testing the null hypothesis of equality of proportions and evaluating confidence intervals for their difference.

The concordant pairs play no role as one would expect since the off-diagonal cell counts alone are sufficient for the difference of the two marginal proportions. Consider the matched-pairs of table 5.1:

Pairs with the same response from cases and controls (Yes-Yes and No-No) are called concordant pairs and are in the main-diagonal (n_{11}, n_{22}). Pairs with different responses (Yes-No and No-Yes) are call discordant pairs and appear in the off-diagonal. We assume that the sampling scheme is multinomial, so

Table 5.1: Table of Matched Pairs Frequencies

Controls			
Cases	Yes	No	Total
Yes	n_{11}	n_{12}	$n_{1.}$
No	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	1

that:

$$n = (n_{11}, n_{12}, n_{21}, n_{22}) \sim Multinomial(N, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}), \sum_{i=1}^{I} \sum_{j=1}^{j} \Pi ij = 1$$

In standard contingency tables analysis the cell probabilities $\pi_{ij}(i, j = 1, 2)$ are estimated by the corresponding sample proportions p_{ij} . Under the symmetry hypothesis, the off-diagonal probabilities are estimated by

$$(p_{12} + p_{21})/2$$

and the correlated proportions $\pi_{1.}$ and $\pi_{.1}$ by

$$p_{11} + (p_{12} + p_{21})/2.$$

Bayesian conjugate analysis proceeds by imposing a Dirichlet prior on the vector of probability parameters:

$$(\pi_{11}, \pi_{12}, \pi_{21}) \sim Dirichlet(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}), \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22} > 0,$$

resulting to a posterior probability of:

$$(\pi_{11}, \pi_{12}, \pi_{21})|n \sim Dirichlet(n_{11} + \alpha_{11}, n_{12} + \alpha_{12}, n_{21} + \alpha_{21}, n_{22} + \alpha_{22})$$

The hypothesis of equality of two correlated proportions ($\pi_{1.} = \pi_{.1}$) can be expressed as:

$$H_0 : \pi_{12} = \pi_{21}$$

and the alternative as:

$$H_0 : \pi_{12} \neq \pi_{21}$$

Delaportas, Kateri and Papaioannou (2001) used Bayes Factors this hypothesis and under this approach and multinomial sampling, with a Dirichlet prior the Bayes Factor is given by:

$$BF_(H_0, H_1) = \frac{\Gamma(\alpha_{12})\Gamma(\alpha_{21})\Gamma(n + \alpha_{12} + \alpha_{22})}{2^n\Gamma(\alpha_{12} + \alpha_{21})\Gamma(n_{12} + \alpha_{12})\Gamma(n_{21} + \alpha_{21})}$$

The above Bayes Factor is equivalent to Mc Nemar's test because the main-diagonal cells do not affect the prior parameters of Dirichlet α_{11}, α_{22}. As in Mc Nemar's test so in this Bayes Factor the total sample size N and the main-diagonal cells do not play any significant role in the calculation of the difference $\pi_{i.} - \pi_{.i}$.

Let set

$$\pi^{\star}_{12} = \pi_{12}/(\pi_{12} + \pi_{21})$$

which is the conditional probability that an observation will fall into a cell (1,2) given that it will fall in the off diagonal cells. Now instead of testing

$$H_0 : \pi_{12} = \pi_{21}$$

we can test

$$H_0 : \pi^{\star}_{12} = \frac{1}{2}$$

versus

$$H_1 : \pi^{\star}_{12} \neq \frac{1}{2}$$

. From all the cell sizes we have $n^{\star} = (n_{12}, n_{21}, n_{11} + n_{22})$ who follow a trinomial distribution with Dirichlet prior again on the vector of cell probabilities. The Dirichlet prior on these probabilities are $\alpha_{12}, \alpha_{21}, \alpha_1 = \alpha_{11} + \alpha_{22}$ in respect.

A well known test, conditional on the sum of the off-diagonal frequencies n, is based on the fact that n_{12} is distributed as binomial with parameters $(n, \pi^{\star}_{12})$. It is straightforward to see that when

$$(\pi_{11}, \pi_{12}, \pi_{21}) \sim Dirichlet(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}), \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22} > 0,$$

holds then the prior on $\pi^{\star}_{12}$ under H_1 is a Beta distribution with parameters α_{12} and α_{21}, denoted $\sim Beta(\alpha_{12}, \alpha_{21})$. In this context, the Bayes Factor

$BF_{(H0,H1)}$ is equal to the Bayes Factor:

$$BF_{01} = \frac{\Gamma(n+2)}{2^n \Gamma(n_{12}+1)\Gamma(n_{21}+1)}.$$

Implementing the Bayesian McNmear test in R with the assistance of Stan we will use a dataset from Agresti's (2001) categorical data analysis Chapter 11. Data came from the General Social Survey, examining the shift from Republicans to Democrats, in presidential votes between the elections of 2004 and the elections of 2008.

The scope of this analysis with McNemar test is to check if there is a substantial shift in the Democrat direction against the Republican direction.Data are presented below:

Table 5.2: Data from Agresti (2001)

2008 Elections			
2004 Elections	Democrat	Republican	Total
Democrat	175	16	191
Republican	54	188	242
Total	229	204	433

According to Delaportas & Kateri paper we will test the proportion:

$$\theta = n_{12}/(n_{12}+n_{21}) = 1/2$$

The H_0 of the Bayes factor in the numerator will evaluate θ which is:

$$H_0 : y \sim Binomial(n, \theta_1)$$

and H_1.

Here the likelihood is being calculated with the frequentist approach in the logarithmic scale. In the other hand in the denominator we have:

$$H_1 : y \sim Binomial(n_2, \theta_2),$$

where the parameter θ follows the Beta Distribution:

$$\theta_2 \sim Beta(1,1)$$

The Bayes Factor equals to :

$$B_{H_0,H_1} = \frac{p(D \mid H_o)}{p(D \mid H_1)}$$

The marginal likelihood have been evaluated with the bridge-sampling method and R reported the following ratio:

$$B_{H_0,H_1} = \frac{-13.07323}{-39.71} = 26.637$$

which this value indicates is a very strong evidence against H_0. So there is a very strong evidence against H_0 thus, we conclude that there is a swift to the democratic part on 2008 elections from 2004.

Here we must note that bridge-sampling method needs a very large number of iterations in sampling method to successfully evaluate the marginal likelihoods.

5.2 Symmetry, Quasi-Symmetry and Marginal Homogeneity Models

The special case of square $I \times J$ contingency table with commensurable classification variables occurs often in social sciences applications, in psychology and sociology and among other fields.

Characteristic of such cases refer to treatments' comparison or "before–after" comparisons applied on the same subjects, cross-classification of responses in matched pairs designs, problems of rater agreement, social mobility tables , or models of preference in opinion surveys.

Under this statistical frame, our focus lies on the off-diagonal cells and the models of symmetry and marginal homogeneity consist of the starting or reference point.

If symmetry is not significant or important for scientific calculating processes , which is usually the case, there is need to consider special models of asymmetry that measure departures from symmetry toward certain direction.

The standard hypothesis of symmetry is defined as :

$$H_0 : \pi_{ij} = \pi_{ji}, i > j, i, j = 1, \ldots, I$$

and when $\pi_{ij} > 0$ symmetry is a logistic regression model with the form of:

$$\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = 0$$

for all $i < j$. When the marginal distributions differ substantially, the symmetry model fits poorly. A generalized model that can accommodate marginal heterogeneity is the quasi-symmetry model:

$$\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \beta_i - \beta_j$$

One parameter is unnecessary, and we set $\beta_1 = 0$ or $\beta c = 0$. The higher the value of $\hat{\beta}_i - \hat{\beta}_j$, relatively more observations fall in the cell in row i and column j than in the cell in row j and column i.

Equivalently in terms of loglinear model with expected cell frequencies the null hypothesis and the symmetry model have the form of:

$$H_0 : \mu_{ij} = \mu_{ji}, i > j, i, j = 1, \ldots, I$$

$$log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

where all $\lambda_{ij}^{XY} = \lambda_{ji}^{XY}$

The main distinction for symmetry models against other models for contingency tables is that they model the off diagonal cells and not the main diagonal.

The latter are kept fixed. Main diagonal for count symmetry models has perfect fit and maximum likelihood for symmetry is:

$$\hat{\mu}_{ij} = \frac{n_{ij} + n_{ji}}{2}, i, j = 1, \ldots, I$$

The hypothesis of marginal homogeneity is:

$$H_0 : \pi_{i+} = \pi_{+i}, i = 1, \ldots, I,$$

states that the marginal distributions of a square contingency table are equal. As we described previously for quasi-symmetry one parameter is redundant, due to $\sum_{i,j} \pi_{ij} = 1$. For an $I \times I$ table, complete symmetry implies marginal homogeneity, while for the special case of 2×2 tables, models symmetry and marginal homogeneity are equivalent and tested by the McNemar test.

The tests proposed for marginal homogeneity are asymptotically chi-squared distributed with $df_{(MH)} = I - 1$. Marginal homogeneity is both equivalent to a loglinear model. However, quasi-symmetry is a useful for studying marginal homogeneity. Caussinus (1966) showed that symmetry is equivalent to quasi-symmetry and marginal homogeneity holding simultaneously.

Equivalently to McNemar's test the distributional sampling we may assume that is multinomial because symmetry and quasi-symmetry models are equal to McNemar test. The prior distribution then may also be Dirichlet for the cell parameters that we want to test.

Here we will give a more precise and specific form of the Multinomial - Dirichlet conjugate analysis. Results for the binomial with beta prior distribution generalize to the multinomial with a Dirichlet prior (Lindley 1964, Good 1965).

With c categories, suppose cell counts $(n_1, \ldots, n_c)$ have a multinomial distribution with $n = \sum_{i=1}^{I} n_i$ and parameters $\pi = (\pi_1, \ldots, \pi_c)'$. Let $\{p_i = n_i/n\}$ be the sample proportions. The likelihood is proportional to:

$$\prod_{i=1}^{c} \pi_c^{n_i}$$

The conjugate density is the Dirichlet, expressed in terms of gamma functions as:

$$g(\pi) = \frac{\Gamma(\sum \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{c} \pi_c^{\alpha_i - 1}$$

for $0 < \pi_i < 1$ all i, $\sum_i \pi_i = 1$, where $\{\alpha_i > 0\}$.Let

$$K = \sum_{i=1}^{c} \alpha_i.$$

The Dirichlet has

$$\mathbb{E}(\pi_i) = \alpha_i / K$$

and

$$Var_{(\pi_i)} = \alpha_i(K - \alpha_i) / \left[K^2(K+1)\right].$$

The posterior density is also Dirichlet, with parameters $\{n_i + \alpha_i\}$, so the posterior mean is:

$$\mathbb{E}(\pi_i \mid n_1, \ldots, n_c) = n_i + \frac{\alpha_1}{(n+K)}.$$

Let

$$\gamma_i = \mathbb{E}(\pi_i) = \alpha_i / K.$$

This Bayesian estimator equals the weighted average:

$$\frac{n}{(n+c)} p_i + \frac{K}{(n+K)} \gamma_i,$$

which is the sample proportion when the prior information corresponds to K trials with α_i outcomes of type $i, i = 1, \ldots, c$.

Vounatsou and Smith (1996) analysed certain structured contingency tables, including symmetry, quasi-symmetry and quasi-independence models for square tables and for triangular tables that result when the category corresponding to the (i, j) cell is indistinguishable from that of the (j, i) cell (a case also studied by Altham 1975).

They assessed goodness of fit using distance measures and by comparing sample predictive distributions of counts to corresponding observed values for malaria disease.

5.3 Kappa Cohen's Coefficient of Agreement

An important statistical inference problem in a range of physical, biological, behavioural, and social sciences is to decide how well one decision-making method agrees with another.

An interesting special case considers only binary decisions, and views one of the decision-making methods as giving objectively true decisions to which the other aspires.

This problem occurs often in medicine, when cheap or easily administered methods for diagnosis are evaluated in terms of how well they agree with a more expensive or complicated "gold standard" method.

For this problem, when both decision-making methods make n independent assessments, the data y take the form of four counts: a observations where both methods decide "one", b observations where the first rater decides "one" but the second decides "zero", c observations where the first rater decides "zero" but the second decides "one", and d observations where both raters decide "zero," with $n = a + b + c + d$.

Cohen's (1960) kappa statistic estimates the level of observed agreement

$$\pi_0 = \frac{a+d}{n}$$

relative to the agreement that would be expected by chance alone which is the overall probability for the first rater to decide "one", times the overall probability for the second rater to decide "one", and added to this the overall probability for the second rater to decide "zero", times the overall probability for the first rater to decide "zero".

$$\pi_e = \frac{(a+b)(a+c)+(b+d)(c+d)}{n^2}$$

and is given by

$$\kappa = \frac{\pi_0 - \pi_e}{1 - \pi_e}$$

Kappa lies on a scale of -1 to +1, with values below 0.4 often interpreted as "poor" agreement beyond chance, values between 0.4 and 0.75 interpreted as

"fair to good" agreement beyond chance, and values above 0.75 interpreted as "excellent" agreement beyond chance. The key insight of kappa as a measure of agreement is its correction for chance agreement.

The Bayesian approach of Kappa coefficient is that we calculate the latent variables of this 2×2 table, which are α , β and γ. The rate α is the rate at which first rater decides "one". This means $(1 - \alpha)$ is the rate at which the first rater decides "zero". The rate β is the rate at which the second rater decides "one" when the first rater also decides "one".

The rate γ is the rate at which the second rater decides "zero" when the first rater decides "zero." The best way to interpret β and γ is that they are the rate of agreement of the second rater with the first rater, for the "one" and "zero" decisions respectively.

Using the rates α , β and γ, it is possible to calculate the probabilities that both raters will decide "one":

$$\pi_a = \alpha\beta,$$

that the first rater will decide "one" but the second will decide "zero":

$$\pi_b = \alpha(1 - \beta),$$

the first will decide "zero" but the second will decide "one":

$$\pi_c = (1 - \alpha)(1 - \gamma),$$

and finally that both raters will decide "zero":

$$\pi_d = (1 - \alpha)\gamma.$$

These probabilities, in turn, describe how the observed data, y, made up of the counts a, b, c, and d, are generated. They come from a Multinomial distribution with n trials, where on each trial there is a π_α probability of generating an a count, π_β probability for b count, and so on.

Now that we have defined the probabilities of each cell in a 2×2 table we have to define the variables that measure the rate of agreement. The first

variable is the variable ξ that measures the rate of agreement between the two raters and is given by:

$$\xi = \alpha\beta + (1-\alpha)\gamma.$$

The second variable ψ is the variable that measures the rate of agreement by chance is given by:

$$\psi = (\pi_\alpha + \pi_b)(\pi_\alpha + \pi_c)(\pi_b + \pi_d)(\pi_c + \pi_d)$$

,and could be expressed in terms of α , β and γ.Finally the κ is the chance - corrected measure of agreement on the -1 to +1 scale, given by:

$$\kappa = \frac{\xi - \psi}{(1-\psi)}.$$

A diet questionnaire was mailed to 537 female American nurses on two separate occasions several months apart. The questions asked included the quantities eaten of more than 100 separate food items. The data from the two surveys.Instead we focus on the percentage of women with concordant responses in the two surveys.

We want to compare the observed concordance rate p_0 with the expected concordance rate p_e assuming the responses of the women in the two surveys were statistically independent.

Suppose there are c response categories and the probability of response in the $i-$th category is α_i, for the first survey and b_i for the second survey.

These probabilities can be estimated from the row and column margins of the following contingency table (Table 5.3).

Table 5.3: Nutrition Data Set

Survey 1	$\leq$ 1 serving/week	$>$ 1 serving/week
$\leq$ 1 serving/week	136	92
$>$ 1serving/week	69	240

The expected concordance rate (p_e) if the survey responses are independent is $\sum_{\alpha_i\beta_i}$.

Implementing the Bayesian framework of Kappa coefficient of agreement as has been described above with have to assign prior distribution to each parameter. Priors to α , β and γ will be $Beta(1,1)$ and Stan after 4000 iterations of Hamiltonian Monte Carlo Simulation reported $\kappa = 0.38$. So we conclude that is a good enough agreement between the two methods of medical assessment. Contrary to the frequentist approach that R reported $\kappa = 0.37$.

5.4 Bayesian Bradley-Terry Model

Bradley and Terry model for paired preferences were introduced by Bradley and Terry (1952) and earlier discussed by Zermelo (1929). The observed data is the outcome of matches between players or teams and we model the matches outcomes.

We will suppose that there are K players. Each contestant will have an ability $\alpha_k \in \mathbb{R}$. The probability that contestant i will beat contestant j is given by :

$$\Pr[i \text{ beats } j] \;=\; \text{logit}^{-1}(\alpha_i - \alpha_j).$$

The log odds function takes the $(0,1)$ initial values of the observations and extend them to $-\infty, \infty$:

$$\text{logit} : (0,1) \to (-\infty, \infty)$$

which is defined by

$$\text{logit}(u) = \log\left(\frac{u}{1-u}\right).$$

Its inverse logit, compresses the log odds to the probability space for interpretation,

$$\text{logit}^{-1} : (-\infty, \infty) \to (0,1),$$

and this inverse logit is given by:

$$\text{logit}^{-1}(v) = \frac{1}{1+\exp(-v)} = \frac{\exp(v)}{1+\exp(v)}.$$

The logistic distribution is the Bernoulli distribution with a parameter on the logit (log odds) scale, where for $y \in \{0, 1\}$ and $\theta \in (0, 1)$,

$$\mathsf{Bernoulli}(y \mid \theta) \;=\; \begin{cases} \theta & \text{if } \; y = 1 \\ 1 - \theta & \text{if } \; y = 0. \end{cases}$$

and for $\alpha \in (-\infty, \infty)$,

$$\mathsf{BernolliLogit}(y \mid \alpha) \;=\; \mathsf{Bernoulli}(y \mid \text{logit}^{-1}(\alpha)).$$

The likelihood is given by:

$$p(y \mid \alpha) = \prod_{n=1}^{N} \mathsf{Bernoulli}\Big(y_n \;\Big|\; \text{logit}^{-1}\big(\alpha_{\text{team1}[n]} - \alpha_{\text{team0}[n]}\big)\Big)$$

All that we need to do, to fit the data with Stan, is pack the data into a list, compile the model and then find the maximum likelihood estimate θ^*, that is, the estimate for the parameter values that maximizes the probability of the match outcomes that were observed.

Now we present the data as they will be passed in Stan (just a reminder that in Stan you cannot insert string variables as factors. Factors might be passed only as integers)

Table 5.4: Data presentation for Bradley Terry model

n	$team^0$	$team^1$	y
1	1	2	0
2	2	1	1
3	1	2	0
4	2	1	1
5	1	2	1
6	2	1	1

The first column, labeled n is the match index. With N matches, $n \in 1, 2, \ldots, N$. The second two columns indicate which teams participated in

the match. The last column is the result $y_n \in \{0, 1\}$, indicating which team won the match.

For example, the fourth row ($n = 4$) records a match between team 2 ($\text{team}^0 = 2$) and team 1 ($\text{team}^1 = 1$) in which team 1 won ($y = 1$).

With teams abilities α_k centered around zero and the total predictor being additive in team abilities, this model has no intercept term. This is because there is symmetry between the team identified as team 0 and the team identified as team 1.

If these identifiers are assigned randomly, the expected difference $\alpha_i - \alpha_j$ is zero. So the Bradley-Terry model is:

$$\log \frac{P_{ij}}{P_{ji}} = \alpha_i - \alpha_j$$

Alternatively we can write:

$$P_{ij} = \frac{e^{(\alpha_i)}}{e^{(\alpha_i)} + e^{(\alpha_j)}}$$

In the Bayesian framework we have to convert our simple likelihood into a proper Bayesian model, and all we need is a prior for the ability parameters. Such a prior will characterise the population of teams in terms of the distribution of their abilities.

$$\alpha_k \sim \mathsf{Normal}(0, 1)$$

Instead of hard centering the coefficients with prior adjustments, the normal prior with location parameter zero will implicitly center the paremeters around zero by assigning them higher density.

The unit scale of the normal prior provides an indication of how much variation there is in player ability. For the posterior fit object,we are taking $\alpha^{(m)}$ from the posterior:

$$p(\alpha \mid y) \propto p(y \mid \alpha) p(\alpha)$$

To calculate Bayesian estimates, we take posterior means, which are guaranteed to minimise expected square error when the model is well specified.

$$\begin{aligned}\hat{\alpha}_k &= \mathbb{E}\left[\alpha_k \mid y\right] \\ &= \int_{-\infty}^{\infty} \alpha_k \; p(\alpha_k|y) \; \mathrm{d}\alpha \\ &\approx \frac{1}{M}\sum_{m=1}^{M} \alpha_k^{(m)}\end{aligned}$$

This is an example of full Bayesian inference, which is nearly always based on calculating conditional expectations of quantities of interest over the posterior.

The second line defining the expectation shows the general form a weighted average of the quantity of interest, α_k, over the posterior distribution $p(\alpha_k|y)$. And last in the third line is the weighted sum from Markov chain Monte Carlo (MCMC) using an average of the posterior draws.

Table 5.5: Bradley Terry Model Analysis

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
α_1	1.0	3282	0.58	0.99	-1.3	0.5	2.8
α_2	1.0	3109	-0.58	0.99	-2.8	-0.5	1.3
log-posterior	1.0	3153	-4.62	1.41	-8.3	-4.3	-3.0

The output from the Stan's Bradley-Terry model (see Appendix for code and output), reported us:$\alpha_1 = 0.58$ and $\alpha_2 = -0.58$. So when team 2 is competing team 1, the probability of team 2 to win is:

$$\hat{P}_{21} = \frac{e^{(-0.58)}}{e^{(-0.58)} + e^{(0.58)}} = 0.239.$$

In the traceplot we see the mcmc diagnostics for the two probabilistic estimation of α's, on each (4) chains. The grey area indicates the warm-ip period.

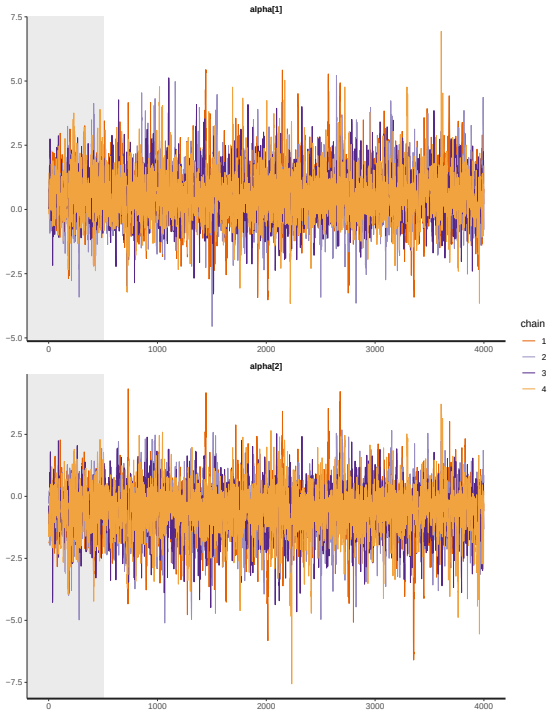

Figure 5.1: MCMC Diagnostics for each 4 chain

Part VI

Conclusion

Chapter 6

Conclusion

6.1 Summary Conclusion

Summarising this thesis we have succeeded a more deep understanding and knowledge of Bayesian thinking.The probabilistic approach gave us an alternative scientific tool from the classical approach of statistics, for computing the model comparisons and hypotheses testing with the powerful programming language R with the use of Stan.

Ending this thesis we summarise the results of this analysis in Bayesian framework for model comparison and hypothesis testing. Bayes factors are a tool for model comparison and hypothesis testing in contingency tables.According to model comparison, one can use it to compare variables (variable selection) and even prior distributions given, in a model.

In the other hand, bayes factors in contingency tables are proper for evaluating independence test given the distributional sampling (or the design of the trial-survey) and as we have seen in chapter 5 for testing McNemar test for dependent proportions.

For generalized linear models we showed that they are suitable when we have a small amount of observations and we want to make a population inference with a probabilistic perspective through bayesian analysis.

The estimated coefficients in generalized linear models follow the Normal distribution and we can safely insert normal distribution as prior information when we have a strong sense of the history of the problem that we are facing.In the other hand a non-informative prior for generalized linear models might be student-t or Cauchy distribtion.

For penalised likelihood criteria Akaike advocated that given a class of competing models for a dataset, one can choose the model that minimises

$$AIC = -2\log(y \mid \theta) + 2\kappa.$$

Two main justifications for the AIC have been advanced. The first, due to Akaike is based on a predictive argument. Suppose that given current data and a set of possible models we want the predictive distribution of a future datum.

Then if the predictive distribution is conditional on a single model and on its estimated parameters, the AIC picks the model that gives the best approximation, asymptotically. But such a predictive distribution is incorrect, because it does not incorporate the uncertainty about parameter values and model form.

Shibata and Katz (1976) have shown that the AIC tends to overestimate the number of parameters needed, even asymptotically. Thus if one must ignore both parameter uncertainty and model uncertainty when making predictions, it may be worthwhile to have a model that is too big.

The second main justification for the AIC, perhaps best described by Akaike (1983), is Bayesian. He wrote that model comparisons based on the AIC are asymptotically equivalent with those based on Bayes Factors.

But this is true only if the precision of the prior is comparable to that of the likelihood, but not in the more usual situation where prior information is small relative to the information provided by the data. In the latter more usual situation, the Schwarz criterion indicates that the model with the highest posterior probability is the one that minimises:

$$BIC = -2\log(y \mid \theta) + 2\log(n)$$

Comparing AIC and BIC indicates that BIC tends to favour simpler models than those by the AIC criterion.

Taking into account the fact that the computation of log likelihood is needed for the evaluation of AIC, BIC we introduced the IC criterion that takes the minimum of the maximum a posteriori that has a built in penalty of the prior information. This criterion as we saw in several examples is adequate for variable selection and also check each model predictive accuracy and is given by:

$$IC_{(m)} \approx D(\hat{\theta}_m, m) + d_m F$$

Ending chapter 4 for model comparison criteria we introduced the leave one out cross validation which a new and very promising method for evaluating model comparison based their predictive accuracy.

The idea that a data set is separated into train and test set give us no penalty but its not proper for variable selection and it cannot be used for statistical analysis that a data set cannot be split (like time-series) but this kind of analysis is beyond the scope of this thesis.

Calculating hypothesis testing in dependent proportions is yet not so much developed in bayesian analysis. We managed to represent some testing and model evaluation in bayesian framework that we find them useful and they are mainly implemented nowadays in Social sciences like McNemar test, Kappa coefficient and Bradley Terry models.

The contribution of Dennis Lindley in bayesian analysis of contingency tables was fundamental and is a milestone for new statistical scientists to lean over his work and take it, one more step further.

Alan Agresti, throughout his books of Categorical Data Analysis which are a benchmark for any statistician gave us a perspective to move forward to bayesian analysis of categorical data and dive into the world of probabilistic approach, along with some other scientists such as Jim Albert, Andrew Gelman and Eric-Jan Wagenmakers.

Bibliography

[1] Alan Agresti (2013). *An Introduction to Categorical Data Analysis 2nd Edition*. Wiley Series in Probability and Statistics.

[2] Alan Agresti & David Hitchcock (2005). *Bayesian inference for categorical data analysis 297-330*. Springer-Verlag.

[3] Jim Albert.(2008) *Bayesian Testing and Estimation of Association in a Two-Way Contingency Table* . Journal of the Statistical Association.

[4] Jim Albert. (2009). *Bayesian Computation with R* . Springer Use R!.

[5] Gustavo G. Bernando & Marcelo S. Lauretto & Julio M. Stern (2012). *The full Bayesian significance test for symmetry in contingency tables*. AIP Conference Proceedings.

[6] Michael Betancourt (2018). *A Conceptual Introduction to Hamiltonian Monte Carlo*. arXiv.

[7] Brandley Carlin & Thomas A. Louis (2019). *Bayesian Methods for Data Analysis*. CRC Press.

[8] Siddhartha Chid.(1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association,1313-1321.*

[9] Petros Dellaportas & Panagiotis Tsiamirtzis (2004). *Bayesian Statistics*. Athens University of Economics & Business Applied Statistics notes,Athens .

[10] Robert E.Kass & Adrian E.Raftery (1995). Bayes Factors. *Journal of the American Statistical Association,773-795.*

[11] Andrew Gelman & Jennifer Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[12] Andrew Gelman & Deborah Nolan (2002). *Teaching Statistics*. Oxford University Press.

[13] Andrew Gelman, John B.Carlin, Hal S.Stern, David B.Dunson, Aki Vehtari & Donald B.Rubin (2013). *Bayesian Data Analysis 3rd Edition*. Chapman & Hall/CRC Press.

[14] Matthew D. Hoffman & Andrew Gelman *The No-U-Turn Sampler : Adaptively Setting Path Lenghts in Hamiltonian Monte Carlo*. Springer,New York 2014.

[15] Tahira Jamil,Alexander Ly,Richard D. Morey, Jonathan Love,Maarten Marsman,Eric-Jan Wagenmakers.(2016). *Default "Gunel and Dickey" Bayes factors for contingency tables*. Springer 638-652.

[16] Harold Jeffreys (1961). *Theory of Probability 3rd Edition*. Oxford University Press.

[17] Maria Kateri (2014). *Contingency Table Analysis*. Springer.

[18] Maria Kateri & Petros Dellaportas (2001). *Bayesian Analysis of Correlated Proportions*. Sankhya.

[19] Maria Kateri & Petros Dellaportas. (2012). Conditional Symmetry Models for Three-Way Contingency Tables. *Journal of Statistical Planning and Inference.*

[20] Maria Kateri & Alan Agresti. (2013). *Bayesian inference about odds ratio structure in ordinal contingency tables*. Wiley Online Library.

[21] John K. Kruschke (2012). Bayesian Estimation Supersedes the t-Test. *Journal of American Psychological Association.*

[22] Ben Lambert (2018). *Bayesian Statistics*. SAGE Publications.

[23] Michael D.Lee & Eric-Jan Wagenmakers (2013). *Bayesian Cognitive Modeling*. Cambridge University Press.

[24] Steven M. Lewis and Adrian E.Raftery (1994). *Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator*. University of Washington.

[25] Jean-Michel Marin & Christian Robert (2011). *Bayesian Essential with R 2nd Edition* . arXiv.

[26] Peter McCullagh & J.A Nelder (1989). *Generalized Linear Models 2nd Edition*. Chapman & Hall CRC Press.

[27] Richard McElreath (2015). *Statistical Rethinking*. CRC Press.

[28] Richard McElreath & Jeremy Koster. *Multinomial analysis of behavor statistical methods*. Springer,138-152, 2017.

[29] Ioannis Ntzoufras.(2009). *Bayesian Modeling Using WinBugs*. Wiley Series in Probability and Statistics.

[30] Alan Turing (1941). *Statistics of Repetitions*. arXiv, 2015.

[31] Aki Vehtari & Andrew Gelman & Johan Gabry (2016). *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC*.

Appendix A

Appendix

For the implementation of the examples on this thesis you can find the full code on the following link:
https://github.com/nikosmatsa/Thesis-Bayesian-Model-Comparison-and-Hypothesis-Testing-for-Contingency-Tables

```
https://github.com/nikosmatsa/Thesis-Bayesian-Model-Comparison-and-Hypothesis\
protect\discretionary{\char\hyphenchar\font}{}{}Testing-for-Contingency-Tables/
blob/master/Thesis%20R%20Code
```