Master's thesis

Nora Røhnebæk Aasen

Saddlepoint approximated methods for computing *p*-values in score tests

Master's thesis in Mathematical Sciences
Supervisor: Thea Bjørnland
June 2023

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Nora Røhnebæk Aasen

Saddlepoint approximated methods for computing *p*-values in score tests

Master's thesis in Mathematical Sciences
Supervisor: Thea Bjørnland
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

The goal of this thesis is to present, both theoretically and through examples, two alternative methods of computing p-values in a score test as opposed to assuming a normal distribution of the score. We consider a double saddlepoint approximation and show that it takes a simple form when working with certain generalized linear models, which yields an easier implementation that can be used for computing p-values of a score test. Furthermore, we discuss an alternative score statistic called effective score, which handles nuisance parameters using projection methods, as opposed to performing conditional inference. Along with saddlepoint approximation, the effective score yields an unconditional test. We compare both alternatives, the double saddlepoint method and the effective score method, to each other and to a regular score test, using both simulated and real data sets.

Sammendrag

I denne oppgaven presenteres, både teoretisk og ved hjelp av eksempler, to alternative metoder for å regne p-verdier i en score test i motsetning til å anta en normalfordeling på score-statistikken. Vi presenterer en dobbel sadelpunkt-approksimasjon og viser at den får et enkelt uttrykk for visse generaliserte lineære modeller, hvilket resulterer i en enklere implementering som kan brukes for å regne ut p-verdier av en score test. Videre diskuterer vi en alternativ score-statisikk kalt effektiv score, som håndterer plageparametre ved hjelp av projeksjonsmetoder, i motsetning til å gjøre betinget inferens. Vi sammenligner begge alternativer, dobbel sadelpunkt og effektiv score, med hverandre samt til vanlig score test, ved hjelp av både simulerte og ekte datasett.

Acknowledgments

This thesis marks the end of my master's degree in mathematics at NTNU, as well as my five years as a student in Trondheim.

First and foremost, I would like to thank my supervisor Thea Bjørnland, who gave me a challenging and interesting topic when I expressed an interest in pursuing a more theoretically angled master's thesis. I would not have been able to write this were it not for your patience, your keen interest, and our detailed discussions throughout the last year. Thanks also to Ingeborg Gullikstad Hem for, on short notice, providing me with an interesting data set and a problem setup related to neurological research that I could use to investigate my methods.

These five years in Trondheim, in particular this last one, has been an intense ride that probably would have derailed at some point were it not for the amazing people around me. Thanks to my amazing roommates, Ingvild and Nora, who are my best friends and who made this time living in Trondheim immensely better. You have been my closest family while living in Trondheim, but I wish also to thank my family at Hamar, who is always there for me when I need help or extra support. Thanks also to all the great people at Matteland I have gotten to know, with a special shout-out to my dear friends Irma and Clara, and thanks in particular to Lesesal 395B and all its lovely inhabitants. Lastly, thanks to Elias, who has been a patient and loving companion both in mathematics and in life. Writing a master thesis with you by my side (literally) has been challenging, but I would never have wished to be without your care, love, and support.

Nora Røhnebæk Aasen
Trondheim, Norway
June 2023

Contents

Chapter 1

Introduction

Score tests, along with likelihood ratio tests and Wald tests, are commonly employed for evaluating parameters or coefficients within a generalized linear model. While the asymptotic distribution of the score test follows a normal distribution, the finite sample distribution is not generally known. Consequently, when applying a score test to a data sample that fails to exhibit sufficient convergence of the score distribution, the test can yield an increased rate of type I errors compared to the desired level of significance. Such a challenge arises in cases of small sample inference where the sample size, denoted as n, is too small for the score to converge adequately. Another circumstance in which this issue arises is when dealing with imbalanced large data sets, as discussed by Johnsen et al. (2023), for instance.

In the article by Johnsen et al. (2023), they investigate saddlepoint approximations to score tests in a logistic regression model for genome-wide association studies. In this master thesis, we build upon their article by presenting their methods in a more general framework, showcasing that the methods are applicable to other models than logistic regression models. We demonstrate that the saddlepoint approximation, subject to certain assumptions, can be reformulated in terms of maximum likelihood estimates which yields an easier implementation. We also demonstrate the relationship between the effective score and the so-called conditional score, in an attempt to build a stronger theoretical foundation for this transformed score statistic.

The layout of this thesis is as follows. We begin with preliminaries in chapter 2, where we focus on hypothesis testing, generalized linear models, and score tests. We emphasize the relationship between the score and a sufficient statistic for the parameters of the generalized linear model, as this later motivates the rewriting of the saddlepoint approximation.

In chapter 3, we consider saddlepoint approximation and begin by offering a concise introduction to the concept, highlighting its relevance in the context of computing p-values for score tests. Saddlepoint approximation allows us to approximate the tail probabilities of a distribution using the cumulant generating function of the distribution. Furthermore, we demonstrate that the saddlepoint approximation can be expressed in terms of maximum likelihood estimates. This connection allows us to implement saddlepoint approximation by utilizing output readily obtained from fitting a generalized linear model in statistical software.

Moving on, we delve into the concept of the "effective score" which has emerged as an alternative to the basic score statistic, offering improved robustness against the impact of nuisance estimates (Hemerik et al., 2020). We explore the relationship between the effective score and the conditional score as a mean of grating deeper insight into the theoretical foundations of the effective score and show that the effective score in conjunction with saddlepoint approximation leads to an alternative test to the regular score test. This approach, which we refer to as an "effective score test", capitalizes on the improved robustness of the effective score and leverages saddlepoint approximation to estimate its unconditional distribution.

The motivation behind these alternative methods is to enhance the accuracy and reliability of parameter inference within generalized linear models, particularly in situations where the regular score test may suffer from inflated type I error rates. To evaluate the efficiency of both the double saddlepoint approximated p-values and the effective score test, we conduct simulations and analyze real-world data sets. In our simulated experiments, we explore the two different implementations of saddlepoint approximation and compare their performance to assess their suitability for practical applications. We also examine the asymptotic properties of each test, investigating aspects such as consistency, efficiency, and the control of type I error rates. Through these analyses, we aim to establish the strengths and limitations of the proposed methods and provide guidance on their appropriate usage.

For the empirical evaluation, we consider two data sets; a small sample data set and a large data set with an imbalanced response. These data sets represent scenarios where the performance of traditional statistical tests may be compromised and approaches such as the effective score test and double saddlepoint approximation could offer significant advantages. By comparing the outcomes of these alternative methods against those of the regular score test, we aim to assess their practical utility and their effectiveness in real-world statistical analyses.

Throughout this thesis, all simulations and plots are executed using the statistical software `R Studio` (Team, 2021), with the graphical representations generated using the `ggplot2` package (Wickham, 2016).

Chapter 2

Preliminaries

In this chapter, we consider the preliminaries needed for this thesis. As mentioned in the introduction, this thesis aims to present methods that can compute p-values for score tests with higher accuracy than a normal approximation. Hence, we are interested in hypothesis testing, more particularly hypothesis testing around parameters in a generalized linear model. We begin by presenting hypothesis testing, and then we look at the generalized linear model. The last thing we look at in this chapter is the score, how it relates to sufficient statistics, and how it can be used to perform a hypothesis test.

2.1 Hypothesis testing

A hypothesis test is defined as a *statement about a parameter*, with the goal of deciding, based on a data sample $Y_1, \dots, Y_n$ from the population, which of two complementary hypotheses is true (Casella and Berger, 2001: p. 373). Assume there exists a population parameter $\boldsymbol{\theta}$, which can be either a scalar or a vector. The hypotheses are formulated in terms of a *null hypothesis* and an *alternative hypothesis*. Often the null hypothesis assumes that $\boldsymbol{\theta}$ has some value $\boldsymbol{\theta}_0$, as follows

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{v.s.} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0. \tag{2.1}$$

We will call this a *two-tailed* hypothesis test. A *one-tailed* test is formulated as one of the following

$$H_0 : \boldsymbol{\theta} \leq \boldsymbol{\theta}_0 \quad \text{v.s.} \quad H_1 : \boldsymbol{\theta} > \boldsymbol{\theta}_0 \tag*{(Right-tailed test)}$$
$$H_0 : \boldsymbol{\theta} \geq \boldsymbol{\theta}_0 \quad \text{v.s.} \quad H_1 : \boldsymbol{\theta} < \boldsymbol{\theta}_0 \tag*{(Left-tailed test)}$$

2.1.1 Performing and evaluating a test

Following Casella and Berger (2001; Ch. 8), a hypothesis test is usually formulated in terms of a test statistic, which is a function of the data sample $W(Y_1, \ldots, Y_n) = W(\boldsymbol{Y})$. The criteria for rejecting or accepting H_0 depends on the test statistic. One criterion that specifies a test is a *critical value* c, where we for example reject H_0 if $W(\boldsymbol{Y}) \geq c$. This is equivalent to defining a rejection region R, where H_0 is rejected if $W(\boldsymbol{Y}) \in R$. Given a critical value c, the corresponding rejection region is

$$R = \{(y_1, \ldots, y_n) \,|\, W(\boldsymbol{y}) \geq c\}.$$

We cannot guarantee that the conclusion of a hypothesis test is correct. However, we can *control* the probability of making certain errors. There are two types of error that can be made (see table 2.1): A *type I error* is when we falsely reject the null hypothesis, while a *type II error* is when we falsely keep the null hypothesis.

The probability of type I error, denoted α, is referred to as the *significance level*, or simply *level*, of the test. The probability of correctly rejecting the null hypothesis, denoted $1 - \beta$ in table 2.1, is referred to as the *power* of the test.

	H_0 true	H_0 false
Keep H_0	$1 - \alpha$	P(Type II error) = β
Reject H_0	P(Type I error) = α	$1 - \beta$

Table 2.1: Table indicating different outcomes of an hypothesis test.

When comparing different tests, one will often first consider to what degree the test is able to control the type I error probability at a set level (Casella and Berger, 2001: p. 385). Let $W(\boldsymbol{Y})$ be any test statistic and R the rejection region for some hypothesis test. The class of tests that are such that

$$P(W(\boldsymbol{Y}) \in R \,|\, H_0 \text{ true}) \leq \alpha$$

make up the set of level α tests for that hypothesis. If a suggested test has a substantially smaller level than the decided significance level α, the test is said to be *conservative*. In general, we want the test to have a probability of type I error as close as possible to the chosen level from below.

An alternative, albeit equivalent, way of deciding whether or not to reject H_0 is by evaluating the p-value of the test. Given some test statistic $W(\boldsymbol{Y})$, we define the p-value of the test as the *probability of observing $W(\boldsymbol{y})$ or something more extreme under the assumption that the null hypothesis is true*. In mathematical terms, this can be written as

$$P(W(\boldsymbol{Y}) \geq W(\boldsymbol{y}) \,|\, H_0 \text{ true}) = p.$$

We reject H_0 if $p \leq \alpha$, meaning p is smaller than the chosen level of the test. By writing conditional on "H_0 true", we mean that the p-value is calculated based on hypothesized true values $\boldsymbol{\theta}_0$ for $\boldsymbol{\theta}$. For simplicity, we introduce the notation

$$P(W(\boldsymbol{Y}) \geq W(\boldsymbol{y})\,;\, H_0) = p,$$

to mean that the probability is computed under the assumption that H_0 is true.

2.1.2 Bonferroni Correction

Consider now that we wish to perform multiple hypotheses tests $\mathcal{H}_1, \ldots, \mathcal{H}_m$, where each $\mathcal{H}_j$ is formulated in terms of a null hypothesis and an alternative hypothesis, and the significance level

$$\alpha_j = P(\text{Making a type I error in test } \mathcal{H}_j)$$

is α for all $i = 1, \ldots, m$. Then the overall probability of making one or more type I error over all tests becomes much larger than the chosen significance level of each test.

There are multiple methods that can be used to compensate for this (Goeman and Solari, 2014), and one way is the Bonferroni method which aims to limit the *familywise error rate (FWER)*. The FWER is defined as the probability of committing *at least one type I error*. The Bonferroni correction proposes to perform each test at significance level $\alpha_j^* = \alpha/m$, where α is the decided, overall significance level and m is the number of tests. The Bonferroni method can be very conservative in some cases, but will always guarantee control of the FWER.

2.1.3 Clopper-Pearson Interval

A $(1-\alpha)100\%$ *confidence interval* for a parameter θ is an interval $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ based on the data $\boldsymbol{Y}$, such that

$$P(\theta \in [L(\boldsymbol{Y}), U(\boldsymbol{Y})]) = 1 - \alpha,$$

for some decided level α.

The Clopper-Pearson interval is an exact confidence interval for some unknown probability of success p in a Binomial distribution. Letting X be the number of successes and n be the number of trials, the Clopper-Pearson interval is given by

$$\left[B\left(\frac{\alpha}{2}; X, n - X + 1\right), B\left(1 - \frac{\alpha}{2}; X + 1, n - X\right)\right],$$

where B refers to quantiles in the *Beta(a,b)* distribution (Thulin, 2014). This expression follows from the relationship between binomial distribution and beta distribution.

2.2 Generalized Linear Models

Following McCullagh and Nelder (1989), let $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ be a vector of length n, where each element is one observation. For each observation y_i there is a set of p covariates $x_{i1}, \ldots, x_{ip}$ that influence the response with magnitude described by the coefficients $\beta_1, \ldots, \beta_p$. Generalized linear models are a wide class of models that can be fitted in this situation. The model has the following three defining properties.

i. The random component

The random component of the model is a random vector $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^T$, whom the observations, $\boldsymbol{y}$, are assumed to be realizations of. The components Y_i are assumed to be independent and have mean $\mu_i = E[Y_i]$. Furthermore, each component is assumed to come from a distribution $Y_i \sim f(y_i)$, either discrete or continuous, such that $f(y_i)$ belongs to the *exponential family*, and can be written in the general form

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - A(\theta_i)}{a(\phi)} + h(y_i, \phi) \right\}, \tag{2.2}$$

for some functions a, A, and h. The parameter θ is called the *canonical parameter* of the distribution, whereas ϕ is called the *dispersion parameter* and is typically assumed to be known or estimated independently of the other parameter.

Remark 2.2.1. Some of the important distributions that belong to the exponential family include normal distribution, binomial distribution with known n, Poisson distribution, and exponential distribution.

ii. The systematic component

For each observed value y_i we assume there is a *linear predictor* defined as

$$\eta_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{2.3}$$

The covariate vector $\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{ip}]^T$ consists of known covariates. The vector of coefficients $\beta_0, \ldots, \beta_p$ are unknown and must be estimated from the data.

iii. The link

The link is a function that relates the systematic component to the random component, such that

$$\eta_i = g(\mu_i).$$

The function g is called the *link function* of the model. It is also possible to write

$$\mu_i = g^{-1}(\eta_i),$$

in which case g^{-1} is usually called the *mean function*. For every distribution f that can be written as in equation (2.2), there is also a unique link function called the *canonical link*. This link function satisfies

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = g(\mu_i) = \theta_i,$$

where θ_i is the canonical parameter in the distribution of Y_i as written in equation (2.2).

2.2.1 Linear model

The classical linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon,$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a special case of a generalized linear model. The matrix $\boldsymbol{X}$ is a $(n \times p)$ matrix with row i equal to $\boldsymbol{x}_i^T$. The components Y_i are assumed to come from a normal distribution, with expected value $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and constant variance σ^2. The linear predictor is

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij},$$

and it can be easily seen that the link function in the linear model then must be the identity, since $\mu_i = \eta_i$.

2.2.2 Logistic regression model

Logistic regression is a classification problem where the observation $\boldsymbol{y}$ is assumed to be a realization from a random vector $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^T$ with independent elements such that $Y_i \sim \text{Bernoulli}(\mu_i)$ for $i = 1, \ldots, n$, meaning

$$Y_i = \begin{cases} 1 & \text{with probability } \mu_i, \\ 0 & \text{with probability } 1 - \mu_i. \end{cases}$$

The probability mass function of Y_i can be written as

$$\begin{aligned} p(y_i; \mu_i) &= \mu_i^{y_i} (1 - \mu_i)^{(1-y_i)} \\ &= \exp\left\{ y_i \ln \mu_i + (1 - y_i) \ln(1 - \mu_i) \right\} \\ &= \exp\left\{ y_i \ln\left(\frac{\mu_i}{1 - \mu_i} \right) + \ln(1 - \mu_i) \right\}, \end{aligned}$$

and we see that this is an exponential family as in equation (2.2) with

$$\begin{aligned}\theta_i &= \ln\left(\frac{\mu_i}{1-\mu_i}\right),\\ A(\theta_i) &= -\ln(1+\exp(\theta_i)),\\ h(y_i,\phi) &= 0,\\ a(\phi) &= 1.\end{aligned}$$

The expected value of a Bernoulli distributed variable is $E[Y_i] = \mu_i$, and given a linear predictor

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij},$$

we can relate the two using the canonical link function

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \ln\left(\frac{\mu_i}{1-\mu_i}\right).$$

This link is called the *logit* link function.

2.2.3 Poisson regression model

In situations where the responses Y_i are count data, meaning they take values $\{0,1,2,\ldots\}$, it is common to model the distribution using a Poisson regression model. That means we assume that the observation $\boldsymbol{y}$ is a realization of the random vector $\boldsymbol{Y} = [Y_1,\ldots,Y_n]$ which consists of independent elements $Y_i \sim \text{Pois}(\mu_i)$. The probability mass function of Y_i is defined as

$$\begin{aligned}p(y_i;\mu_i) &= \frac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i)\\ &= \exp\left\{y_i \ln\mu_i - \mu_i - \ln y_i!\right\}.\end{aligned}$$

Once again we recognize the exponential family from equation (2.2), with

$$\begin{aligned}\theta_i &= \ln\mu_i,\\ A(\theta_i) &= \exp(\theta_i),\\ h(y_i,\phi) &= -\ln y_i!,\\ a(\phi) &= 1.\end{aligned}$$

The expected value of a Poisson distributed variable is $E[Y_i] = \mu_i$. Furthermore, we have some linear predictor

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}.$$

The canonical link function for a generalized linear model with Poisson distributed response is the *log*-function. Hence, the linear predictor relates to the expected value of Y_i through

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \ln\mu_i.$$

2.2.4 General assumptions and notaion

This thesis attempts to present a more general, theoretical tool. However, we wish to make some general assumptions that are needed for many of the theoretical results before proceeding.

From subsection 2.2.2 and subsection 2.2.3, it is seen that both the Bernoulli and Poisson distributions could be written in general terms as

$$f(y_i;\theta_i) = \exp\{y_i\theta_i - A(\theta_i) + h(y_i)\}. \tag{2.4}$$

In this thesis, we will restrict ourselves to distributions belonging to the so-called *regular exponential family*. By this, we will mean that the distribution of Y_i can be written as in equation (2.4) and that the linear predictor is *regular*. The linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ is regular if it consists of a regular covariate matrix, $\boldsymbol{X}$, with rank equal to the dimension of the covariates, and covariates $\boldsymbol{\beta}$ that are unconstrained.

Furthermore, we always assume that the linear predictor η_i and the canonical parameter θ_i are linked by a canonical link function, meaning $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. From these assumptions, we have that the distribution in equation (2.4) is equivalent to

$$f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i) = \exp\{y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - A(\boldsymbol{x}_i^T\boldsymbol{\beta}) + h(y_i)\}. \tag{2.5}$$

For simplicity, we will let $(Y_1,\boldsymbol{x}_1),\dots,(Y_n,\boldsymbol{x}_n)$ denote the random sample where each Y_i is a random variable with distribution $f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i)$.

2.3 Likelihood theory and the score

In this section, we introduce the likelihood function and the score, and we explain how the coefficients, or parameters, in the linear predictor are estimated by maximizing the likelihood function, or, equivalently, the log-likelihood function.

2.3.1 Likelihood function

Following Fahrmeir et al. (2013), let $(y_1,\boldsymbol{x}_1)\dots,(y_n,\boldsymbol{x}_n)$ be a set of independent observations, and let each y_i be a realization from the random variable $Y_i \sim f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i)$. The results discussed in this section hold even if we do not assume that all elements of $\boldsymbol{Y} = [Y_1,\dots,Y_n]^T$ have the same distribution, as long as they depend on the same parameters $\boldsymbol{\beta}$. However, if the probability distribution $f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i)$ is equal for all Y_i, then $\boldsymbol{Y}$ is a *random sample* of independent and identically distributed (IID) variables, with joint distribution given by

$$f(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{X}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i).$$

For notational simplicity we introduce $f(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{X})$ to mean $f(y_1,\dots,y_n;\boldsymbol{\beta},\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$.

The likelihood function is then defined as

$$\mathcal{L}(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i) = f(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{X}).$$

Note that the likelihood is a function of the parameters, depending on the observed data, as opposed to the probability distribution, which is a function of the data, depending on the parameters.

The *log-likelihood* is defined as $\ell(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \ln(\mathcal{L}(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}))$, and can therefore be written as

$$\ell(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \sum_{i=1}^{n} \ln f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i).$$

Assuming that $(Y_1,\boldsymbol{x}_1),\dots,(Y_n,\boldsymbol{x}_n)$ is a random sample from a regular exponential family, the expression for the log-likelihood can be written generally as

$$\ell(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \sum_{i=1}^{n} \left\{\theta_i y_i - A(\theta_i) + h(y_i)\right\}. \tag{2.6}$$

Equivalently, since we always assume canonical link, meaning $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, this can be rewritten as

$$\ell(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \sum_{i=1}^{n} \boldsymbol{\beta}^T\boldsymbol{x}_i y_i - \mathcal{A}_i(\boldsymbol{\beta}) + h(y_i), \tag{2.7}$$

where we define

$$\mathcal{A}_i(\boldsymbol{\beta}) = A(\boldsymbol{x}_i^T\boldsymbol{\beta}). \tag{2.8}$$

Both expressions will be used throughout the thesis, depending on what is most practical. However, we emphasize that the expressions in equation (2.6) and equation (2.7) are the same.

The *score* is defined as the derivative of the log-likelihood and can be written as

$$s(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\beta}} \ln f(y_i;\boldsymbol{\beta},\boldsymbol{x}_i). \tag{2.9}$$

Since $\boldsymbol{\beta}$ is a $(p+1)$-dimensional vector, the score is the gradient of the log-likelihood with respect to $\boldsymbol{\beta}$, and therefore itself a $(p+1)$-dimensional vector, which can be written as,

$$s(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \begin{bmatrix} s_{\beta_0}(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) \\ \vdots \\ s_{\beta_p}(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) \end{bmatrix}.$$

2.3.2 Maximum Likelihood Estimators

In order to estimate the parameters $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]^T$ we can maximize the likelihood or, equivalently, the log-likelihood function, since the logarithm is a strictly increasing function. This will yield an estimate which is a realization of the *maximum likelihood estimator (MLE)* $\hat{\boldsymbol{\beta}}$. For linear models, this estimator has a known, analytic solution which is

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

However, for generalized linear models, the maximum likelihood estimator does not generally have a closed-form solution.

The maximum likelihood estimator is the estimator that maximizes the likelihood function, meaning

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}).$$

This is equivalent to solving

$$s(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) = \begin{bmatrix} s_{\beta_0}(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \\ \vdots \\ s_{\beta_p}(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{2.10}$$

since the derivative of the log-likelihood will be zero at its maximum point. Finding the maximum likelihood estimator for $\boldsymbol{\beta}$ is done by solving the $(p+1)$-dimensional, non-linear system of equations given in equation (2.10). The most common method of doing this is by using the Fisher scoring algorithm (Fahrmeir et al., 2013: p. 324), however we will not derive this in detail.

For a regular exponential family, the log-likelihood will be a strictly concave function. Hence, if there exist a maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ that satisfy

$$s(\hat{\boldsymbol{\beta}}; \boldsymbol{Y}, \boldsymbol{X}) = \boldsymbol{0},$$

it will be unique (Butler, 2007: p.148).

2.4 Score testing

In section 2.1 we said that a hypothesis test is usually formulated in terms of a test statistic $W(Y_1, \ldots, Y_n)$. For generalized linear models, we are often interested in performing inference about one or more of the parameters $\beta_0, \ldots, \beta_p$. The score can then be used as a test statistic, in which case we call it a *score statistic*. The score statistic can be particularly advantageous when the goal is to test some simpler model, where one or more of the coefficients $\beta_0, \ldots, \beta_p$ are 0, against a larger and more complex model. This is because the score test only requires us to fit the smaller model, and thus only compute the maximum likelihood estimators for a subset of the coefficients.

Let $(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n)$ be a random sample, and let the score vector be defined as before, meaning

$$s(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) = \begin{bmatrix} s_{\beta_0}(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \\ \vdots \\ s_{\beta_p}(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \beta_0} \ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \\ \vdots \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \end{bmatrix}. \tag{2.11}$$

Note that when $\boldsymbol{S}(\boldsymbol{\beta}) = s(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X})$ depends on an unrealized random vector $\boldsymbol{Y}$, the score can itself be thought of as a random vector with expectation, variance, and distribution of its own.

Following Lindsey (1996; p. 188-189), the expectation of $\boldsymbol{S}(\boldsymbol{\beta})$, with the parameters assumed to be fixed at their true values, is then computed as

$$\begin{aligned} E[\boldsymbol{S}(\boldsymbol{\beta})] &= E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X})\right] \\ &= E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \ln f(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{X})\right] \\ &= E\left[\frac{\frac{\partial}{\partial \boldsymbol{\beta}} f(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{X})}{f(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{X})}\right] \\ &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial \boldsymbol{\beta}} f(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{X})}{f(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{X})} f(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{X}) d\boldsymbol{Y} \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} \int_{\mathcal{Y}} f(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{X}) d\boldsymbol{Y} \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} 1 \\ &= \mathbf{0}. \end{aligned} \tag{2.12}$$

The variance of the score can be shown to be the expected Fisher information matrix, still under the assumption that $\boldsymbol{\beta}$ is fixed at its true value. First note that since $E[\boldsymbol{S}(\boldsymbol{\beta})] = \mathbf{0}$, we get

$$\text{Var}[\boldsymbol{S}(\boldsymbol{\beta})] = E[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^T].$$

It is a well-known result (Casella and Berger, 2001: p. 338) that

$$E[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^T] = -E\left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right],$$

under certain regularity conditions. In particular, this holds for every probability distribution belonging to the exponential family. Hence,

$$\begin{aligned}
\text{Var}[\boldsymbol{S}(\boldsymbol{\beta})] &= -E\left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] \\
&= -E\left(\frac{\partial^2 \ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right) \\
&= -E\begin{bmatrix} \frac{\partial^2}{\partial \beta_0^2}\ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) & \dots & \frac{\partial^2}{\partial \beta_0 \partial \beta_p}\ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \beta_p \partial \beta_0}\ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) & \dots & \frac{\partial^2}{\partial \beta_p^2}\ell(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \end{bmatrix} \\
&= \mathcal{I}(\boldsymbol{\beta}),
\end{aligned}$$

with $\mathcal{I}(\boldsymbol{\beta})$ being the *expected Fisher information matrix*.

As noted by Lindsey (1996; p. 215), the exact distribution of the score statistic is only tractable in certain simple cases, but not generally. However, the asymptotic distribution can be shown to be a *multivariate normal distribution*. This follows from the central limit theorem since the score vector is the sum of independent random variables. In particular, from equation (2.9) we have

$$\begin{aligned}
\boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \ln f(Y_i; \boldsymbol{\beta}, \boldsymbol{X}) \\
&= \sum_{i=1}^{n} S_i(\boldsymbol{\beta}),
\end{aligned}$$

where each $S_i(\boldsymbol{\beta})$ is a random variable depending on Y_i. Hence, under the assumption that $\boldsymbol{\beta}$ is fixed at the true parameter values, we have that, asymptotically,

$$\boldsymbol{S}(\boldsymbol{\beta}) \sim \mathcal{N}_{p+1}(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta})).$$

Nuisance parameters

When performing inference about a regression model we often have only one parameter that we are interested in investigating, whereas the remaining parameters are parameters that are of no intrinsic interest besides adding meaning to the model (McCullagh and Nelder, 1989: p. 245). These other parameters are called *nuisance parameters*. Consider now that instead of $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, we define our problem in terms of a scalar[1] *parameter of interest*, γ, and a vector of nuisance parameters, $\boldsymbol{\beta}$, which in mathematical terms can be written as,

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\gamma.$$

[1] Some results in this section generalize to a vector parameter of interest, but we only consider testing one parameter at a time and therefore introduce this notation from the start.

The full covariate matrix then becomes a $(n \times p + 1)$ matrix $\boldsymbol{V} = [\boldsymbol{X}, \boldsymbol{Z}]$, where $\boldsymbol{X}$ is $(n \times p)$ and $\boldsymbol{Z}$ is $(n \times 1)$. The full parameter vector is a $(p + 1 \times 1)$ vector $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$, where $\boldsymbol{\beta}$ is $(p \times 1)$ and γ is a scalar.

The hypothesis test we are interested in only concerns our parameter of interest, γ, and since score tests are optimal for testing if the true model should have fewer covariates, as mentioned earlier, the hypothesis will often be

$$H_0 : \gamma = 0 \quad \text{v.s.} \quad H_1 : \gamma \neq 0. \tag{2.13}$$

Hence, we are interested in testing if the parameter of interest γ should be included in the linear predictor, or not.

For the score statistic, we then get a natural partition, keeping in mind that the distribution of the score is a normal distribution only asymptotically and with the parameters $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$ fixed at their true parameter values,

$$\boldsymbol{S}(\boldsymbol{\psi}) = \begin{bmatrix} \boldsymbol{S}_\beta(\boldsymbol{\psi}) \\ S_\gamma(\boldsymbol{\psi}) \end{bmatrix} \sim \mathcal{N}_{p+1}\left(\begin{bmatrix} \boldsymbol{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \boldsymbol{I}_{\beta\gamma} \\ \boldsymbol{I}_{\gamma\beta} & I_{\gamma\gamma} \end{bmatrix} \right).$$

Following Smyth (2003), if the nuisance parameters $\boldsymbol{\beta}$ are known, the score test statistic under H_0 from equation (2.13) becomes simply

$$S_\gamma(\boldsymbol{\psi}_0) \overset{H_0}{\approx} \mathcal{N}(0, I_{\gamma\gamma}(\boldsymbol{\psi}_0)).$$

We write $\boldsymbol{\psi}_0$ to highlight that the null-value of γ, which in our case is 0, is used in place of the parameter to compute the p-value. A low p-value indicates that our distributional assumptions could be wrong, namely that $\gamma_0 = 0$ is not the true value of the parameter, and this would lead to a rejection of the null hypothesis.

Generally, we do not have known nuisance parameters. They must therefore be estimated, and the score test uses the maximum likelihood estimators, estimated under the null hypothesis, fixing the nuisance parameters at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. This is equivalent to $\boldsymbol{S}_\beta(\boldsymbol{\psi}_0) = 0$. The score test statistic therefore becomes

$$S_\gamma(\boldsymbol{\psi}_0) \mid \boldsymbol{S}_\beta(\boldsymbol{\psi}_0) = 0 \overset{H_0}{\approx} \mathcal{N}(0, I_{\gamma\gamma} - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\gamma}), \tag{2.14}$$

where we see that the distribution accounts for the fact that we use estimators $\hat{\boldsymbol{\beta}}$, as opposed to the true parameter values $\boldsymbol{\beta}$, which are unknown to us.

Remark 2.4.1. The conditional distribution of the score $S_\gamma(\boldsymbol{\psi}_0)$ as given in equation (2.14) follows from the known, conditional distribution of normally distributed variables (Härdle and Simar, 2015: 186). Given two normal random vectors, X_1 and X_2 such that

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right).$$

Then the conditional distribution of $X_2 \mid X_1 = x_1$ has the known expression

$$X_2 \mid X_1 = x_1 \sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}).$$

2.4.1 Score tests for a regular exponential family

In this section, we restrict ourselves to a random sample from a regular exponential family, $(Y_1, \boldsymbol{v}_1), \ldots, (Y_n, \boldsymbol{v}_n)$, assuming an underlying generalized linear model with the canonical link function, meaning

$$\theta_i = \eta_i = \boldsymbol{v}_i^T \boldsymbol{\psi} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \gamma z_i.$$

Note that we are keeping with the notation above, in terms of distinguishing between nuisance parameters and the parameter of interest.

Consider again the log-likelihood as in equation (2.6). Note that the score then takes the form

$$\begin{aligned} s(\boldsymbol{\psi}; \boldsymbol{Y}, \boldsymbol{V}) &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\psi}} \{\theta_i Y_i - A(\theta_i) + h(Y_i)\} \\ &= \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \boldsymbol{\psi}} \frac{\partial}{\partial \theta_i} \{\theta_i Y_i - A(\theta_i) + h(Y_i)\} \\ &= \sum_{i=1}^{n} \boldsymbol{v}_i (Y_i - A'(\theta_i)) \\ &= \boldsymbol{V}^T \boldsymbol{Y} - \boldsymbol{V}^T A'(\boldsymbol{\theta}). \end{aligned}$$

We will note two important things before moving on.

Proposition 2.4.2. *The expression $A'(\boldsymbol{\theta})$ is the expectation of $\boldsymbol{Y}$.*

Proof. We start by computing the derivative of $f(y_i; \theta_i)$ with respect to θ_i, which is

$$\begin{aligned} \frac{\partial}{\partial \theta_i} f(y_i; \theta_i) &= \frac{\partial}{\partial \theta_i} \exp\{\theta_i y_i - A(\theta_i) + h(y_i)\} \\ &= (y_i - A'(\theta_i)) \exp\{\theta_i y_i - A(\theta_i) + h(y_i)\} \\ &= (y_i - A'(\theta_i)) f(y_i; \theta_i). \end{aligned}$$

Then, integrating on the left side yields

$$\int \frac{\partial}{\partial \theta_i} f(y_i; \theta_i) dy_i = \frac{\partial}{\partial \theta_i} \int f(y_i; \theta_i) dy_i = \frac{\partial}{\partial \theta_i} 1 = 0.$$

and integration on the right side yields

$$\int y_i - A'(\theta_i) f(y_i; \theta_i) dy_i = \int y_i f(y_i; \theta_i) dy_i - \int A'(\theta_i) f(y_i; \theta_i) dy_i = E[Y_i] - A'(\theta_i).$$

Hence, we get $E[Y_i] = A'(\theta_i)$ after rearranging the terms and it is clear that

$$E[\boldsymbol{Y}] = \begin{bmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{bmatrix} = \begin{bmatrix} A'(\theta_1) \\ \vdots \\ A'(\theta_n) \end{bmatrix} = A'(\boldsymbol{\theta}).$$

□

From this, it is clear that we can write the score in terms of the response vector $\boldsymbol{Y}$ and its expected value $\boldsymbol{\mu}$, as

$$\boldsymbol{S}(\boldsymbol{\psi}) = \boldsymbol{V}^T(\boldsymbol{Y} - \boldsymbol{\mu}). \tag{2.15}$$

Proposition 2.4.3. *The first term in the score, $\boldsymbol{V}^T\boldsymbol{Y}$, is a sufficient statistic for $\boldsymbol{\psi}$.*

According to Casella and Berger (2001; p.272), a sufficient statistic can be defined as follows.

Definition 2.4.4. Denote the data $\boldsymbol{Y}$, and let $T(\boldsymbol{Y})$ be a statistic, that is, a function of the data. Then $T(\boldsymbol{Y})$ is said to be *sufficient* for a parameter $\boldsymbol{\psi}$ if the conditional distribution of $\boldsymbol{Y}$ given the statistic $T(\boldsymbol{Y})$ does not depend on $\boldsymbol{\psi}$.

Informally, sufficient statistics can be thought of as transformations of the data $\boldsymbol{Y}$ that maintain all relevant information about the parameter $\boldsymbol{\psi}$.

To evaluate if a statistic is sufficient, the following theorem can be used (Casella and Berger, 2001: p. 276).

Theorem 2.4.5. *(Factorization Theorem) Let $f(\boldsymbol{y};\boldsymbol{\psi})$ denote the joint distribution of a sample $\boldsymbol{Y}$ and let $T(\boldsymbol{Y})$ be a statistic of the data. Then $T(\boldsymbol{Y})$ is sufficient for $\boldsymbol{\psi}$ if and only if*

$$f(\boldsymbol{y};\boldsymbol{\psi}) = g(T(\boldsymbol{y});\boldsymbol{\psi})h(\boldsymbol{y}),$$

for some functions g and h, and for all sample points $\boldsymbol{y}$ and parameter points $\boldsymbol{\psi}$.

From this, it follows directly that $T(\boldsymbol{Y}) = \boldsymbol{V}^T\boldsymbol{Y}$ is a sufficient statistic for $\boldsymbol{\psi}$ in our model, that is, where the distribution of Y_i is from a regular exponential family and the link function is a canonical link. To see that this is true, we write

$$\begin{aligned}
f(\boldsymbol{Y};\boldsymbol{\psi},\boldsymbol{V}) &= \exp\left\{\sum_{i=1}^{n}\{\theta_i Y_i - A(\theta_i) + h(Y_i)\}\right\} \\
&= \exp\{\boldsymbol{\theta}^T\boldsymbol{Y} - A(\boldsymbol{\theta}) + h(\boldsymbol{Y})\} \\
&= \exp\{(\boldsymbol{V}\boldsymbol{\psi})^T\boldsymbol{Y} - \mathcal{A}(\boldsymbol{\psi}) + h(\boldsymbol{Y})\} \\
&= \exp\{\boldsymbol{\psi}^T\boldsymbol{V}^T\boldsymbol{Y} - \mathcal{A}(\boldsymbol{\psi})\}\exp\{h(\boldsymbol{Y})\} \\
&= \exp\{\boldsymbol{\psi}^T T(\boldsymbol{Y}) - \mathcal{A}(\boldsymbol{\psi})\}\exp\{h(\boldsymbol{Y})\} \\
&= g(T(\boldsymbol{Y});\boldsymbol{\psi})h(\boldsymbol{Y}),
\end{aligned}$$

where $\mathcal{A}$ is defined as in equation (2.8), and depends only on the parameter $\boldsymbol{\psi}$ and the known covariates $\boldsymbol{V}$.

Remark 2.4.6. The sufficient statistic, $T(\boldsymbol{Y}) = \boldsymbol{V}^T\boldsymbol{Y}$, will be complete as well in many situations, in particular for regular exponential families (Casella and Berger, 2001: p.288).

Thus we get a second interpretation of the score, namely that it is a centered sufficient statistic for the parameter vector $\boldsymbol{\psi}$, that is

$$\boldsymbol{S}(\boldsymbol{\psi}) = \boldsymbol{T} - E[\boldsymbol{T}], \tag{2.16}$$

for $\boldsymbol{T} = T(\boldsymbol{Y}) = \boldsymbol{V}^T\boldsymbol{Y}$.

Returning now to the hypothesis testing framework, we begin by introducing a simpler notation,

$$s(\boldsymbol{\psi}; \boldsymbol{Y}, \boldsymbol{V}) = \boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_\beta \\ U_\gamma \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \\ \boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \end{bmatrix},$$

where we write the expression for the score as in equation (2.15). It can be shown that the expected Fisher information matrix also gets a general expression

$$\mathcal{I}(\boldsymbol{\psi}) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \boldsymbol{I}_{\beta\gamma} \\ \boldsymbol{I}_{\gamma\beta} & I_{\gamma\gamma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} \end{bmatrix},$$

where W is a diagonal matrix with diagonal element i being the variance of Y_i, for $i = 1, \dots, n$.

Assume now that we want to compute the p-value of the right-tailed score test given by,

$$H_0 : \gamma \leq 0 \quad \text{v.s.} \quad H_1 : \gamma > 0.$$

Given a set of observations $(y_1, \boldsymbol{v}_1), \dots, (y_n, \boldsymbol{v}_n)$, the observed score becomes

$$u = \boldsymbol{Z}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0),$$

where $\hat{\boldsymbol{\mu}}_0$ is the expected value $E[\boldsymbol{Y}]$ computed under H_0, meaning that we use the maximum likelihood estimators for $\boldsymbol{\beta}$ and set $\gamma = 0$. The p-value of the score test is

$$p = P(U_\gamma \geq u \,|\, \boldsymbol{U}_\beta = \boldsymbol{0}; H_0). \tag{2.17}$$

However, since the score also can be written as in equation (2.16), we have that,

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_\beta \\ U_\gamma \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_\beta - E[\boldsymbol{T}_\beta] \\ T_\gamma - E[T_\gamma] \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_\beta - \boldsymbol{X}^T\boldsymbol{\mu} \\ T_\gamma - \boldsymbol{Z}^T\boldsymbol{\mu} \end{bmatrix},$$

since the expectation $E[\boldsymbol{T}] = E[\boldsymbol{V}^T\boldsymbol{Y}] = \boldsymbol{V}^T\boldsymbol{\mu}$. The variance of $\boldsymbol{U}$ is clearly equal to the variance of $\boldsymbol{T}$, since $\text{Var}(E[X]) = 0$ for all random variables X. For fixed parameters under H_0, $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\gamma = 0$, we have that $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}_0$.

Therefore, we get

$$\boldsymbol{U}_\beta = \boldsymbol{0} \implies \boldsymbol{T}_\beta - \boldsymbol{X}^T\hat{\boldsymbol{\mu}}_0 = 0 \implies \boldsymbol{T}_\beta = \boldsymbol{X}^T\hat{\boldsymbol{\mu}}_0,$$

where we have simply used the definition of $\boldsymbol{U}_\beta$ in terms of a sufficient statistic and the fact that $\boldsymbol{\mu}$ is known for fixed parameters.

Similarly, given $t_\gamma = \boldsymbol{Z}^T\boldsymbol{y}$ defined as the observed value of T_γ, we have

$$u = \boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T\hat{\boldsymbol{\mu}}_0 = t_\gamma - \boldsymbol{Z}^T\hat{\boldsymbol{\mu}}_0 \implies t_\gamma = u + \boldsymbol{Z}^T\hat{\boldsymbol{\mu}}_0.$$

Hence, the p-value of the score can also be computed with respect to the sufficient statistic $\boldsymbol{T} = [\boldsymbol{T}_\beta, T_\gamma]^T$, meaning

$$\begin{aligned} p &= P(U_\gamma \geq u \,|\, \boldsymbol{U}_\beta = \boldsymbol{0}; H_0) \\ &= P(U_\gamma + \boldsymbol{Z}^T\hat{\boldsymbol{\mu}}_0 \geq u + \boldsymbol{Z}^T\hat{\boldsymbol{\mu}}_0 \,|\, \boldsymbol{T}_\beta - \boldsymbol{X}^T\hat{\boldsymbol{\mu}}_0 = \boldsymbol{0}; H_0) \\ &= P(T_\gamma \geq t_\gamma \,|\, \boldsymbol{T}_\beta = \boldsymbol{X}^T\hat{\boldsymbol{\mu}}_0; H_0). \end{aligned} \tag{2.18}$$

Usually, $\boldsymbol{U}$ is assumed to be multivariate normally distributed, in which case $\boldsymbol{T}$ will also be multivariate normally distributed, and then the p-value of the score test can be easily computed from equation (2.17), or, equivalently, from equation (2.18). However, in this thesis we will not assume that the distribution of the score statistic has properly converged to a normal distribution, meaning that the p-value of the score test can not be computed in a straightforward manner. In chapter 3 and 4 we, therefore, present two alternative methods for computing the p-value of a score test, and in chapter 5 and 6 we compare these to the standard way of computing the p-value from a score test using simulated and real examples. It will be useful to consider p-value calculations both in terms of $\boldsymbol{U}$ (equation (2.17)) and in terms of T (equation (2.18)).

Chapter 3

Saddlepoint Approximation

Saddlepoint approximation is a technique for estimating the density and cumulative probability of a random variable. The method was first presented in a statistical setting by Daniels (1954). For this thesis, we will use saddlepoint approximation to estimate the p-values of a score test. However, we begin by establishing the relevant framework in a general setting. The results presented here concern univariate and continuous random variables. The material is based on Butler (2007), and we refer there for generalizations and further details.

3.1 Cumulant Generating Function

An important tool for constructing the saddlepoint approximation is the *cumulant generating function (CGF)*. This is defined as the logarithm of the *moment generating function (MGF)*.

Definition 3.1.1 (Moment generating function)**.** Let Y be a random variable with density $f(y)$ and support $\mathcal{Y}$. Then the moment generating function of Y is defined as

$$M_Y(s) = E(e^{sY}) = \begin{cases} \int\limits_{\mathcal{Y}} e^{sy} f(y)dy & \text{if } Y \text{ is continuous,} \\ \sum\limits_{k\in\mathcal{Y}} e^{sk} P(Y = k) & \text{if } Y \text{ is discrete.} \end{cases}$$

for $s \in (a, b) \subset \mathbb{R}$ such that the expression is well-defined. The n'th derivative evaluated in 0 is called the n'th *moment* of X, and it can be shown that

$$M_Y^{(n)}(0) = E(Y^n).$$

The cumulant generating function is defined as

$$K_Y(s) = \ln M_Y(s), \quad s \in (a, b).$$

The n'th derivative evaluated in 0 for the cumulant generating function is called the n'th *cumulant*. In particular, the first two cumulants are

$$\begin{aligned} \kappa_1 &= K_Y'(0) = E(Y) && \text{(Mean)} \\ \kappa_2 &= K_Y''(0) = \text{Var}(Y) && \text{(Variance)} \end{aligned}$$

The CGF K is a strictly convex function for any random variable Y, meaning $K_Y''(s) > 0$ for all $s \in (a, b)$, and the variance as the second cumulant is therefore well-defined. A proof can be found in Johnsen et al. (2023; Appendix C).

Properties

We consider three computational properties of the cumulant generating function, as well as one that relates to the regular exponential family.

Proposition 3.1.2. *Given an independent sample* $Y_1, \dots, Y_n$*, where each* Y_i *has CGF* $K_{Y_i}(s)$*, then* $Y = \sum_{i=1}^{n} Y_i$ *has CGF*

$$K_Y(s) = \sum_{i=1}^{n} K_{Y_i}(s).$$

Proof. By definition we have,

$$\begin{aligned} K_Y(s) &= \ln M_Y(s) \\ &= \ln E[e^{Ys}]. \end{aligned}$$

Since Y is the sum of independent random variables Y_i we get,

$$\begin{aligned} \ln E[e^{Ys}] &= \ln E[e^{s(Y_1+\ldots+Y_n)}] \\ &= \ln E[e^{sY_1} \ldots e^{sY_n}] \\ &= \ln E[e^{sY_1}] + \ldots + \ln E[e^{sY_n}] \\ &= K_{Y_1}(s) + \ldots + K_{Y_n}(s) \\ &= \sum_{i=1}^{n} K_{Y_i}(s). \end{aligned}$$

□

Proposition 3.1.3. *Given a random variable Y with CGF $K_Y(s)$, then, for some $b \in \mathbb{R}$, bY has CGF*

$$K_{bY}(s) = K_Y(sb).$$

Proof. Once again we use the definition of CGF to compute,

$$\begin{aligned} K_{bY}(s) &= \ln M_{bY}(s) \\ &= \ln E[e^{s(bY)}] \\ &= \ln E[e^{(sb)Y}] \\ &= \ln M_Y(sb) \\ &= K_Y(sb). \end{aligned}$$

□

Proposition 3.1.4. *Given a random variable Y with CGF $K_Y(s)$, then, for some $b \in \mathbb{R}$, $b + Y$ has CGF*

$$K_{b+Y}(s) = sb + K_Y(s).$$

Proof. Again, by writing out the definition we get,

$$\begin{aligned} K_{b+Y}(s) &= \ln M_{b+Y}(s) \\ &= \ln E[e^{s(b+Y)}] \\ &= \ln E[e^{sb+sY}] \\ &= \ln E[e^{sb}] + \ln E[e^{sY}] \\ &= sb + \ln M_Y(s) \\ &= sb + K_Y(s). \end{aligned}$$

□

CGF for a regular exponential family

Let Y_i be from a regular exponential family and assume we are working with a generalized linear model with a canonical link, where x_i is a known covariate and β is a (scalar) parameter[1]. In other words, $\theta_i = \eta_i = x_i\beta$. Then

$$f(y_i; \beta, x_i) = \exp\{\beta t(y_i) - \mathcal{A}_i(\beta) + h(y_i)\},$$

where $t(y_i) = x_i y_i$, and $\mathcal{A}_i(\beta) = A(x_i\beta)$ as before. The function $\mathcal{A}_i$ is sometimes called the *cumulant function* or the *log-partition function* of the probability distribution.

Proposition 3.1.5. *The cumulant generating function of the sufficient statistic $T_i = T(Y_i) = x_iY_i$ is given by*

$$K_{T_i}(s; \beta, x_i) = \mathcal{A}_i(\beta + s) - \mathcal{A}_i(\beta).$$

[1] This result generalizes to parameter vectors as well, but we only prove it for a univariate parameter.

Proof. First, we note that

$$\begin{aligned} 1 &= \int_{\mathcal{Y}} \exp\{\beta t(y_i) - \mathcal{A}_i(\beta)\} \exp\{h(y_i)\} dy_i \\ \implies \exp\{\mathcal{A}_i(\beta)\} &= \int_{\mathcal{Y}} \exp\{\beta t(y_i)\} \exp\{h(y_i)\} dy_i. \end{aligned} \tag{3.1}$$

From the definition of the moment generating function, we then get

$$\begin{aligned} M_{T_i}(s;\beta,x_i) &= E_{Y_i}[e^{sT(Y_i)}] \\ &= \int_{\mathcal{Y}} \exp\{st(y_i)\} \exp\{\beta t(y_i) - \mathcal{A}_i(\beta)\} \exp\{h(y_i)\} dy_i \\ &= \exp\{-\mathcal{A}_i(\beta)\} \int_{\mathcal{Y}} \exp\{st(y_i) + \beta t(y_i)\} \exp\{h(y_i)\} dy_i \\ &= \exp\{-\mathcal{A}_i(\beta)\} \int_{\mathcal{Y}} \exp\{(s+\beta) t(y_i)\} \exp\{h(y_i)\} dy_i \\ &= \exp\{\mathcal{A}_i(\beta + s) - \mathcal{A}_i(\beta)\}, \end{aligned}$$

where the last equality comes from what we showed in equation (3.1). Since the cumulant generating function is defined as the logarithm of the MGF, it follows easily that

$$K_{T_i}(s;\beta,x_i) = \mathcal{A}_i(\beta + s) - \mathcal{A}_i(\beta),$$

as desired. □

From proposition 3.1.2 it follows that the sufficient statistic $T = T(\boldsymbol{Y}) = \boldsymbol{X}^T\boldsymbol{Y} = \sum\limits_{i=1}^{n} x_i Y_i$ has cumulant generating function

$$K_T(s;\beta,\boldsymbol{X}) = \sum_{i=1}^{n} K_{T_i}(s;\beta,x_i) = \sum_{i=1}^{n} \mathcal{A}_i(\beta + s) - \mathcal{A}_i(\beta).$$

Define the function $\mathcal{A}(-) = \sum\limits_{i=1}^{n} \mathcal{A}_i(-)$, then

$$K_T(s;\beta,\boldsymbol{X}) = \mathcal{A}(\beta + s) - \mathcal{A}(\beta). \tag{3.2}$$

From the properties of the cumulant generating function we know that the first and second cumulant is the mean and variance respectively. Thus we have that

$$\begin{aligned} E[T(\boldsymbol{Y})] &= K_T'(0;\beta,\boldsymbol{X}) = \mathcal{A}'(\beta) = \boldsymbol{X}^T\boldsymbol{\mu} \\ \text{Var}[T(\boldsymbol{Y})] &= K_T''(0;\beta,\boldsymbol{X}) = \mathcal{A}''(\beta) = \mathcal{I}(\beta). \end{aligned}$$

We note also here that $\mathcal{A}$ must be a strictly convex function when evaluated in s, as the CGF is a strictly convex function, and $\mathcal{A}(\beta)$ in equation (3.2) is merely a constant shift with respect to s.

Example 3.1.6. Consider $Y_i \sim \text{Pois}(\mu_i)$ from a Poisson regression model with canonical link, and linear predictor $\eta_i = x_i\beta$. The cumulant generating function for Y_i, a Poisson distributed random variable, is known and given by

$$K_{Y_i}(s) = \mu_i(\exp(s) - 1).$$

The score vector corresponding to this model is given by

$$U = \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}),$$

and since the score is also a random variable for an unrealized response vector $\boldsymbol{Y}$, it too has a cumulant generating function, which can be computed as

$$\begin{aligned} K_U(s) &= \sum_{i=1}^{n} K_{Y_i}(x_i s) - x_i\mu_i s \\ &= \sum_{i=1}^{n} \mu_i(\exp(x_i s) - 1) - s x_i \mu_i. \end{aligned}$$

This follows directly from the computational properties shown earlier. However, since $U = T - E[T]$, we can also compute the CGF of the score by using proposition 3.1.5. We note that $\mathcal{A}_i(\beta) = \exp(x_i\beta)$ from how $\mathcal{A}$ was defined in equation (2.8), and the fact that $A(\theta_i) = \exp(\theta_i)$, as showed in subsection 2.2.3. Hence,

$$\begin{aligned} K_U(s) &= \sum_{i=1}^{n} K_{T_i}(s) - sE[T_i] \\ &= \sum_{i=1}^{n} \mathcal{A}_i(\beta + s) - \mathcal{A}_i(\beta) - sK'_{T_i}(0) \\ &= \sum_{i=1}^{n} \exp(x_i\beta + x_i s) - \exp(x_i\beta) - s\mathcal{A}'_i(\beta + s) \\ &= \sum_{i=1}^{n} \exp(x_i\beta)(\exp(x_i s) - 1) - s x_i \exp(x_i\beta) \\ &= \sum_{i=1}^{n} \mu_i(\exp(x_i s) - 1) - s x_i \mu_i, \end{aligned}$$

where the last equality follows from the fact that $\exp(x_i\beta) = \exp(\theta_i) = \mu_i$, which was showed in subsection 2.2.3. We see that both methods yield the same CGF for U, as expected.

3.2 Saddlepoint approximation

In this section, we aim to motivate and explain the saddlepoint approximation. We will present the saddlepoint approximation for a univariate density, mainly for illustrative purposes of how the method works. As mentioned in at the beginning of this chapter, the saddlepoint approximation is a general method that works in multiple frameworks. Therefore, we do not assume anything related to generalized linear models, unless explicitly stated.

3.2.1 Saddlepoint density

Saddlepoint approximation can be used to estimate the density, $f(y)$, of a random variable Y that is otherwise intractable. This is done by utilizing the cumulant generating function $K_Y(s)$. The saddlepoint approximation to the density of Y, denoted $\hat{f}$, is given by

$$\hat{f}(y) = \frac{1}{\sqrt{2\pi K_Y''(\hat{s})}} \exp(K_Y(\hat{s}) - \hat{s}y), \tag{3.3}$$

where $\hat{s} = \hat{s}(y)$ is the unique solution to

$$K_Y'(\hat{s}) = y \tag{3.4}$$

The point $\hat{s}$ is referred to as the *saddlepoint* associated with value y, and equation (3.4) is called the *saddlepoint equation*. The name saddlepoint approximation comes from the fact that the function

$$s \mapsto K_Y(s) - sy$$

has a saddlepoint at $\hat{s}(y)$, where $K_Y'(\hat{s}) = y$.

Example 3.2.1. It can be easily shown that the saddlepoint density is exact for a random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$. The CGF of a normally distributed variable is

$$K_Y(s) = \mu s + \frac{\sigma^2 s^2}{2},$$

and the saddlepoint then becomes

$$K_Y'(\hat{s}) = \mu + \sigma^2 \hat{s} = y \implies \hat{s} = \frac{y - \mu}{\sigma^2}.$$

Additionally,

$$K_Y''(\hat{s}) = \sigma^2.$$

By simply inserting this into equation (3.3), we get

$$\begin{aligned}
\hat{f}(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \mu \frac{y-\mu}{\sigma^2} + \frac{\sigma^2 \left(\frac{y-\mu}{\sigma^2}\right)^2}{2} - \frac{y-\mu}{\sigma^2} y \right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 - \frac{\mu^2 - y\mu - y\mu + y^2}{\sigma^2} \right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 - \left(\frac{y-\mu}{\sigma}\right)^2 \right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 \right\},
\end{aligned}$$

which we recognize as the density of a normal random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$.

The saddlepoint approximation can be thought of as a refinement of the central limit approximation with higher-order expansion terms (Butler, 2007: p. 60-61), which leads to greater accuracy than what we would expect from a normal approximation. Furthermore, the saddlepoint approximation uses more information about the distribution. Whereas the central limit approximation only uses the first two cumulants (the mean and variance of the distribution), the saddlepoint approximation depends on the cumulant generating function through a saddlepoint in every evaluation point $y \in \mathcal{Y}$.

Comparing the accuracy, the central limit converges with rate $\mathcal{O}(n^{-1/2})$, and is more accurate near the mean of the distribution. The saddlepoint approximation converges with rate $\mathcal{O}(n^{-1})$, and in some cases up to $\mathcal{O}(n^{-3/2})$, and achieves almost uniformly well approximations over the entire distribution.

3.2.2 Saddlepoint distribution

The saddlepoint method can easily be applied to a cumulative distribution. Let Y be a continuous random variable and let $K_Y(s)$ denote the cumulant generating function of Y. Then there exists a saddlepoint approximation for the distribution of Y, introduced by Lugannani and Rice (1980), which is given by

$$\hat{P}(Y \leq y) = \Phi(\hat{\omega}) + \phi(\hat{\omega})(1/\hat{\omega} - 1/\hat{u}), \quad y \neq \mu \tag{3.5}$$

where Φ and ϕ are the normal distribution and normal density, respectively,

$$\hat{\omega} = \text{sgn}(\hat{s})\sqrt{2(\hat{s}y - K_Y(\hat{s}))}, \quad \hat{u} = \hat{s}\sqrt{K_Y''(\hat{s})},$$

and $\hat{s}$ is the saddlepoint as defined in equation (3.4).

From this expression, the earlier mentioned ties to the central limit result become more visible. Indeed, this approximation will also be exact for a normally distributed random variable, as it was for the density which we saw in example 3.2.1. Furthermore, we note the singularity in μ. This follows from the fact that $\hat{s} = 0$ when $y = \mu$, and hence $1/\hat{u}$ will be undefined. However, a limiting value can be used in the singularity and is shown derived in Butler (2007; p. 68-69).

Remark 3.2.2. We have so far assumed that Y is continuous and univariate. Extensions to discrete random variables or random vectors are not very difficult for densities. However, for the cumulative distribution, this is not straightforward. For discrete random variables, the trouble arises from the fact that saddlepoint approximation is a continuous approximation. In Butler (2007), multiple continuity corrected expressions are presented. However, we will only consider the continuous expression which is given in equation (3.5). The continuous approximation applied to a discrete random variable with unit step length will correspond to a so-called mid-p-value (Butler, 2007: p.188).

3.2.3 Double saddlepoint distribution

Consider now the saddlepoint approximation for the conditional distribution of a continuous random variable Y given some random vector $\boldsymbol{X} = \boldsymbol{x}$. This cumulative conditional distribution has a saddlepoint approximation derived by Skovgaard (1987), which is defined as

$$\hat{P}(Y \leq y | \boldsymbol{X} = \boldsymbol{x}) = \Phi(\hat{\omega}) + \phi(\hat{\omega})(1/\hat{\omega} - 1/\hat{u}), \quad \hat{s} \neq 0, \tag{3.6}$$

where Φ and ϕ are the normal distribution and normal density, as before, and

$$\begin{aligned} \hat{\omega} &= \operatorname{sgn}(\hat{s})\sqrt{2([K(\hat{\boldsymbol{r}}_0, 0) - \hat{\boldsymbol{r}}_0^T \boldsymbol{x}] - [K(\hat{\boldsymbol{r}}, \hat{s}) - \hat{\boldsymbol{r}}^T \boldsymbol{x} - \hat{s}y])}, \\ \hat{u} &= \hat{s}\sqrt{\frac{|K''(\hat{\boldsymbol{r}}, \hat{s})|}{|K''_{rr}(\hat{\boldsymbol{r}}_0, 0)|}}. \end{aligned}$$

The function $K(\boldsymbol{r}, s)$ is the joint cumulant generating function of $(\boldsymbol{X}, Y)$, and K''_{rr} is the twice derivative only with respect to $\boldsymbol{r}$ of the joint CGF. The saddlepoints are defined as the points $(\hat{\boldsymbol{r}}, \hat{s})$ and $(\hat{\boldsymbol{r}}_0, 0)$ which satisfy

$$\begin{aligned} K'(\hat{\boldsymbol{r}}, \hat{s}) &= (\boldsymbol{x}, y) \\ K'(\hat{\boldsymbol{r}}_0, 0) = K_X(\hat{\boldsymbol{r}}_0) &= \boldsymbol{x}, \end{aligned}$$

respectively. Note that the joint CGF evaluated in $K(\boldsymbol{r}, 0)$ is equal to the CGF of $\boldsymbol{X}$, $K_X(\boldsymbol{r})$ (Butler, 2007: p. 108).

Example 3.2.3. In this example, we illustrate how the double saddlepoint approximation can be used to compute the p-value of a score test in a simple Poisson regression model.

Consider responses $Y_i \sim \text{Pois}(\mu_i)$ for $i = 1, \ldots, n$ in a Poisson regression model with canonical link function and linear predictor $\eta_i = x_i\beta + z_i\gamma$. This is a simple model with one nuisance parameter and one parameter of interest, where we are interested in testing the null-hypothesis $H_0 : \gamma \leq 0$ against $H_1 : \gamma > 0$ using a score test. Note that the calculations will be practically identical for a vector of nuisance parameters $\boldsymbol{\beta}$.

The score test statistic is given by

$$\boldsymbol{U} = \begin{bmatrix} U_\beta \\ U_\gamma \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \\ \boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \end{bmatrix},$$

and the joint CGF of (U_β, U_γ) is given by

$$\begin{aligned} K_U(r, s) &= \sum_{i=1}^n K_{Y_i}(x_i r + z_i s) + \mu_i(x_i r + z_i s) \\ &= \sum_{i=1}^n K_{Y_i}(\boldsymbol{v}_i^T \boldsymbol{s}) + \mu_i(\boldsymbol{v}_i^T \boldsymbol{s}), \end{aligned}$$

where $\boldsymbol{v}_i = [x_i, z_i]^T$ and $\boldsymbol{s} = [r, s]^T$. Since the double saddlepoint method is defined in terms of the left-tail probability, and we are interested in performing a right-tail test, we compute

$$P(U_\gamma \geq u | U_\beta = 0; H_0) \approx 1 - \hat{P}(U_\gamma \leq u | U_\beta = 0; H_0),$$

where the probability distribution is approximated in terms of the Skovgaard approximation from equation (3.6).

We must first compute the saddlepoints, for which we need the gradient of $K_U(\boldsymbol{s})$

$$\nabla K_U(\boldsymbol{s}) = \sum_{i=1}^{n} \mu_i \exp(\boldsymbol{v}_i \boldsymbol{s}) \boldsymbol{v}_i - \boldsymbol{v}_i \mu_i.$$

The saddlepoints are determined by,

$$\begin{aligned} \nabla K_U(\hat{r}, \hat{s}) &= (0, u) \\ \nabla_r K_U(\hat{r}_0, 0) &= 0, \end{aligned} \tag{3.7}$$

where u will be the observed value of the score U_γ. It is easily seen that the latter saddlepoint $\hat{r}_0$ must be zero as $\nabla_r K_U(\mathbf{0}) = 0$, and this will be the unique solution to the saddlepoint equation as the CGF is a convex function. The saddlepoint $(\hat{r}, \hat{s})$ is found by numerically solving the two-dimensional, non-linear system of equations given in equation (3.7).

The expression also uses the twice derivative of the CGF, the Hessian, which in this case is given by

$$H(\boldsymbol{s}) = \sum_{i=1}^{n} \mu_i \exp(\boldsymbol{v}_i \boldsymbol{s}) \boldsymbol{v}_i^T \boldsymbol{v}_i,$$

and

$$H_{rr}(r) = \sum_{i=1}^{n} \mu_i \exp(x_i r) x_i^2.$$

The p-value of the score test is therefore given by

$$p = 1 - [\Phi(\hat{\omega}) + \phi(\hat{\omega})(1/\hat{\omega} - 1/\hat{u})], \quad \hat{s} \neq 0,$$

where

$$\begin{aligned} \hat{\omega} &= \text{sgn}(\hat{s}) \sqrt{2[\hat{s}u - K(\hat{r}, \hat{s})]}, \\ \hat{u} &= \hat{s} \sqrt{\frac{|H(\hat{r}, \hat{s})|}{|H_{rr}(0)|}}. \end{aligned}$$

3.3 Saddlepoint approximations in a regular exponential family

In Butler (2007; Ch. 5), it is shown that the saddlepoint approximation takes a rather simple form for sufficient statistics in a regular exponential family. We extend their ideas to score statistics, and show how a saddlepoint approximated distribution of the score can be easily computed using output that is usually provided when fitting a generalized linear model in statistical software such as R Team (2021).

3.3.1 Distribution function

Proposition 3.3.1. *Assume that $(Y_1, x_1), \dots, (Y_n, x_n)$ is a random sample from a regular exponential family with log-likelihood as in equation (2.7), and let the linear predictor $\eta_i = x_i\beta$ be linked to the response with a canonical link function. Then the saddlepoint approximation of $P(T \leq t)$ for a sufficient statistic $T = \boldsymbol{X}^T\boldsymbol{Y}$ and observed value $t = \boldsymbol{X}^T\boldsymbol{y}$ is given by*

$$\hat{P}(T \leq t) = \Phi(\hat{\omega}) + \phi(\hat{\omega})(1/\hat{\omega} - 1/\hat{u}), \quad t \neq E[T], \tag{3.8}$$

with

$$\hat{\omega} = \operatorname{sgn}(\hat{\beta} - \beta)\sqrt{-2\ln\frac{\mathcal{L}(\beta)}{\mathcal{L}(\hat{\beta})}}, \quad \hat{u} = (\hat{\beta} - \beta)\sqrt{\mathcal{I}(\hat{\beta})},$$

where Φ and ϕ denote the standard normal distribution and density, respectively, $\hat{\beta}$ is the maximum likelihood estimator of the parameter β, and $\mathcal{I}$ denotes the expected Fisher information matrix, which is the variance of T.

In Butler (2007), this relationship between maximum likelihood estimates and the saddlepoint approximation is only shown for densities, but an analog proof for distributions is easily constructed using a similar approach.

Proof. As shown in proposition 3.1.5, the cumulant generating function of T is given by

$$K_T(s; \beta, \boldsymbol{X}) = \mathcal{A}(\beta + s) - \mathcal{A}(\beta).$$

The saddlepoint equation (3.4) is defined as

$$K_T'(\hat{s}; \beta, \boldsymbol{X}) = \mathcal{A}'(\beta + \hat{s}) = t.$$

However, from subsection 2.3.2, we know that the maximum likelihood estimator of β is uniquely given by

$$\left.\frac{\partial \ell(\beta; \boldsymbol{y}, \boldsymbol{X})}{\partial \beta}\right|_{\hat{\beta}} = 0$$
$$\implies t - \mathcal{A}'(\hat{\beta}) = 0$$
$$\implies \mathcal{A}'(\hat{\beta}) = t.$$

Since the point that satisfy $\mathcal{A}'(-) = t$ is unique, we must have

$$\hat{\beta} = \beta + \hat{s}.$$

Combining what we have noted so far, it is possible to rewrite the saddlepoint approximation in equation (3.5) for the special case $P(T \leq t)$. This is done as follows,

$$\begin{aligned}
\hat{\omega} &= \text{sgn}(\hat{s})\sqrt{2(\hat{s}t - K(\hat{s}))} \\
&= \text{sgn}(\hat{\beta} - \beta)\sqrt{2[(\hat{\beta} - \beta)t - (\mathcal{A}(\beta + \hat{\beta} - \beta) - \mathcal{A}(\beta))]} \\
&= \text{sgn}(\hat{\beta} - \beta)\sqrt{2[\hat{\beta}t - \beta t - (\mathcal{A}(\hat{\beta}) - \mathcal{A}(\beta))]} \\
&= \text{sgn}(\hat{\beta} - \beta)\sqrt{2[\hat{\beta}t - \mathcal{A}(\hat{\beta}) - (\beta t - \mathcal{A}(\beta))]} \\
&= \text{sgn}(\hat{\beta} - \beta)\sqrt{2[\ell(\hat{\beta}; \boldsymbol{y}) - \ell(\beta; \boldsymbol{y})]} \\
&= \text{sgn}(\hat{\beta} - \beta)\sqrt{-2\ln\frac{\mathcal{L}(\beta)}{\mathcal{L}(\hat{\beta})}}.
\end{aligned}$$

and

$$\begin{aligned}
\hat{u} &= \hat{s}\sqrt{K''(\hat{s})} \\
&= (\hat{\beta} - \beta)\sqrt{\mathcal{A}''(\beta + \hat{\beta} - \beta)} \\
&= (\hat{\beta} - \beta)\sqrt{\mathcal{A}''(\hat{\beta})} \\
&= (\hat{\beta} - \beta)\sqrt{\mathcal{I}(\hat{\beta})}.
\end{aligned}$$

The approximation therefore becomes

$$\hat{P}(T \leq t) = \Phi(\hat{\omega}) + \phi(\hat{\omega})(1/\hat{\omega} - 1/\hat{u}), \quad t \neq E[T],$$

where

$$\hat{\omega} = \text{sgn}(\hat{\beta} - \beta)\sqrt{-2\ln\frac{\mathcal{L}(\beta)}{\mathcal{L}(\hat{\beta})}}, \quad \hat{u} = (\hat{\beta} - \beta)\sqrt{\mathcal{I}(\hat{\beta})}.$$

□

Corollary 3.3.1.1. *Under the assumptions from proposition 3.3.1, and given the score $U = T - E[T]$, the saddlepoint approximation of $P(U \leq u)$ will be the same as for $P(T \leq t)$ as given in equation (3.8).*

Proof. This follows directly from the justification we did in equation (2.18). However, we can also show it by writing out a similar proof as we did above, so we do that in order to establish this relationship more firmly.

We have that the CGF of U is given by

$$\begin{aligned} K_U(s;\beta,\boldsymbol{X}) &= K_T(s) - K_T'(0)s \\ &= \mathcal{A}(\beta+s) - \mathcal{A}(\beta) - \mathcal{A}'(\beta)s. \end{aligned}$$

Hence the saddlepoint equation becomes

$$K_U'(\hat{s};\beta,\boldsymbol{X}) = \mathcal{A}'(\beta+\hat{s}) - \mathcal{A}'(\beta) = u,$$

which is equivalent to

$$\mathcal{A}'(\beta+\hat{s}) = u + \mathcal{A}'(\beta) = t,$$

since $u = t - \mathcal{A}'(\beta)$.

Furthermore, we have also that the maximum likelihood estimator $\hat{\beta}$ is uniquely determined by

$$u = t - \mathcal{A}'(\hat{\beta}) = 0 \implies \mathcal{A}'(\hat{\beta}) = t.$$

Again, we get the relationship

$$\hat{\beta} = \beta + \hat{s},$$

and we compute as before

$$\begin{aligned} \hat{\omega} &= \operatorname{sgn}(\hat{s})\sqrt{2(\hat{s}u - K_U(\hat{s}))} \\ &= \operatorname{sgn}(\hat{\beta}-\beta)\sqrt{2[(\hat{\beta}-\beta)u - (\mathcal{A}(\beta+\hat{\beta}-\beta) - \mathcal{A}(\beta) - \mathcal{A}'(\beta)(\hat{\beta}-\beta))]} \\ &= \operatorname{sgn}(\hat{\beta}-\beta)\sqrt{2[\hat{\beta}u - \beta u - \mathcal{A}(\hat{\beta}) + \mathcal{A}(\beta) + \boldsymbol{x}^T\boldsymbol{\mu}(\hat{\beta}-\beta)]} \\ &= \operatorname{sgn}(\hat{\beta}-\beta)\sqrt{2[\hat{\beta}u + \boldsymbol{x}^T\boldsymbol{\mu}\hat{\beta} - \mathcal{A}(\hat{\beta}) - (\beta u + \boldsymbol{x}^T\boldsymbol{\mu}\beta - \mathcal{A}(\beta))]} \\ &= \operatorname{sgn}(\hat{\beta}-\beta)\sqrt{2[\hat{\beta}(\boldsymbol{x}^T\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\mu} + \boldsymbol{x}^T\boldsymbol{\mu}) - \mathcal{A}(\hat{\beta}) - (\beta(\boldsymbol{x}^T\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\mu} + \boldsymbol{x}^T\boldsymbol{\mu}) - \mathcal{A}(\beta))]} \\ &= \operatorname{sgn}(\hat{\beta}-\beta)\sqrt{2[\hat{\beta}t - \mathcal{A}(\hat{\beta}) - (\beta t - \mathcal{A}(\beta))]} \\ &= \operatorname{sgn}(\hat{\beta}-\beta)\sqrt{2[\ell(\hat{\beta};\boldsymbol{y}) - \ell(\beta;\boldsymbol{y})]} \\ &= \operatorname{sgn}(\hat{\beta}-\beta)\sqrt{-2\ln\frac{\mathcal{L}(\beta)}{\mathcal{L}(\hat{\beta})}}. \end{aligned}$$

and

$$\begin{aligned} \hat{u} &= \hat{s}\sqrt{K_U''(\hat{s})} \\ &= (\hat{\beta}-\beta)\sqrt{\mathcal{A}''(\beta+\hat{\beta}-\beta)} \\ &= (\hat{\beta}-\beta)\sqrt{\mathcal{A}''(\hat{\beta})} \\ &= (\hat{\beta}-\beta)\sqrt{\mathcal{I}(\hat{\beta})}. \end{aligned}$$

□

Note that the distribution only depends on the parameter and its maximum likelihood estimate. Thus, it makes sense that the saddlepoint approximation is the same, as both U and T are sufficient for the parameters, and therefore no information is lost when moving between these two.

We note also that both the saddlepoint $\hat{s}$ and the maximum likelihood estimators depend on the observed point u (or t), and that the saddlepoint approximation only gives us the probability $P(U \leq u)$ (or $P(T \leq t)$) for a specific, observed point, as opposed to the cumulative distribution function over the entire domain of U (T).

3.3.2 Double saddlepoint distribution in a regular exponential family

The Skovgaard approximation can be rewritten in a similar manner as we did above. Given a random sample $(Y_1, \boldsymbol{v}_1), \dots, (Y_n, \boldsymbol{v}_n)$ from a regular exponential family, with known covariate matrix $\boldsymbol{V} = [\boldsymbol{X}, \boldsymbol{Z}]$ and parameter vector $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$ consisting of a parameter of interest and a vector of nuisance parameters. Then the canonical parameter and the linear predictor are related by

$$\boldsymbol{\theta} = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{Z}.$$

The joint distribution of $y_1, \dots, y_n$ can then be written as

$$\begin{aligned} f(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{X}, \gamma, \boldsymbol{Z}) &= \exp\{\boldsymbol{\theta}^T\boldsymbol{y} - A(\boldsymbol{\theta}) + h(\boldsymbol{y})\} \\ &= \exp\{(\boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{Z})^T\boldsymbol{y} - \mathcal{A}(\boldsymbol{\beta}, \gamma) + h(\boldsymbol{y})\} \\ &= \exp\{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \gamma\boldsymbol{Z}^T\boldsymbol{y} - \mathcal{A}(\boldsymbol{\beta}, \gamma) + h(\boldsymbol{y})\} \\ &= \exp\{\boldsymbol{\beta}^T\boldsymbol{t}_\beta + \gamma t_\gamma - \mathcal{A}(\boldsymbol{\beta}, \gamma) + h(\boldsymbol{y})\}. \end{aligned}$$

We have that $\boldsymbol{T} = [\boldsymbol{T}_\beta, T_\gamma]^T = [\boldsymbol{X}^T\boldsymbol{Y}, \boldsymbol{Z}^T\boldsymbol{Y}]^T$ is a sufficient statistic for the parameter vector $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$. Then the joint CGF of $\boldsymbol{T} = [\boldsymbol{T}_\beta, T_\gamma]^T$ is given by (Butler, 2007: p.162)

$$K_T(\boldsymbol{r}, s; \boldsymbol{\psi}, \boldsymbol{V}) = \mathcal{A}(\boldsymbol{r} + \boldsymbol{\beta}, s + \gamma) - \mathcal{A}(\boldsymbol{\beta}, \gamma).$$

Similar arguments as was done in subsection 3.3 yields an alternative form of the double saddlepoint approximation for the distribution $P(T_\gamma \leq t_\gamma | \boldsymbol{T}_\beta = \boldsymbol{t}_\beta)$, for an observation $(y_1, \boldsymbol{v}_1), \dots, (y_n, \boldsymbol{v}_n)$ yielding $\boldsymbol{t} = [\boldsymbol{t}_\beta, t_\gamma]^T$ as the observed value of $\boldsymbol{T}$.

Letting

$$\mathcal{I}(\boldsymbol{\psi}) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \boldsymbol{I}_{\beta\gamma} \\ \boldsymbol{I}_{\gamma\beta} & I_{\gamma\gamma} \end{bmatrix}$$

be the expected Fisher information matrix, and let $\hat{\boldsymbol{\beta}}, \hat{\gamma}$ be the maximum likelihood estimators of $\boldsymbol{\beta}, \gamma$, respectively. Lastly, denote by $\hat{\boldsymbol{\beta}}_\gamma$ the maximum likelihood estimator of $\boldsymbol{\beta}$ with γ fixed at its true parameter value. Then

$$\hat{P}(T_\gamma \leq t_\gamma | \boldsymbol{T}_\beta = \boldsymbol{t}_\beta) = \Phi(\hat{\omega}) + \phi(\hat{\omega}) \left(\frac{1}{\hat{\omega}} - \frac{1}{\hat{u}} \right), \quad t_\gamma \neq E[T_\gamma],$$

where

$$\hat{\omega} = \operatorname{sgn}(\hat{\gamma} - \gamma)\sqrt{-2\ln\frac{\mathcal{L}(\hat{\beta}_\gamma)}{\mathcal{L}(\hat{\beta}, \hat{\gamma})}}$$
$$\hat{u} = (\hat{\gamma} - \gamma)\sqrt{\frac{|\mathcal{I}(\hat{\beta}, \hat{\gamma})|}{|\boldsymbol{I}_{\beta,\beta}(\hat{\beta}_\gamma)|}}.$$

Equivalently, we get that the score statistic $\boldsymbol{U} = [\boldsymbol{U}_\beta, U_\gamma]^T$ has a double saddlepoint approximated, conditional distribution given by

$$\hat{P}(U_\gamma \leq u | \boldsymbol{U}_\beta = 0) = \Phi(\hat{\omega}) + \phi(\hat{\omega})\left(\frac{1}{\hat{\omega}} - \frac{1}{\hat{u}}\right), \quad u \neq 0, \tag{3.9}$$

where $\hat{\omega}$ and $\hat{u}$ are as above.

We now consider an example of how this method can be used to compute the p-value of a score test. Here, we only explain the procedure, while implemented examples are considered in chapter 5 and 6. Keep in mind that this will correspond to what we did in example 4.3.1, only with a different expression.

Example 3.3.2. Assume that we have a set of observations $(y_1, \boldsymbol{v}_1), \ldots, (y_n, \boldsymbol{v}_n)$, with each y_i being a realization of a random variable $Y_i \sim \text{Pois}(\mu_i)$ for $i = 1, \ldots, n$. Let the linear predictor be given by $\eta_i = \boldsymbol{v}_i^T \boldsymbol{\psi} = x_i\beta + z_i\gamma$, and linked to the response through a canonical link function. We are interested in testing the null-hypothesis $H_0 : \gamma \leq 0$ against $H_1 : \gamma > 0$ using a score test.

We use statistical software such as R to fit two models: one under the null hypothesis and one for the alternative hypothesis.

In order to compute

$$\hat{\omega} = \operatorname{sgn}(\hat{\gamma} - \gamma)\sqrt{-2\ln\frac{\mathcal{L}(\hat{\beta}_\gamma)}{\mathcal{L}(\hat{\beta}, \hat{\gamma})}},$$

we use the log-likelihood $\mathcal{L}(\hat{\beta}, \hat{\gamma})$ from the full model, and the log-likelihood $\mathcal{L}(\hat{\beta}_\gamma)$ from the null-model. The parameter value of γ is 0, as we estimate the distribution under H_0, and $\hat{\gamma}$ is the MLE of γ from the full model.

Furthermore,

$$\hat{u} = (\hat{\gamma} - \gamma)\sqrt{\frac{|\mathcal{I}(\hat{\beta}, \hat{\gamma})|}{|I_{\beta,\beta}(\hat{\beta}_\gamma)|}},$$

is computed by inserting the expected Fisher information matrix $\mathcal{I}(\hat{\beta}, \hat{\gamma})$ from the full model, and the expected Fisher information matrix $I_{\beta,\beta}(\hat{\beta}_\gamma)$ from the null-model.

Lastly, we plug $\hat{\omega}$ and $\hat{u}$ into the expression

$$p = 1 - \hat{P}(U_\gamma \leq u \,|\, U_\beta = 0) = 1 - [\Phi(\hat{\omega}) + \phi(\hat{\omega}) \left(\frac{1}{\hat{\omega}} - \frac{1}{\hat{u}}\right)], \quad u \neq 0,$$

to get the p-value of the right-tailed score test for a given observation u, and the corresponding maximum likelihood estimates.

We note the lack of reliance upon the observed score in equation (3.9). Consider if we, instead of evaluating a p-value against a significance level α, wanted to specify our test in terms of a critical value c such that $P(U \geq c \,|\, U_\beta = 0) = \alpha$ and we reject all observations $u > c$. It is not clear how we could use the method from equation (3.9) to determine the critical value c, as the method is fully specified in terms of the maximum likelihood estimates, $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$, for a given observation $(y_1, \boldsymbol{v}_1), \ldots, (y_n, \boldsymbol{v}_n)$. This goes to show that we are not approximating the distribution of U as a function $F(u) \approx \hat{P}(U_\gamma \leq u \,|\, U_\beta = 0)$, but rather that we are computing a p-value of a score test, and that this p-value would correspond to $p = 1 - F(u)$, for an observation $(y_1, \boldsymbol{v}_1), \ldots, (y_n, \boldsymbol{v}_n)$ yielding the observed score u, or, equivalently, the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$.

Chapter 4

Effective Score

We will now leave saddlepoint approximations for a moment and instead consider something we will call the *effective score*. As before, let the score vector be given by

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_\beta \\ U_\gamma \end{bmatrix},$$

where the parameter vector $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$ consists of a vector of nuisance parameters $\boldsymbol{\beta}$, and a scalar parameter of interest γ.

In the paper by Johnsen et al. (2023), the double saddlepoint approximated score test as discussed in chapter 3 is compared to a method presented in Dey et al. (2017) which uses saddlepoint techniques on the a statistic defined as

$$\tilde{U}_\gamma = U_\gamma - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta, \tag{4.1}$$

where $\boldsymbol{I}_{\gamma\beta}$ and $\boldsymbol{I}_{\beta\beta}$ are submatrices of the expected Fisher information matrix, as before. According to Johnsen et al. (2023; p.3), $\tilde{U}_\gamma$ can be interpreted as a reparametrization of the regression model

$$(\boldsymbol{\beta}, \gamma) \to (\boldsymbol{\alpha}, \gamma),$$

for some parameter $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\boldsymbol{\beta}, \gamma)$ depending on the original parameters, such that the new parametrization yields *local orthogonality* under the null hypothesis $\gamma = \gamma_0$, meaning $\boldsymbol{I}_{\alpha\gamma} = \boldsymbol{I}_{\gamma\alpha} = 0$ in the expected information matrix of this reparametrized model. This is precisely the motivation for the construction of $\tilde{U}_\gamma$, henceforth referred to as *effective score*, namely to remove some of the dependency between the elements of the score vector, in order to allow for unconditional inference. This “alternative” formulation of the score appears in many different frameworks in the literature.

In a hypothesis testing framework, the articles by Hall and Mathiason (1990; p. 82) and Marohn (2002; p. 341) define something called *effective score*, as in equation (4.1). The effective score was also recently used in a paper by Hemerik et al. (2020) as an alternative to the regular score in a sign-flipping test.

The alternative score presented in equation (4.1) also appears in an article by Waterman and B. G. Lindsay (1996; p. 4). Their paper shows how the *conditional score* (McCullagh and Nelder, 1989)(B. Lindsay, 1982) can be approximated using projection methods. Their method results in something they call the tth order *projected score*, and the first-order, linear approximation turns out to be the effective score.

Without going into detail, we will also acknowledge that the alternative score from equation (4.1) is a familiar concept in semi-parametric and non-parametric settings (Bickel et al., 1993)(Choi et al., 1996). When the nuisance parameters are of potentially infinite dimension or are functions, the effective score can be used to perform inference on the parameter of interest. In this case, the most common name seems to be *efficient score*, also used in Johnsen et al. (2023).

In this chapter, we will attempt to derive and explain this other score statistic by considering the work of Waterman and B. G. Lindsay (1996). First, we show how to derive the conditional score in a special case. Secondly, we link the conditional score to what is called the projected score, and show that the first-order projected score is the effective score. Lastly, we show how the effective score can be used in place of the regular score as the test statistic in a score test, and consider some asymptotic properties of the effective score test. We will denote the alternative score as the *effective score* as it seems to be the most common consensus within the parametric testing framework. We will denote the score from section 2.4 as the *regular* score, to emphasize the difference and avoid confusion.

4.1 Conditional score

The conditional score is defined as (B. Lindsay, 1982: p.504)

$$U_c = U_\gamma - E[U_\gamma | \boldsymbol{T}_\beta], \tag{4.2}$$

where U_γ is a score with respect to a parameter of interest γ, and $\boldsymbol{T}_\beta$ is a sufficient statistic for a vector of nuisance parameters $\boldsymbol{\beta}$. We recognize the notation from our above discussion of generalized linear models. However, we emphasize that conditional scores are not limited to only these models. Note also that $E[U_\gamma|\boldsymbol{T}_\beta]$ does not mean the expectation of U_γ conditioned on some observation $\boldsymbol{T}_\beta = \boldsymbol{t}_\beta$, but rather conditioned on the stochastic variable $\boldsymbol{T}_\beta$, and we, therefore, think of $E[U_\gamma|\boldsymbol{T}_\beta]$ as a random variable.

Following McCullagh and Nelder (1989; Ch. 7.2.2) assume that for each fixed $\gamma = \gamma_0$, there exists a sufficient and complete statistic $\boldsymbol{T}_\beta(\gamma_0)$ for $\boldsymbol{\beta}$. We distinguish between the statistic $\boldsymbol{T}_\beta(\gamma_0)$ being the same for all choices of γ_0, and when it depends on γ_0. In Waterman and B. G. Lindsay (1996), they denoted these two cases a *Type I* problem and *Type II* problem, respectively. They also state that generalized linear models with a canonical link, where Y_i is from a regular exponential family are models of Type I structure. The conditional score as defined in equation (4.2) is well-defined for both Type I and Type

II problems. However, for Type I problems, meaning $\boldsymbol{T}_\beta(\gamma_0) = \boldsymbol{T}_\beta$ is the same for all choices of γ_0, the conditional score can be computed as the derivative of the conditional log-likelihood (Waterman and B. G. Lindsay, 1996: p. 1).

Derivation of the conditional score

We now derive the conditional score as the derivative of the conditional log-likelihood, in a Type I problem. We start by noting that the joint density of $Y_1, \ldots, Y_n$ can be expressed as

$$f_Y(\boldsymbol{Y}; \boldsymbol{\beta}, \gamma) = f_{Y,T_\beta}(\boldsymbol{Y}; \boldsymbol{\beta}, \gamma) = f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma) f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma).$$

The first equality follows from the fact that $\boldsymbol{T}_\beta$ is fully determined by $\boldsymbol{Y}$. From this, we define the *conditional log-likelihood* as

$$\begin{aligned} \ell_c(\gamma; \boldsymbol{Y}, \boldsymbol{T}_\beta) &= \ln f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma) \\ &= \ln f_Y(\boldsymbol{Y}; \boldsymbol{\beta}, \gamma) - \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma). \end{aligned} \tag{4.3}$$

Note that ℓ_c is a function only of γ whereas the regular log-likelihood is a function of the full parameter vector $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$. Denote by $\tilde{U}_c$ the derivative of the conditional log-likelihood, that is

$$\begin{aligned} \tilde{U}_c &= \frac{\partial}{\partial \gamma} \ell_c(\gamma; \boldsymbol{Y}, \boldsymbol{T}_\beta) \\ &= \frac{\partial \ln f_Y(\boldsymbol{Y}; \boldsymbol{\beta}, \gamma)}{\partial \gamma} - \frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma}. \end{aligned} \tag{4.4}$$

We claim that $\tilde{U}_c = U_c$ from equation (4.2).

The first expression in equation (4.4) is recognizable as the regular score for γ,

$$U_\gamma = \frac{\partial}{\partial \gamma} \ell(\boldsymbol{\beta}, \gamma; \boldsymbol{Y}).$$

In order to justify our claim, what remains to show is that

$$\frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma} = E[U_\gamma | \boldsymbol{T}_\beta].$$

We start by noting that

$$\begin{aligned} E[U_\gamma | \boldsymbol{T}_\beta] &= E\left[\tilde{U}_c + \frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma} \middle| \boldsymbol{T}_\beta \right] \\ &= E[\tilde{U}_c | \boldsymbol{T}_\beta] + E\left[\frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma} \middle| \boldsymbol{T}_\beta \right], \end{aligned} \tag{4.5}$$

where we use the fact that

$$U_\gamma = \tilde{U}_c + \frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma}$$

from a rearrangement of equation (4.4). The first term in equation (4.5) is zero, under the assumption that integral and derivative are interchangeable, since

$$\begin{aligned}
E[\tilde{U}_c|\boldsymbol{T}_\beta] &= E\left[\frac{\partial}{\partial\gamma}\ell_c(\gamma; \boldsymbol{Y}, \boldsymbol{T}_\beta)\middle|\boldsymbol{T}_\beta\right] \\
&= E\left[\frac{\frac{\partial}{\partial\gamma} f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma)}{f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma)}\middle|\boldsymbol{T}_\beta\right] \\
&= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial\gamma} f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma)}{f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma)} f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma) d\boldsymbol{Y} \\
&= \frac{\partial}{\partial\gamma}\int_{\mathcal{Y}} f_{Y|T_\beta}(\boldsymbol{Y}; \boldsymbol{T}_\beta, \gamma) d\boldsymbol{Y} \\
&= \frac{\partial}{\partial\gamma} 1 \\
&= 0.
\end{aligned}$$

This is similar to the argument that is used to show that the regular score function has expected value zero, which we showed in equation (2.12). For the remaining term in equation (4.5), we get

$$E\left[\frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma}\middle|\boldsymbol{T}_\beta\right] = \frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma},$$

since $E[h(\boldsymbol{T}_\beta)|\boldsymbol{T}_\beta] = h(\boldsymbol{T}_\beta)$ (Karr, 1993: Example 8.16). Hence,

$$E[U_\gamma|\boldsymbol{T}_\beta] = 0 + \frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma}.$$

To summarize, we have shown, by computing the derivative of the conditional log-likelihood (4.3), that

$$\begin{aligned}
\tilde{U}_c &= \frac{\partial \ln f_Y(\boldsymbol{Y}; \boldsymbol{\beta}, \gamma)}{\partial \gamma} - \frac{\partial \ln f_{T_\beta}(\boldsymbol{T}_\beta; \boldsymbol{\beta}, \gamma)}{\partial \gamma} \\
&= U_\gamma - E[U_\gamma|\boldsymbol{T}_\beta] = U_c,
\end{aligned}$$

which is what we wanted.

Remark 4.1.1. From the calculations above, we get a computational understanding of the conditional score. However, as noted by B. Lindsay (1982) there is also another, perhaps more intuitive interpretation. We can think of the conditional expectation $E[U_\gamma|\boldsymbol{T}_\beta]$ as an orthogonal projection of U_γ onto the probability subspace generated by $\boldsymbol{T}_\beta$, which is (Karr, 1993: p.227)

$$S_{\boldsymbol{T}_\beta} \overset{\text{def}}{=} \{X \in L^2 : X = h(\boldsymbol{T}_\beta) \text{ for some } h : \mathbb{R}^p \to \mathbb{R}\}.$$

The conditional score is then the residual of the projection of U_γ onto the subspace $S_{\boldsymbol{T}_\beta}$, which will be orthogonal to $\boldsymbol{T}_\beta$. The visual interpretation of this can be seen in Figure 4.1.

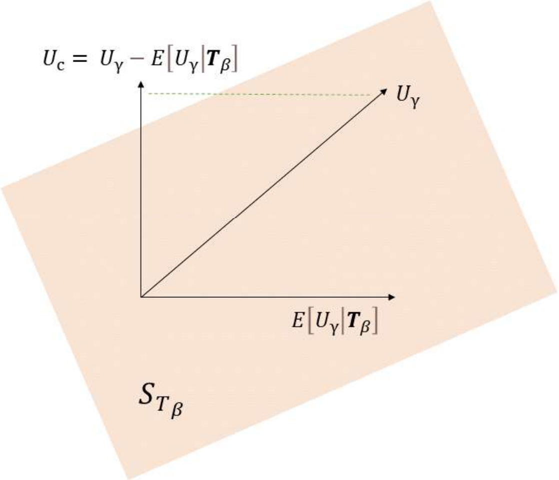

$U_c = U_\gamma - E[U_\gamma|\boldsymbol{T}_\beta]$
U_γ
$E[U_\gamma|\boldsymbol{T}_\beta]$
S_{T_β}

Figure 4.1: An illustration of how the conditional expectation $E[U_\gamma|\boldsymbol{T}_\beta]$ is projected out in order to make a new, orthogonal parameter U_c.

4.2 Projected score

The conditional expectation, $E[U_\gamma|\boldsymbol{T}_\beta]$, is typically intractable, and, consequently, U_c will also be intractable. However, Waterman and B. G. Lindsay (1996) suggests a method for approximating the conditional score by exploiting that this is a projection onto the subspace $S_{\boldsymbol{T}_\beta}$.

They propose to use the so-called *Bhattacharya basis*, to construct a space called Λ_{CLR}, which approximates $S_{\boldsymbol{T}_\beta}$. Assume that $Y_1, \dots, Y_n$ is a random sample from a parametric family[1], and let our vector of parameters be $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$ with dimension $p+1$. Then the basis for the subspace Λ_{CLR}, the Bhattacharyya basis, is defined as a set of sets $\mathcal{B}_t = (\mathcal{V}_0^T, \dots, \mathcal{V}_t^T)^T$, where $\mathcal{V}_0$ is the constant function 1 and $\mathcal{V}_t$ is the column vector containing all t'th order derivatives on the form

$$V_{i_1,\dots,i_t} = \frac{\partial^t \mathcal{L}(\boldsymbol{\beta}, \gamma; \boldsymbol{y})}{\partial\beta_{i_1}, \dots, \partial\beta_{i_t}} \Big/ \mathcal{L}(\boldsymbol{\beta}, \gamma; \boldsymbol{y}), \quad 1 \le i_j \le p,$$

where $\mathcal{L}(\boldsymbol{\beta}, \gamma; \boldsymbol{y})$ is the likelihood function.

Example 4.2.1. Let $Y_1, \dots, Y_n$ be a random sample from a parametric model with two nuisance parameters, that is $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$, and denote its likelihood by $\mathcal{L} = \mathcal{L}(\boldsymbol{\beta}, \gamma; \boldsymbol{y})$.

[1] For instance a generalized linear model with known covariates.

Then the Bhattacharryya basis for $t = 1, 2$ will be

$$\begin{aligned}
\mathcal{B}_1 &= (\mathcal{V}_0^T, \mathcal{V}_1^T)^T = \left(1, \frac{\partial \mathcal{L}}{\partial \beta_1} \Big/ \mathcal{L}, \frac{\partial \mathcal{L}}{\partial \beta_2} \Big/ \mathcal{L}\right)^T \\
\mathcal{B}_2 &= (\mathcal{V}_0^T, \mathcal{V}_1^T, \mathcal{V}_2^T)^T \\
&= \left(1, \frac{\partial \mathcal{L}}{\partial \beta_1} \Big/ \mathcal{L}, \frac{\partial \mathcal{L}}{\partial \beta_2} \Big/ \mathcal{L}, \frac{\partial^2 \mathcal{L}}{\partial \beta_1^2} \Big/ \mathcal{L}, \frac{\partial^2 \mathcal{L}}{\partial \beta_1 \partial \beta_2} \Big/ \mathcal{L}, \frac{\partial^2 \mathcal{L}}{\partial \beta_2 \partial \beta_1} \Big/ \mathcal{L}, \frac{\partial^2 \mathcal{L}}{\partial \beta_2^2} \Big/ \mathcal{L}\right)^T
\end{aligned}$$

As seen from example 4.2.1, the length of the Bhattacharyya basis grows quickly, since $\mathcal{V}_t$ has length p^t, where p is the dimension of the nuisance parameter $\boldsymbol{\beta}$.

The Bhattacharya basis is used to make what is called the *projected score*, which can be defined in terms of the score vector $\boldsymbol{U}$ and a matrix M_t given by

$$M_t = E\left(\begin{bmatrix} U_\gamma \\ \mathcal{B}_t \end{bmatrix} \begin{bmatrix} U_\gamma & \mathcal{B}_t^T \end{bmatrix}\right) = \begin{bmatrix} j_{\gamma\gamma} & j_{\gamma\beta} \\ j_{\beta\gamma} & j_{\beta\beta} \end{bmatrix}$$

As noted by Waterman and B. G. Lindsay (1996; p.3), the constant $\mathcal{V}_0$ is redundant and can be ignored in the computations. According to Waterman and B. G. Lindsay (1996; p. 4), the t'th order projected score is given by

$$U_t = U_\gamma - j_{\gamma\beta} j_{\beta\beta}^{-1} \mathcal{B}_t.$$

Since the matrix M_1 is the expected Fisher information matrix (Waterman and B. G. Lindsay, 1996: p. 4), it is easily seen that the first-order projected score is the effective score

$$\tilde{U}_\gamma = U_\gamma - \boldsymbol{I}_{\gamma\beta} \boldsymbol{I}_{\beta\beta}^{-1} \boldsymbol{U}_\beta. \tag{4.6}$$

In conclusion, the conditional score is an intractable score that is orthogonal to the space $S_{\boldsymbol{T}_\beta}$ spanned by the sufficient statistic $\boldsymbol{T}_\beta$. The effective score is an approximation of the conditional score, namely the residual of U_γ after projecting onto a subspace spanned by the nuisance scores which we have shown is a linear approximation to the spaces $S_{\boldsymbol{T}_\beta}$. This intuition corresponds with Hall and Mathiason (1990; p. 82), who writes that "[effective scores are] obtained as the residual from projection of [U_γ on $\boldsymbol{U}_\beta$], that part of the score for γ which is orthogonal to the score for $\boldsymbol{\beta}$".

4.3 Effective score and effective score test

Let $(Y_1, \boldsymbol{v}_1), \ldots, (Y_n, \boldsymbol{v}_n)$ be a random sample from a regular exponential family and assume the responses come from a generalized linear model using a canonical link function. We have shown earlier that the score then takes the general form

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_\beta \\ U_\gamma \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \\ \boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \end{bmatrix},$$

with expectation zero and variance given by the expected Fisher information matrix

$$\mathcal{I}(\boldsymbol{\psi}) = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} \end{bmatrix}.$$

For these models the effective score takes the general form

$$\begin{aligned} \tilde{U}_\gamma &= U_\gamma - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta \\ &= U_\gamma - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{U}_\beta \\ &= \boldsymbol{Z}^T(\boldsymbol{Y}-\boldsymbol{\mu}) - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{\mu}) \\ &= (\boldsymbol{Z}^T - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T)(\boldsymbol{Y}-\boldsymbol{\mu}) \\ &= \tilde{\boldsymbol{Z}}^T(\boldsymbol{Y}-\boldsymbol{\mu}), \end{aligned}$$

where

$$\tilde{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z}.$$

The expectation and variance of $\tilde{U}_\gamma$ can be found through straightforward calculations, always keeping $\boldsymbol{\psi}$ fixed at its true parameter value,

$$\begin{aligned} E(\tilde{U}_\gamma) &= E(U_\gamma - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta) \\ &= E(U_\gamma) - E(\boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta) \\ &= E(U_\gamma) - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}E(\boldsymbol{U}_\beta) \\ &= 0. \\ \text{Var}(\tilde{U}_\gamma) &= \text{Var}(U_\gamma - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta) \\ &= \text{Var}(U_\gamma) + \text{Var}(\boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta) - 2\text{Cov}(U_\gamma, \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta) \\ &= \boldsymbol{I}_{\gamma\gamma} + \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\text{Var}(\boldsymbol{U}_\beta)(\boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1})^T - 2\text{Cov}(U_\gamma, \boldsymbol{U}_\beta)(\boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1})^T \\ &= \boldsymbol{I}_{\gamma\gamma} + \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\gamma} - 2\boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\gamma} \\ &= \boldsymbol{I}_{\gamma\gamma} - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\gamma} \\ &= \tilde{\boldsymbol{Z}}^T\boldsymbol{W}\tilde{\boldsymbol{Z}}. \end{aligned}$$

The final line in the calculations of the variance can be found by straightforward verification and is also the notation used by Johnsen et al. (2023).

The motivation behind formulating the effective score is to reduce the dependency on the nuisance parameters, in order to perform unconditional inference. It is easily shown that the effective score is indeed uncorrelated with the score vector $\boldsymbol{U}_\beta$, still keeping $\boldsymbol{\psi}$ fixed at its true parameter value

$$\begin{aligned}\mathrm{Cov}(\tilde{U}_\gamma, \boldsymbol{U}_\beta) &= \mathrm{Cov}(U_\gamma - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta, \boldsymbol{U}_\beta)\\ &= \mathrm{Cov}(U_\gamma, \boldsymbol{U}_\beta) - \mathrm{Cov}(\boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta, \boldsymbol{U}_\beta)\\ &= \boldsymbol{I}_{\gamma\beta} - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\mathrm{Cov}(\boldsymbol{U}_\beta, \boldsymbol{U}_\beta)\\ &= \boldsymbol{I}_{\gamma\beta} - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}I_{\beta\beta}\\ &= \boldsymbol{I}_{\gamma\beta} - \boldsymbol{I}_{\gamma\beta}\\ &= \boldsymbol{0}.\end{aligned}$$

Hence, we formulate the effective score test as an unconditional test

$$P(\tilde{U}_\gamma \leq \tilde{u}),$$

where we allow the nuisance parameters $\boldsymbol{\beta}$ to be fixed at their maximum likelihood estimators $\hat{\boldsymbol{\beta}}$, without conditioning in the distribution. In a sense, we can think of it as if the conditioning happened already at "likelihood-level", instead of in the distribution.

As noted by Hall and Mathiason (1990; p. 78), the effective score test and the regular score test are equivalent under the assumption that the score vector $\boldsymbol{U}$ is multivariate normally distributed. To see this, assume now that $\boldsymbol{U}$ is multivariate normal. Then

$$\tilde{U}_\gamma = U_\gamma - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta \tag{4.7}$$

is a linear combination of normally distributed random variables, and therefore also normally distributed. The normal distribution is uniquely determined by the mean vector and covariance matrix, and we have already shown that the effective score has expectation and variance

$$E(\tilde{U}_\gamma) = 0, \qquad \mathrm{Var}(\tilde{U}_\gamma) = I_{\gamma\gamma} - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\gamma}.$$

Hence,

$$\tilde{U}_\gamma \overset{d.}{=} U_\gamma | \boldsymbol{U}_\beta = \boldsymbol{0},$$

when $\boldsymbol{U}$ is multivariate normal.

Furthermore, the observed value $\tilde{u} = \tilde{\boldsymbol{Z}}_0^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)$ will be equal to the observed value $u = \boldsymbol{Z}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)$, because $\boldsymbol{U}_{\hat{\beta}} = \boldsymbol{0}$ when we plug in the MLE for $\boldsymbol{\beta}$. Therefore, under the assumption that $\boldsymbol{U}$ is multivariate normal, we have

$$P(\tilde{U}_\gamma \leq \tilde{u}) = P(\tilde{U}_\gamma \leq u) = P(U_\gamma \leq u \,|\, \boldsymbol{U}_\beta = \boldsymbol{0}).$$

This relation is probably also why Hall and Mathiason (1990; p. 78) notes that the effective score test "will rarely differ much from the [regular score] test in practice". However, if we now instead apply a saddlepoint approximation to compute the distribution of $\tilde{U}_\gamma$, the tests may differ. We illustrate how an effective score test with saddlepoint approximation can be computed in the following example.

Example 4.3.1. Let $(y_1, \boldsymbol{v}_1), \dots, (y_n, \boldsymbol{v}_n)$ be a set of observations where each y_i is a realization of the random variable $Y_i \sim \text{Pois}(\mu_i)$ for $i = 1, \dots, n$. The linear predictor[2] $\eta_i = x_i\beta + z_i\gamma$ is, as before, linked to the response by a canonical link function, and we are interested in testing the null-hypothesis $H_0 : \gamma \leq 0$ against $H_1 : \gamma > 0$ using a score test.

We start by considering the effective score, which is defined as

$$\tilde{U}_\gamma = \tilde{\boldsymbol{Z}}^T(\boldsymbol{Y} - \boldsymbol{\mu}).$$

The saddlepoint approximation of the distribution $P(\tilde{U}_\gamma \leq u)$ can be computed using the cumulant generating function as in subsection 3.2.2. Then we must first compute the CGF of $\tilde{U}_\gamma$, which is straightforward using the propositions from section 3.1, and we get

$$\begin{aligned} K_{\tilde{U}_\gamma}(s) &= \sum_{i=1}^{n} K_{Y_i}(\tilde{z}_i s) - s\tilde{z}_i\mu_i \\ &= \sum_{i=1}^{n} \mu_i(\exp(\tilde{z}_i s) - 1) - s\tilde{z}_i\mu_i. \end{aligned}$$

To compute the saddlepoint, we solve the equation

$$K'_{\tilde{U}_\gamma}(\hat{s}) = \sum_{i=1}^{n} \mu_i \exp(\tilde{z}_i\hat{s})\tilde{z}_i - \tilde{z}_i\mu_i = u,$$

and this equation can be solved numerically by minimizing the function

$$s \mapsto K_{\tilde{U}_\gamma}(s) - us.$$

We also need the second derivative of the CGF, which is

$$K''_{\tilde{U}_\gamma}(s) = \sum_{i=1}^{n} \mu_i \exp(\tilde{z}_i s)\tilde{z}_i^2.$$

We note that both $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{Z}}$, through $\boldsymbol{W}$, depend on the parameters $\boldsymbol{\psi} = [\beta, \gamma]^T$, and must therefore be computed under the null hypothesis and with the maximum likelihood estimate in place of β. The saddlepoint approximation becomes

$$\hat{P}(\tilde{U}_\gamma \leq u) = \Phi(\hat{\omega}) + \phi(\hat{\omega})(1/\hat{\omega} - 1/\hat{u}), \quad \tilde{u} \neq 0$$

with

$$\hat{\omega} = \text{sgn}(\hat{s})\sqrt{2[\hat{s}u - K_{\tilde{U}_\gamma}(\hat{s})]}, \quad \hat{u} = \hat{s}\sqrt{K''_{\tilde{U}_\gamma}(\hat{s})},$$

and the p-value of our right-tailed test is then $p = 1 - \hat{P}(\tilde{U}_\gamma \leq u)$.

[2] The example would work well also for a vector of nuisance parameters $\boldsymbol{\beta}$, instead of a scalar as considered here.

For a regular score U_γ from an exponential family, we saw in subsection 3.3 that it is possible to use the special case of saddlepoint approximation expressed in terms of maximum likelihood estimates and not cumulant generating functions. We note that $\tilde{U}_\gamma = \tilde{\boldsymbol{Z}}^T(\boldsymbol{Y} - \boldsymbol{\mu})$ appears as a score function for an exponential family, with expectation 0 and variance $\tilde{\boldsymbol{Z}}^T\boldsymbol{W}\tilde{\boldsymbol{Z}}$. We, therefore, suspect that the saddlepoint implementation that uses maximum likelihood estimates can be used for the effective score as well. This is done by first fitting a null model for estimating the coefficients $\boldsymbol{\beta}$, and then a "full" model where we manually set the intercept to be $\beta_0^{(i)} = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, for $i = 1, \ldots, n$. We then use the output from these two models in the expression from equation (2.18). The approximation becomes

$$\hat{P}(\tilde{U}_\gamma \leq u) = \Phi(\hat{\omega}) + \phi(\hat{\omega})(1/\hat{\omega} - 1/\hat{u}), \quad u \neq 0,$$

where

$$\hat{\omega} = \operatorname{sgn}(\hat{\gamma} - \gamma)\sqrt{-2\ln\frac{\mathcal{L}(\gamma)}{\mathcal{L}(\hat{\gamma})}}, \quad \hat{u} = (\hat{\gamma} - \gamma)\sqrt{\mathcal{I}(\hat{\gamma})},$$

with $\gamma = 0$ and with $\hat{\gamma}$ being the MLE from the "full model". A pseudo-algorithm of how this could be done is given in algorithm 1.

Algorithm 1: Effective score test with saddlepoint approximation using glm-output

Input : A vector of responses $\boldsymbol{y}$, a matrix of nuisance covariates $\boldsymbol{X}$, and a vector covariate of interest $\boldsymbol{Z}$

Output: The right-tail p-value of the hypothesis test $\gamma = 0$

```
effectiveScoreTest ($\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}$)
$\textbf{\textit{mod0}} \leftarrow glm(\boldsymbol{y} \sim \boldsymbol{X})$
$\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}} \leftarrow \textbf{\textit{mod0}}$
$\tilde{\boldsymbol{\beta}}_0 \leftarrow g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$,                ($g$ is the canonical link function)
Compute $\tilde{\boldsymbol{Z}}$ using the fitted values
$\textbf{\textit{mod1}} \leftarrow glm(\boldsymbol{y} \sim \tilde{\boldsymbol{Z}} + \text{offset}(\tilde{\boldsymbol{\beta}}_0))$
$\hat{\gamma}, \mathcal{I}(\hat{\gamma}) \leftarrow \textbf{\textit{mod1}}$
$\hat{\omega} = \text{sgn}(\hat{\gamma})\sqrt{-2(\text{loglik}(\textbf{\textit{mod0}}) - \text{loglik}(\textbf{\textit{mod1}}))}$
$\hat{u} = \hat{\gamma}\sqrt{\mathcal{I}(\hat{\gamma})}$
$p$-val $= \text{pnorm}(\hat{\omega}) + \text{dnorm}(\hat{\omega})(1/\hat{\omega} - 1/\hat{u})$
```

4.3.1 Asymptotic properties

In subsection 2.4, we said that the score has been shown to be asymptotically multivariate normally distributed. From this we get important asymptotic results concerning the effective score $\tilde{U}_\gamma$ as well. First of all, we note that

$$\boldsymbol{U} \xrightarrow{\text{d}} \mathcal{N}_{p+1} \implies \tilde{U}_\gamma \xrightarrow{\text{d}} \mathcal{N}_1.$$

This follows from the same argument as in equation (4.7), in particular, that $\tilde{U}_\gamma$ is, asymptotically, a linear combination of normally distributed random variables. From this, it follows that the effective score $\tilde{U}_\gamma$ and the nuisance score $\boldsymbol{U}_\beta$ will be asymptotically independent, as uncorrelatedness implies independence for normally distributed random variables.

Furthermore, we have that the *unconditional* test using $\tilde{U}_\gamma$ will be asymptotically the same as the conditional test $U_\gamma | \boldsymbol{U}_\beta = \mathbf{0}$. Note also that this will hold in practice when we use a saddlepoint approximation to determine $P(\tilde{U}_\gamma \leq u)$, as saddlepoint approximations are exact for the normal distribution. Hence, asymptotically, a saddlepoint approximated, effective score test will be the same as a conditional score test computed with a normal distribution.

Lastly, we will also note that the effective score is asymptotically equal to the conditional score. To see this, consider $\boldsymbol{U} = \boldsymbol{T} - E[\boldsymbol{T}]$ as simply a centered sufficient statistic, and note that $\boldsymbol{T}$ must also be normal if $\boldsymbol{U}$ is multivariate normal. Furthermore, we have that $\text{Var}[\boldsymbol{U}] = \text{Var}[\boldsymbol{T}] = \mathcal{I}(\boldsymbol{\psi})$. The conditional expectation $E[U_\gamma | \boldsymbol{T}_\beta]$ can be written out explicitly when both U_γ and $\boldsymbol{T}_\beta$ are normal. We get,

$$E[U_\gamma | \boldsymbol{T}_\beta] = E[U_\gamma] - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}(\boldsymbol{T}_\beta - E[\boldsymbol{T}_\beta]) = 0 - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta.$$

Hence, the conditional score will asymptotically be given by

$$U_c = U_\gamma - E[U_\gamma | \boldsymbol{T}_\beta] = U_\gamma - \boldsymbol{I}_{\gamma\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{U}_\beta\boldsymbol{U}_\beta = \tilde{U}_\gamma,$$

which is the efficient score.

Chapter 5

Simulated examples

The goal of this thesis is to present two new methods of computing p-values for score tests. From chapter 3 we saw that a double saddlepoint approximation could be used to compute more accurate p-values of a score test, and from chapter 4 we defined an effective score test, also based on a saddlepoint approximation. We have also seen that the saddlepoint approximation can be implemented in two different ways, which yields a total of four "different" methods which should in theory be pairwise equivalent. In this chapter, we consider simulated examples to investigate and compare what we have only presented theoretically so far.

We begin by comparing the two different implementations, either using the cumulant generating function or using maximum likelihood estimates. We will call them the CGF-method and MLE-method, for short. We have seen that the double saddlepoint approximation should be equivalent regardless of which method we use to implement it, but since the effective score is not exactly the same as a regular score, we are particularly interested to see if the effective score test will work with the implementation that uses maximum likelihood estimates, as outlined in algorithm 1, since this has not been shown explicitly.

Thereafter, we focus on the difference between the tests, and compare double saddlepoint approximated score test and effective score test against the regular score test. We are interested in investigating the ability of the tests to control the type I error at a given significance level. We consider first a Poisson regression example with an imbalanced response and next a small sample example with n less than 100, where we fit a logistic regression model. A complete overview of the different simulations for this chapter is found in table 5.1. The code used to generate the different data sets as well as the code for implementation of each test can be found in Appendix A and B, respectively.

There is a singularity in the saddlepoint method, which was discussed in chapter 3. The problem occurs when the realization is close to the expected value of its corresponding random variable. As we have not implemented anything to compensate for this, all methods can be seen to have some illogical results near the expected value, which is 0 for the score. For a regular score test we have that the p-value is 0.5 for an observed value of 0. We will therefore see that p-values that in theory should lie near 0.5 can give irregular results when computed with either the double saddlepoint or effective score method.

For all examples, we let $(Y_1, \boldsymbol{v}_1), \ldots, (Y_n, \boldsymbol{v}_n)$ be a random sample from a regular exponential family, with the response and covariates linked through a generalized linear model with canonical link function. As before, denote by $\boldsymbol{V} = [\boldsymbol{X}, \boldsymbol{Z}]$ the covariate matrix, and let the parameter vector be given by $\boldsymbol{\psi} = [\boldsymbol{\beta}, \gamma]^T$. For our examples, we consider a score test to evaluate the hypothesis

$$H_0 : \gamma \leq 0 \quad \text{v.s.} \quad H_1 : \gamma > 0. \tag{5.1}$$

Hence, we wish to perform a right-tail hypothesis test. We implement the methods on a simulated data set $(y_1, \boldsymbol{v}_1), \ldots, (y_n, \boldsymbol{v}_n)$, where $\boldsymbol{v}_i = [\boldsymbol{x}_i, z_i]$ are both drawn from a random distribution. The covariate Z_i is always drawn from a Gamma(1,3) distribution and the parameter of interest is always 0, meaning we only consider models for which the null hypothesis is true. The nuisance covariates vary somewhat for the different examples. See table 5.1 or Appendix A for details.

Sim.	Model	Sample size	Nuisance Covariates	Number of simulations	Goal
1	Logistic	$n = 50$ $n = 500$	$p = 6$ All continuous	5000	Compare methods of saddlepoint implementation.
2	Logistic	$n = 50$ $n = 500$	$p = 3$ and $p = 10$ Mix of continuous and discrete	1000	Investigate the effect of different number of nuisance parameters.
3	Poisson	$n = 150$ $n = 500$ $n = 1000$	$p = 6$ All continuous	50000	Controlling type I error at significance level, $\alpha = 0.0005$, with imbalanced response.
4	Logistic	$n = 35$ $n = 50$ $n = 100$	$p = 6$ All continuous	50000	Controlling type I error at significance level, $\alpha = 0.0005$, with small sample size.

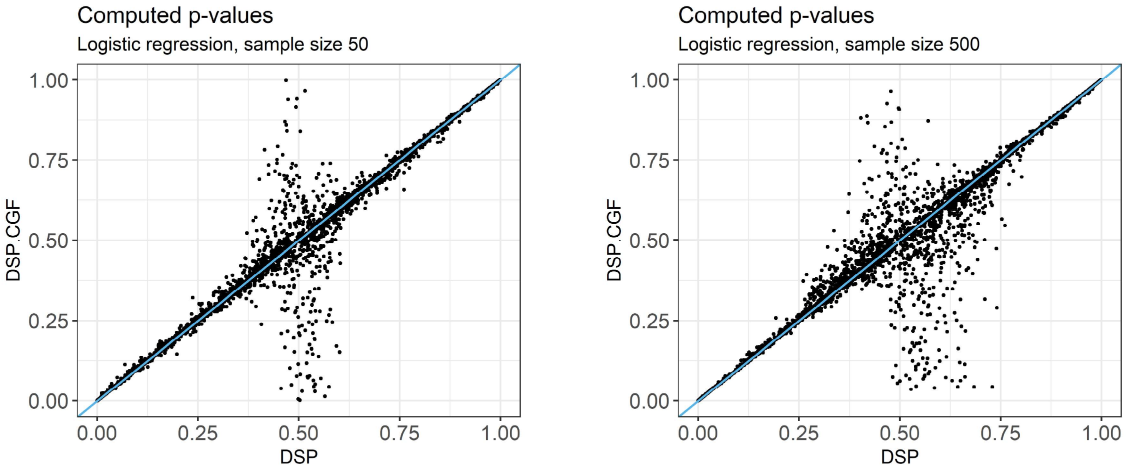
Table 5.1: Overview of the setup for each simulation done in this chapter. The covariate of interest, $\boldsymbol{Z}$, is always sampled from a Gamma$(1, 3)$ distribution.

5.1 Comparing implementation methods

In this section, we investigate how the double saddlepoint approximated score test and effective score test behaves for the two different implementations. We fit a logistic regression model and consider sample sizes of $n = 50$ and $n = 500$. We perform 5000 simulations, meaning we compute 5000 p-values for each method.

5.1.1 Double saddlepoint approximated score test

We consider first a double saddlepoint approximated score test. In Figure 5.1 we have plotted the 5000 computed p-values from the CGF-method against the MLE-method when performing the hypothesis test from equation (5.1). In theory, the two methods should agree, but it can be seen that they are quite different in practice. As the differences seem to grow near the singularity in $p = 0.5$, we find it plausible that the discrepancy between the methods is a result of instability in the vector optimization in the algorithm, which we perform to find the saddlepoints, rather than some inherent difference between the methods.



(a) Sample size $n = 50$

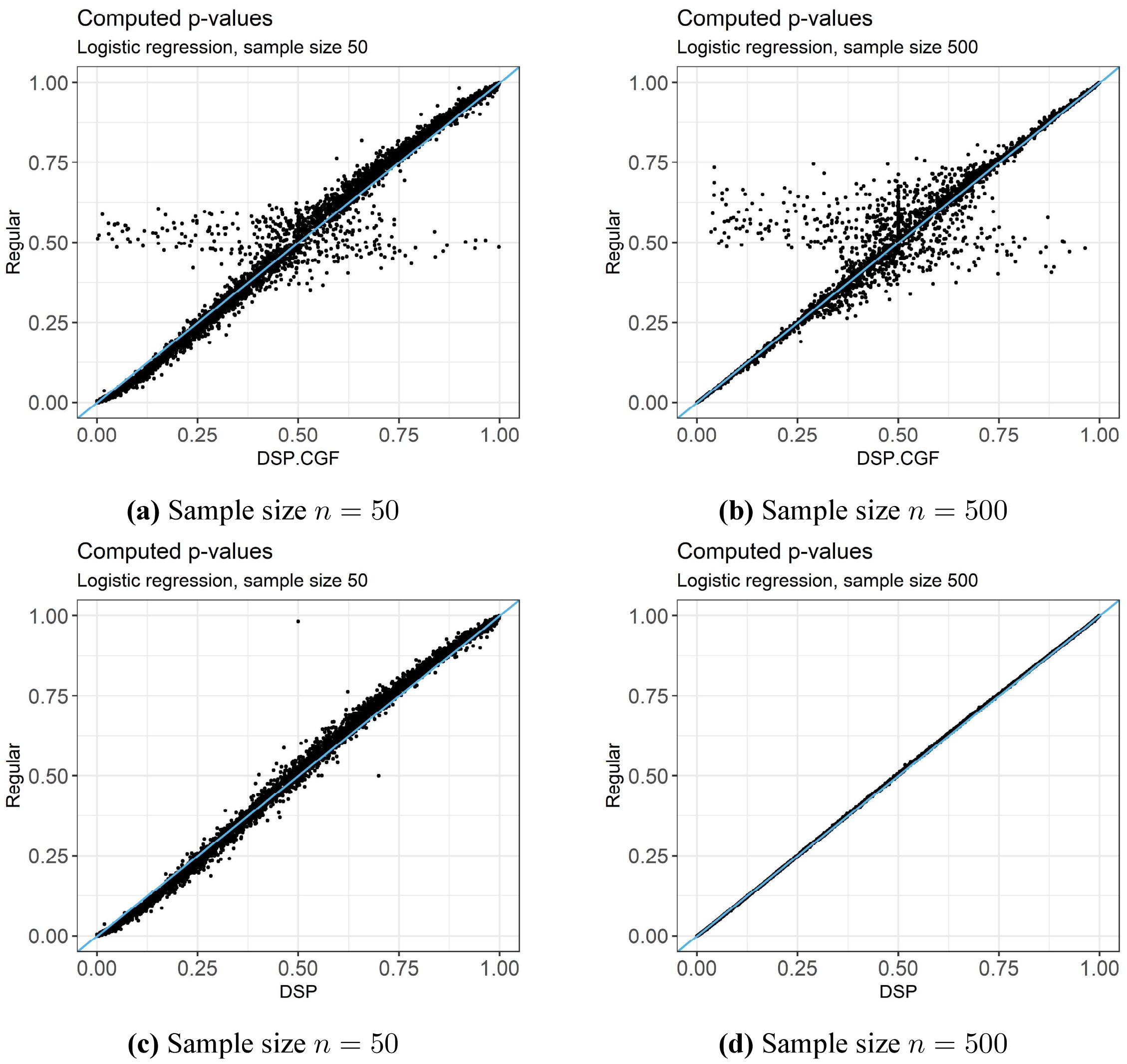
(b) Sample size $n = 500$

Figure 5.1: The computed p-values from both implementations of double saddlepoint plotted against one another. The x-axis is the MLE-method, and the y-axis is the CGF-method. The blue line across the diagonal displays $y = x$.

We compare both methods also to a regular score test, and those plots can be seen in Figure 5.2. Here, we have once again plotted the p-values from the different methods against each other. The instability of the CGF-method is once again clearly visible, which is seen in subfigures 5.2a and 5.2b. Trial and error also showed that this method is very sensitive to the initializing vector used in the saddlepoint optimization algorithm, where we use `optim` to find the saddlepoints

$$K'(\hat{\boldsymbol{r}}, \hat{s}) = (\boldsymbol{x}, y), \quad K'(\hat{\boldsymbol{r}}_0, 0) = K_X(\hat{\boldsymbol{r}}_0) = \boldsymbol{x},$$

as described in section 3.2.3. By contrast, the MLE-method seems quite consistent, despite a few outliers when $n = 50$ (subfigure 5.2c). For $n = 500$ (subfigure 5.2d) it seems practically identical to a regular score test, indicating that the distribution of the score has converged to a normal distribution.



(a) Sample size $n = 50$

(b) Sample size $n = 500$

(c) Sample size $n = 50$

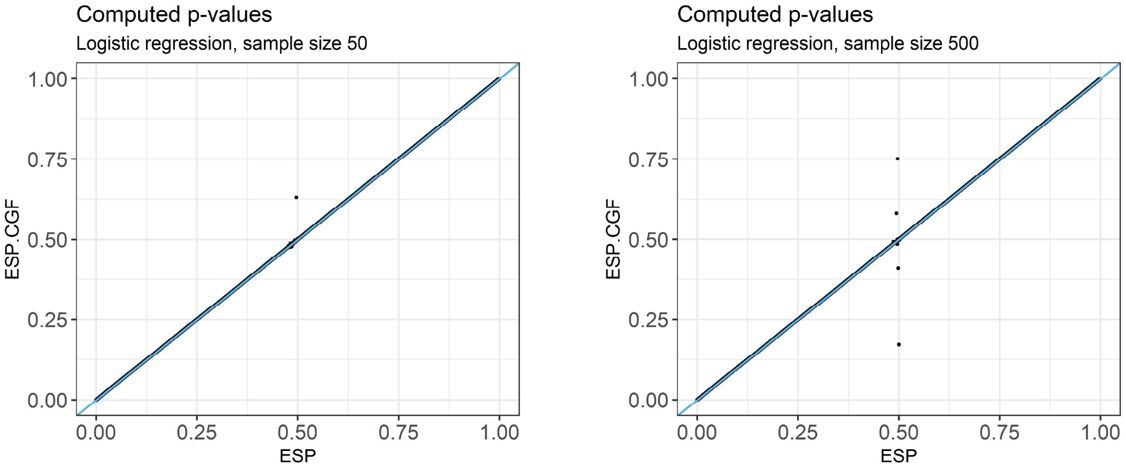
(d) Sample size $n = 500$

Figure 5.2: The computed p-values from double saddlepoint implemented with CGF-method (upper) and MLE-method (lower). The regular score test is plotted on the y-axis, whereas the double saddlepoint methods are on the x-axis.

Both implementations outputted some values that were "invalid", meaning that they were either `NA` or outside the interval $[0, 1]$. In the plots in Figure 5.2 these outputs have been manually set to 0.5, and the lack of outliers in the plots indicates that these invalid outputs are points that the regular score test will evaluate at around 0.5 as well. However, we note that there is an outlier in subfigure 5.2c, at approximately (DSP, Regular)$= (0.5, 1)$. One could speculate if this happened because the algorithm may also be unstable at p-values very close to 0 or 1, which would be problematic as we are particularly interested in computing tail probabilities for hypothesis testing. However, as this was not a recurring problem in the simulations, we have not investigated this further.

5.1.2 Effective score test

We now consider the effective score test for the same setup and data sets as in the previous subsection. From figure 5.3 we see that the two implementations seem to agree, except for a few outliers near the singularity in $p = 0.5$. This is interesting, as we were not certain beforehand that the two implementations would correspond perfectly. However, the simulations give confirmation that the two algorithms are practically equivalent.



(a) Sample size $n = 50$

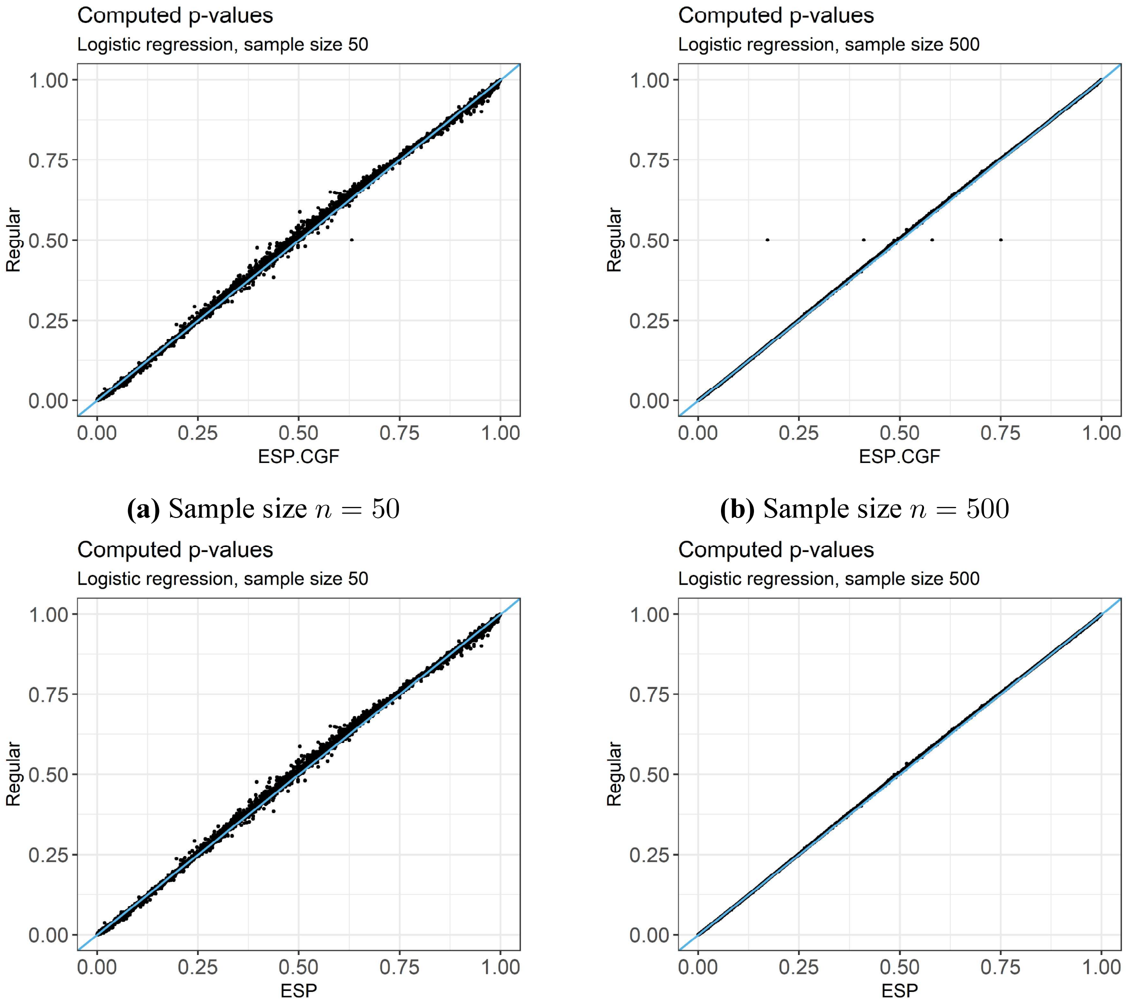
(b) Sample size $n = 500$

Figure 5.3: The computed p-values from both implementations of effective score test using saddlepoint approximation plotted against one another. None of the points have been manually set to 0.5 in this plot, and the outliers are likely due to instability in the saddlepoint optimization near the singularity.

We compare also the two implementations to a regular score test, and those plots can be seen in Figure 5.4. There are four outliers in subfigure 5.4b, that are not present in subfigure 5.4d, indicating that the CGF-method is slightly less robust than the MLE-method near the singularity for this example. However, four points are not sufficient in order to conclude that this is a general trend. Also here we see a clear convergence to a regular score test when $n = 500$ for both implementations, supporting the asymptotic results from subsection 4.3.1.

5.2 Investigating the effect of the number of nuisance parameters

We investigate how the methods are affected by the number of nuisance parameters, p. We compare two models, either with $p = 3$ or $p = 10$, including the intercept, and once again



(a) Sample size $n = 50$

(b) Sample size $n = 500$

(c) Sample size $n = 50$

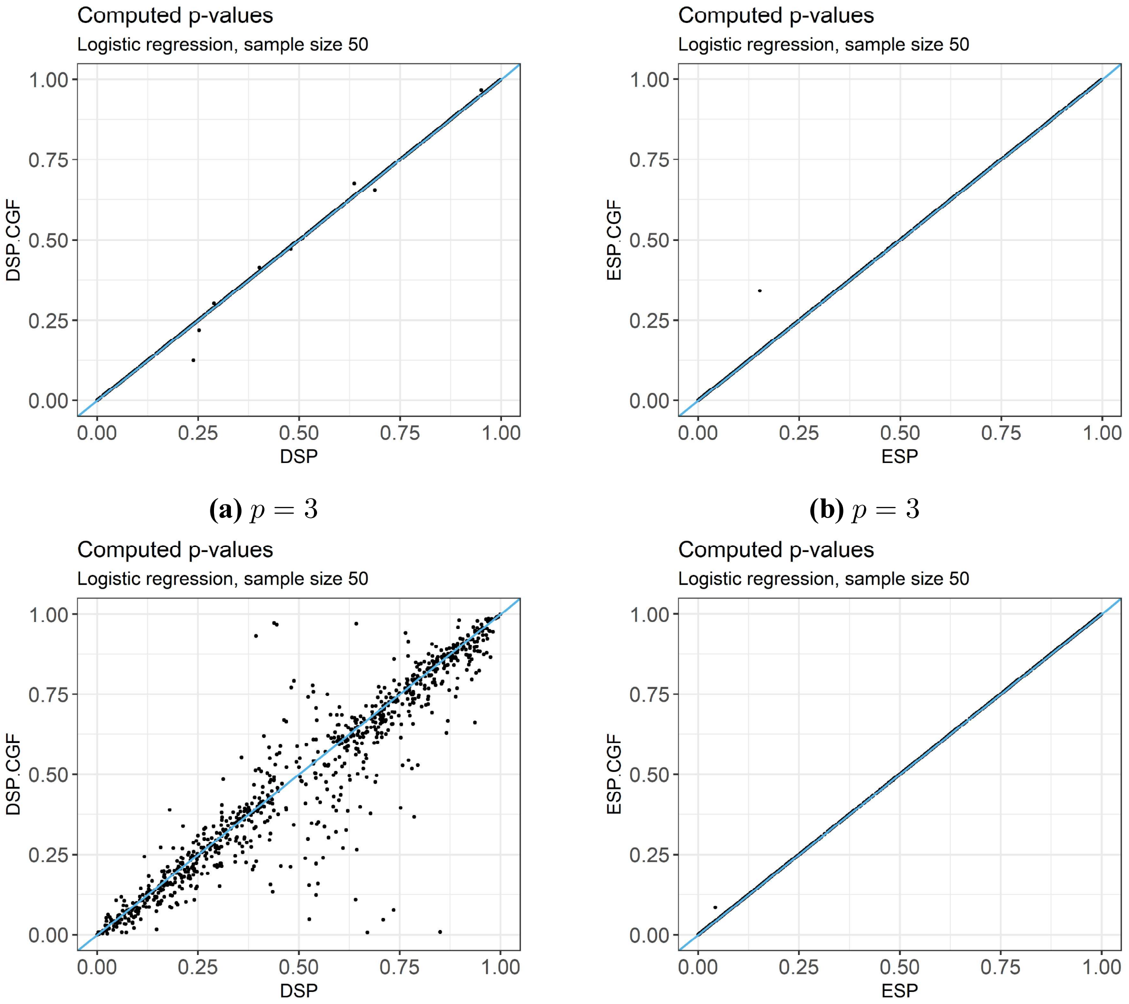
(d) Sample size $n = 500$

Figure 5.4: The computed p-values from effective score test with saddlepoint approximation implemented with CGF (upper) and glm-output (lower).

consider a logistic regression model[1]. We simulated 1000 data sets $(y_1, \boldsymbol{v}_1), \dots, (y_n, \boldsymbol{v}_n)$ and computed p-values using a double saddlepoint approximated score test and an effective score test, using both methods of implementation, and then compared the results based on whether the data set contained a smaller or larger set of covariates. We are interested in investigating the performance of the algorithms, not the actual inference. We remark that general conclusions cannot be drawn from one example, but the difference is still quite remarkable and interesting.

In Figure 5.5, we plot the computed p-values by the different implementations, meaning we plot the CGF-method against the MLE-method for the two tests and for the two sets of covariates. We see from subfigure 5.5c, where the double saddlepoint approximated score

[1] Testing with Poisson regression yielded similar results, but on a smaller scale as there was a higher overall stability of the algorithm.



(a) $p = 3$

(b) $p = 3$

(c) $p = 10$

(d) $p = 10$

Figure 5.5: Plotting the CGF-method against the MLE-method for small (upper) and large (lower) number of covariates. To the left is a plot with p-values computed by the two implementations of double saddlepoint approximated score test, and to the right are the two implementations of effective score test.

test is applied to a data set with 10 covariates, a similar trend as in Figure 5.1, namely that there is a quite high disagreement between the implementation methods and that the singularity in 0.5 interferes with the optimization algorithm. What is remarkable is that for $p = 3$, as seen in subfigure 5.5a, this instability is almost completely gone. This indicates a strong correlation between the number of covariates and the stability of the double saddlepoint approximated score test, in particular when implemented manually using the cumulant generating function. This remarkable difference may also be due to the fact that the CGF-method contains a quite naive attempt to do vector optimization by someone[2] who is not well-versed in the realm of numerical optimization, whereas the MLE-method indirectly uses more robust optimization algorithms implemented in packages in `R Studio`

[2] that would be the author

(Team, 2021). For the effective score test, the number of covariates seems to have no effect on the stability of the implementation methods, as the p-values are seemingly identical. It makes sense that the CGF-method is more stable for the effective score test, as it handles the nuisance parameters outside of the saddlepoint approximation. Therefore, the optimization used to determine the saddlepoint is always a convex optimization problem of one variable no matter the number of nuisance covariates.

We also consider the time spent by each algorithm, as well as the number of `NA`'s produced by each method. Here, we include the regular score test for reference. Note that the actual times themselves may vastly differ depending on processing power. They can, however, be used for relative comparison between the methods.

	Regular	DSP	DSP.CGF	ESP	ESP.CGF
$n = 50$	2.37	5.11	10.19	4.7	2.81
$n = 500$	3.98	10.15	51.10	12.18	7.98

(a) Running time with $p = 3$.

	Regular	DSP	DSP.CGF	ESP	ESP.CGF
$n = 50$	2.76	6.24	10.79	5.41	3.46
$n = 500$	5.91	12.55	59.26	13.82	11.06

(b) Running time with $p = 10$

Table 5.2: Accumulated time each algorithm spends over the course of 1000 simulations with number of nuisance parameters $p = 3$ (upper) and $p = 10$ (lower). The headings denote the methods Regular (score test), DSP (Double SaddlePoint implemented with MLE-method), DSP.CGF (Double SaddlePoint implemented with CGF-method), ESP (Effective (SaddlePoint) score test implemented with MLE-method), ESP.CGF (Effective (SaddlePoint) score test implemented with CGF-method).

	Regular	DSP	DSP.CGF	ESP	ESP.CGF
$n = 50$	0	23	23	20	15
$n = 500$	0	0	1	0	1

(a) Invalid values for $p = 3$

	Regular	DSP	DSP.CGF	ESP	ESP.CGF
$n = 50$	0	59	110	37	36
$n = 500$	0	1	133	1	1

(b) Invalid values for $p = 10$

Table 5.3: Number of `NA`'s or other invalid p-values out of 1000 simulations with number of nuisance parameters $p = 3$ (upper) and $p = 10$ (lower).

Unsurprisingly, we once again see that the double saddlepoint method implemented with the CGF-method takes the most time (table 5.2), and also produces the most invalid values (5.3). Furthermore, we see that the effective score test implemented with MLE-method is slower than the double saddlepoint approximated score test implemented with MLE-method when $n = 500$. In Johnsen et al. (2023; p.2755), they write that "The [effective score] test is considerably faster to compute [than the double saddlepoint approximated score test]". This is true for the CGF-methods, which is what Johnsen et al. (2023) implemented, but we see that for this example the same does not hold for the MLE-methods. This is not necessarily surprising, as the effective score test requires matrix multiplication in order to determine $\tilde{\boldsymbol{Z}}$, as well as the manually determined intercept, or offset, in the "full" model (see algorithm 1 for details). These matrix operations will become more tedious when either n or p becomes larger. Despite this, the effective score test using CGF-method seems to be the fastest implementation of all, thus supporting the idea that the effective score test can be used as a computationally cheaper alternative to a double saddlepoint approximated score test if implemented efficiently.

Lastly, we see that all methods produce more invalid values for small n, except for the double saddlepoint approximated score test with CGF-method when $p = 10$, indicating that, in general, the methods are more stable for large sample sizes.

5.3 Comparison of tests

For this section, we consider two examples where we are interested in comparing how the three different score tests are able to control the level of a hypothesis test

$$H_0 : \gamma \leq 0 \quad \text{v.s.} \quad H_1 : \gamma > 0. \tag{5.2}$$

In order to investigate the ability of each test to control level, we generate 50000 data sets $(y_i, \boldsymbol{x}_i, z_i)$, with $i = 1, \ldots, n$, with nuisance covariates $X_{i1} \ldots, X_{i5}$ generated from a predetermined, continuous[3] distribution and $Z_i \sim \text{Gamma}(1, 3)$. Each response Y_i is simulated from a generalized linear model with a linear predictor

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \gamma_0 z_i, \tag{5.3}$$

where $\gamma_0 = 0$. We then apply the regular score test, the double saddlepoint approximated score test, and the effective score test, and compare the computed p-values as well as the number of p-values below the decided significance level $\alpha = 0.005$, which are the number of p-values that would have led to a rejection of the null hypothesis. The null-hypothesis is clearly true for this example, and a test should therefore reject H_0 in approximately $\alpha = 0.005$ times of the cases for us to say that it is able to control the significance level. Based on the discussions in section 5.1 we only implement the MLE-methods, as it seems that the tests are indifferent to the implementation method up to algorithmic error.

[3] The covariates can very well be discrete as well, but in the simulated examples we wished to generate 50000 samples while keeping an imbalanced response, and this was easier to achieve with continuous covariates.

5.3.1 Poisson regression with imbalanced response

For the first example, we consider a Poisson regression model where the responses are imbalanced, by which we mean that there is a prevalence of zeros in the responses. Let $Y_i \sim \text{Pois}(\mu_i)$, with $i = 1, \dots, n$, with μ_i relating to the linear predictor from equation (5.3) by

$$\eta_i = \ln(\mu_i).$$

The simulated covariates and chosen coefficients yielded on average $\mu_i \approx 0.1$. We will consider three different sample sizes $n = 150, 500, 1000$.

We consider first a plot where the p-values from the different tests are plotted against one another. Considering the smallest sample size, $n = 150$ first, we see from Figure 5.6 that the regular score test computes more significant p-values than the other two tests. This makes sense as the normal distribution often tends too fast towards zero in the tail, which will yield inflated p-values when the distribution of the score has not converged to a normal distribution yet. The effective score test and double saddlepoint approximated score test seem to agree quite well.

In Figure 5.7 we investigate the proportion of tests that lead to a rejection of the null-hypothesis for the three different tests and sample sizes, and consider this to be an estimate of the level of each respective test. We illustrate the uncertainty in our simulations with Clopper-Pearson confidence intervals. Letting K_j be the number of significant p-values computed from test j, and let $S = 50000$ be the number of simulated data sets. Then K_j can be thought of as a random variable from a binomial distribution $\text{Binom}(p_j, S)$ with p_j being the unknown, true probability of making a type I error for that model when using test j. The confidence intervals are found using a beta distribution, as described in subsection 2.1.3. We see that the effective score test and double saddlepoint approximated score test are both able to control the level of the test, whereas the regular score test is not able to control the level at all. The normal approximation is better for larger sample size, but even at $n = 1000$ it is not yet near the chosen level of the test, which is marked in a red, dashed line.

We also note that from Figure 5.8, all three methods seem to converge towards the same test as the sample size $n \to \infty$. This is consistent with the discussion in subsection 4.3.1 as well as the property that saddlepoint methods are exact for normally distributed random variables.

5.3.2 Logistic regression with small sample size

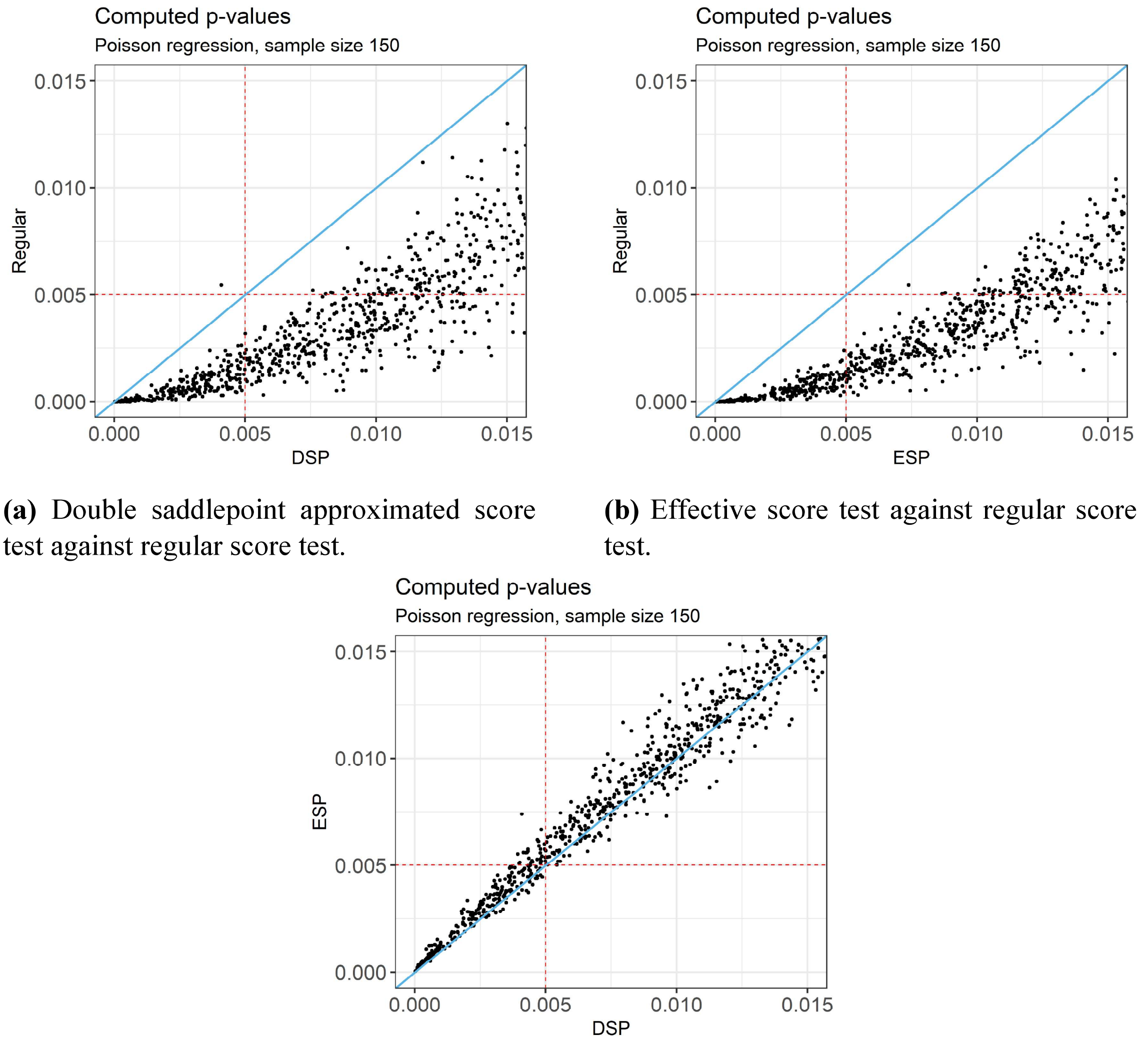
We now consider a logistic regression model, and this time we consider smaller data sets of sample sizes $n = 35, 50, 100$. Letting $Y_i \sim \text{Binom}(\mu_i)$ for $i = 1, \dots, n$, with μ_i relating to the linear predictor from equation (5.3) through

$$\eta_i = \ln\left(\frac{\mu_i}{1-\mu_i}\right).$$

The simulated covariates and chosen coefficients for this example yielded an average value of $\mu_i \approx 0.32$.

We compare the methods as in the previous example, by first plotting each p-values against one another for the smallest sample size $n = 35$. As can be seen from Figure 5.9, the regular score test in general produces more significant p-values than the other two tests. However, at this small sample size, all methods seem less consistent with each other overall, and we note that the effective score test seems to disagree more with the double saddlepoint approximated score test in this example, as compared to the Poisson regression model with larger sample size.

From Figure 5.10, we again consider the proportion of tests that lead to a rejection of the null-hypothesis, which we consider to be an estimation of the level of the test. We include Clopper-Pearson confidence intervals to signify uncertainty in our estimations. We see that both the effective score test and the regular score test produce inflated p-values, and the only test that is able to control the level of the test for this example is the double saddlepoint approximated score test. From this we can speculate that the projection to eliminate nuisance parameters is not precise enough for small sample sizes, thus making the effective score test in practice as bad as the regular score test. However, for all sample sizes $n = 35, 50, 100$, the effective score test is closer to the decided significance level than the regular score test. Also for this example, the asymptotic tendency of the tests to become similar as $n \to \infty$ can be seen in Figure 5.11.



(a) Double saddlepoint approximated score test against regular score test.

(b) Effective score test against regular score test.

(c) Double saddlepoint approximated score test against effective score test.

Figure 5.6: Comparison of the computed p-values from each of the three tests. The plot is enhanced in the tail, in order to compare the smallest p-values. The red, dashed line marks the chosen significance level.

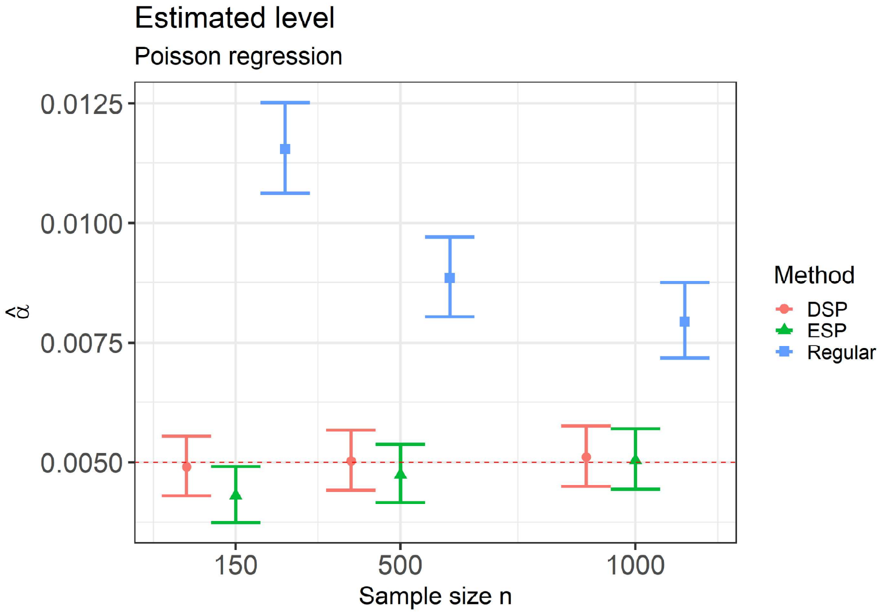
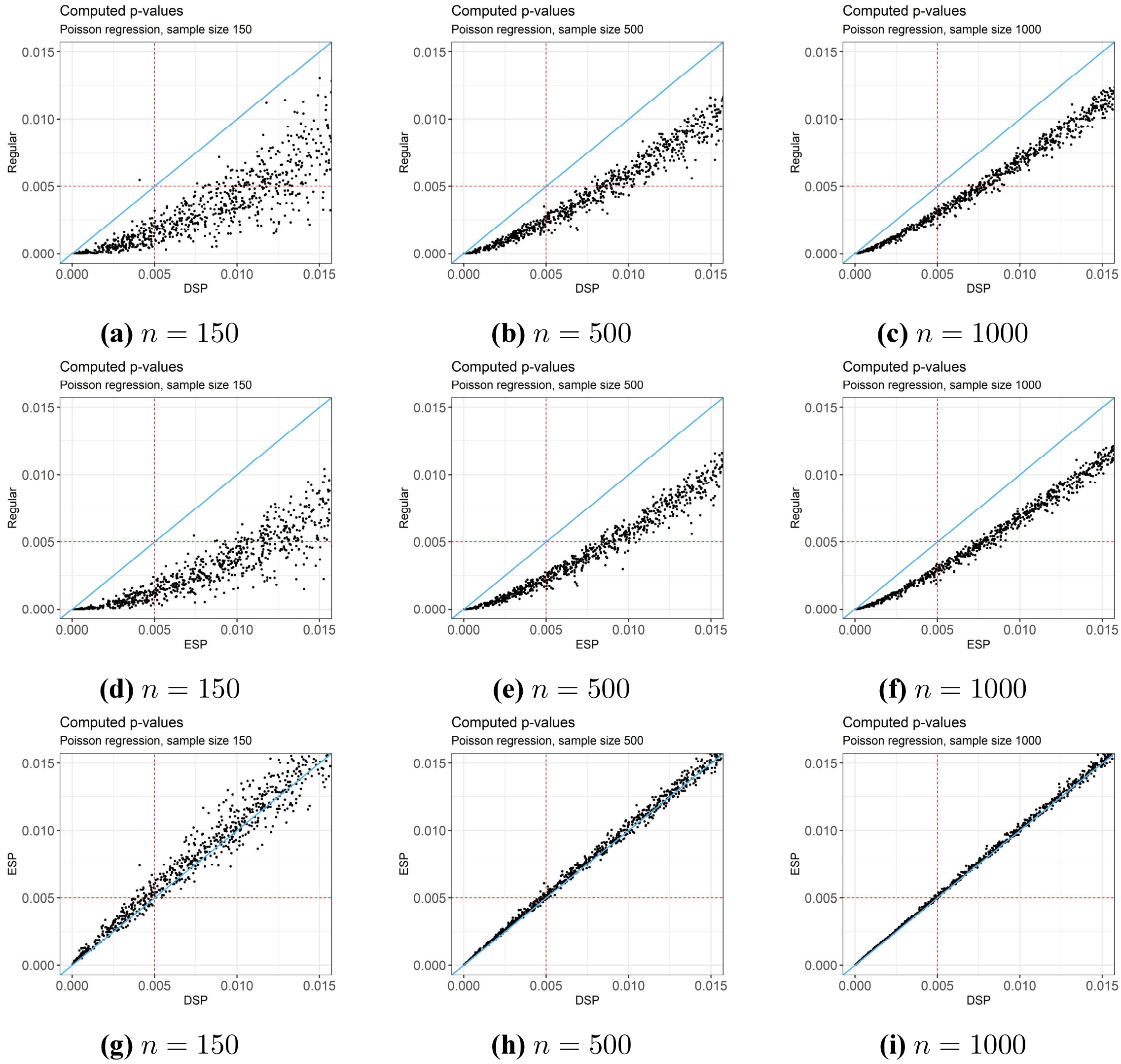


Figure 5.7: Estimated level over 50000 simulations of each test for the three different sample sizes, with Clopper-Pearson confidence intervals. The red, dashed line marks the chosen significance level.

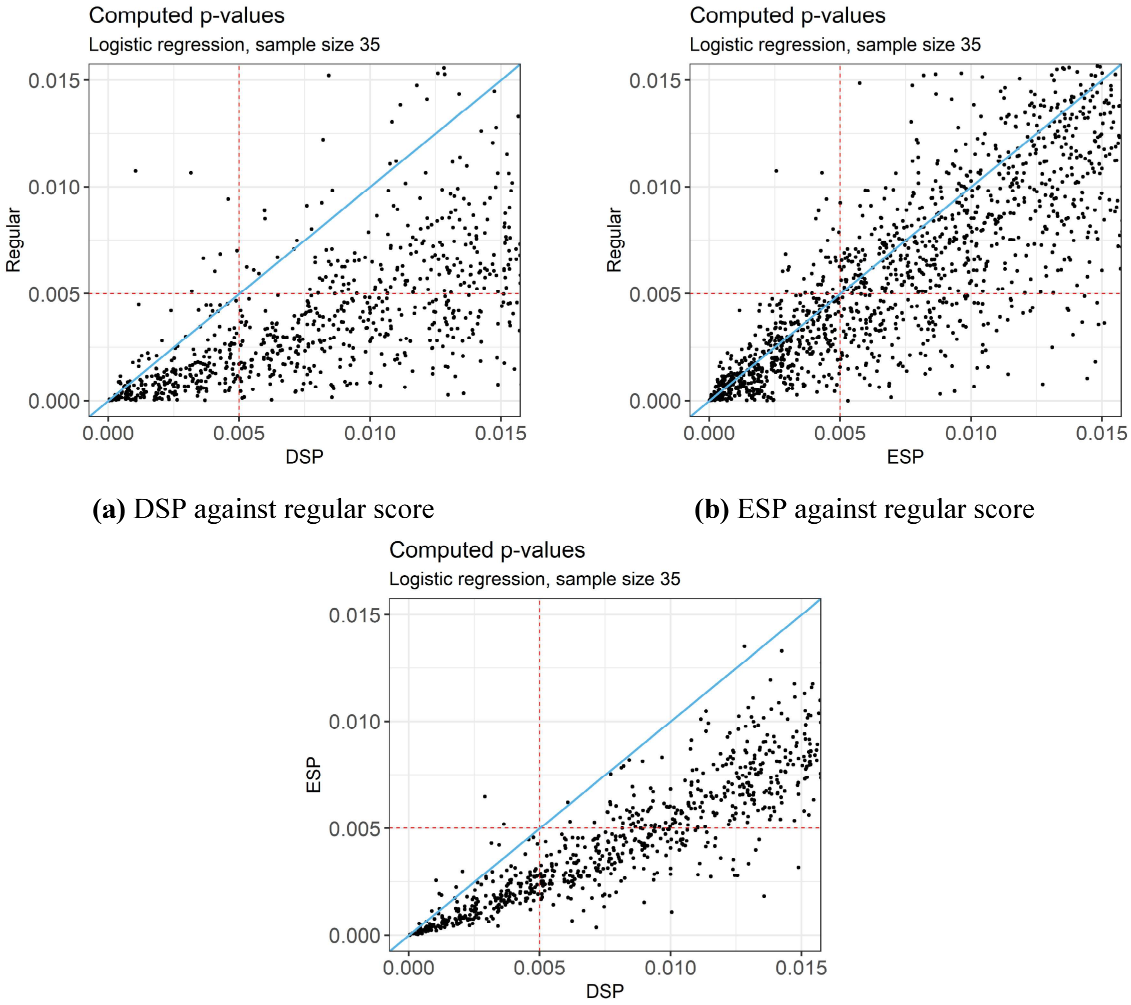


(a) $n = 150$ **(b)** $n = 500$ **(c)** $n = 1000$

(d) $n = 150$ **(e)** $n = 500$ **(f)** $n = 1000$

(g) $n = 150$ **(h)** $n = 500$ **(i)** $n = 1000$

Figure 5.8: Comparing the p-values near the tail computed by double saddlepoint approximated score test against regular score test (upper), effective score test against regular score test (middle) and double saddlepoint approximated score test against effective score test (lower) for the three different sample sizes.



(a) DSP against regular score

(b) ESP against regular score

(c) DSP against ESP

Figure 5.9: Copmarision of the computed p-values for each of the three methods for $n =$ 35.The plot is enhanced in the tail, in order to compare the smallest p-values. The red, dashed line marks the chosen significance level.

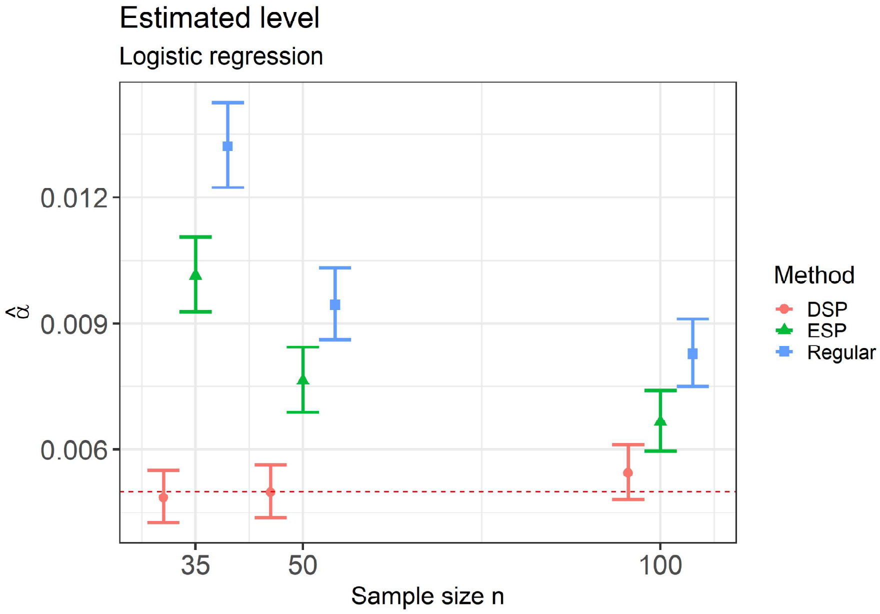
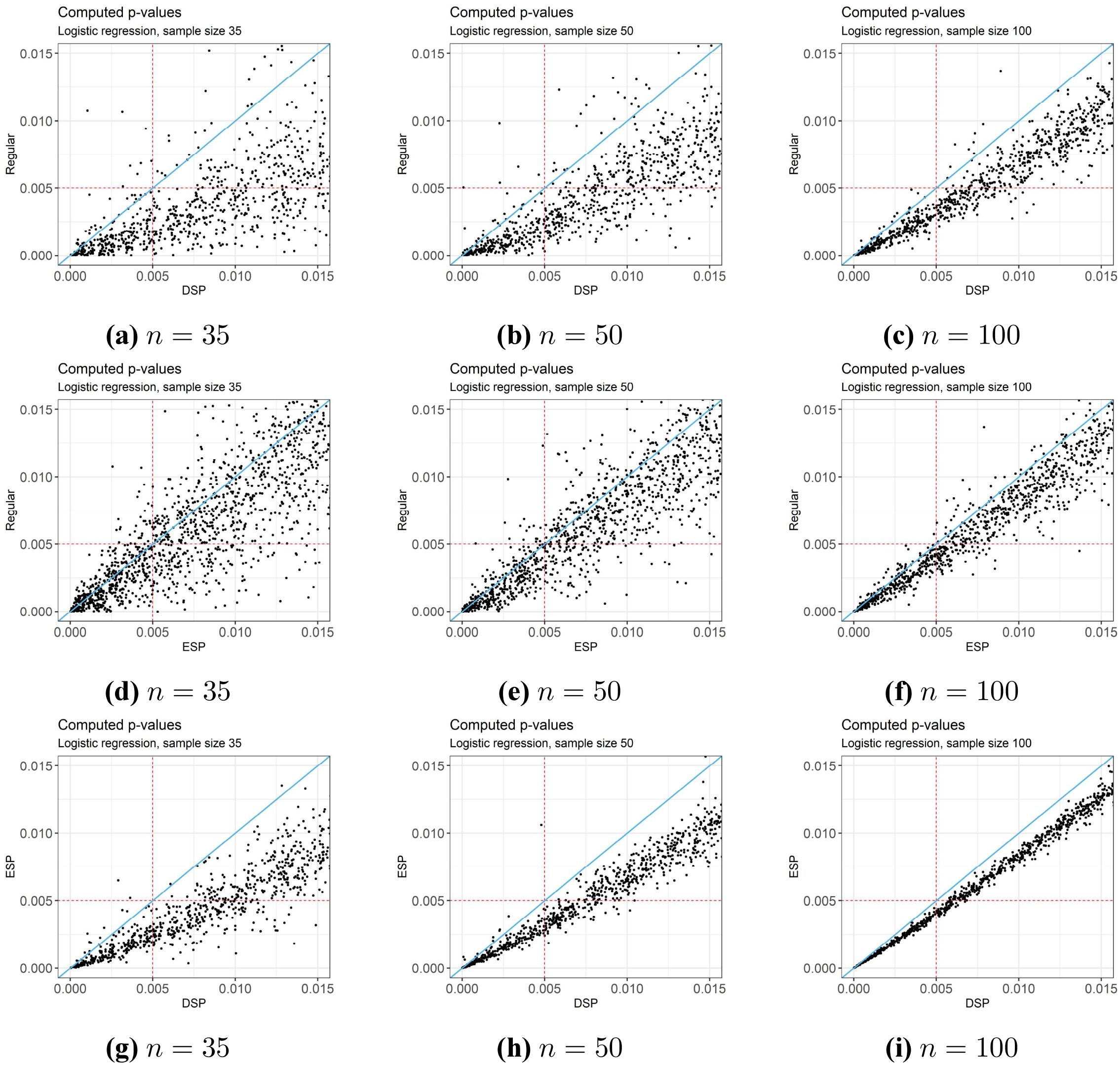


Figure 5.10: Estimated level over 50000 simulations of each test for the three different sample sizes, with Clopper-Pearson confidence intervals. The red, dashed line marks the chosen significance level.



(a) $n = 35$

(b) $n = 50$

(c) $n = 100$

(d) $n = 35$

(e) $n = 50$

(f) $n = 100$

(g) $n = 35$

(h) $n = 50$

(i) $n = 100$

Figure 5.11: Comparing the p-values near the tail computed by double saddlepoint approximated score test against regular score test (upper), effective score test against regular score test (middle) and double saddlepoint approximated score test against effective score test (lower) for the three different sample sizes.

Chapter 6

Applying the methods to real data sets

In this chapter, we apply the discussed methods to two real data sets. We consider a small data set first, and then a large data set with an imbalanced response. We emphasize that it is, as before, the three score tests and a comparison and evaluation of those that is the goal of this thesis and therefore also this chapter, as opposed to performing complete inference of the data at hand.

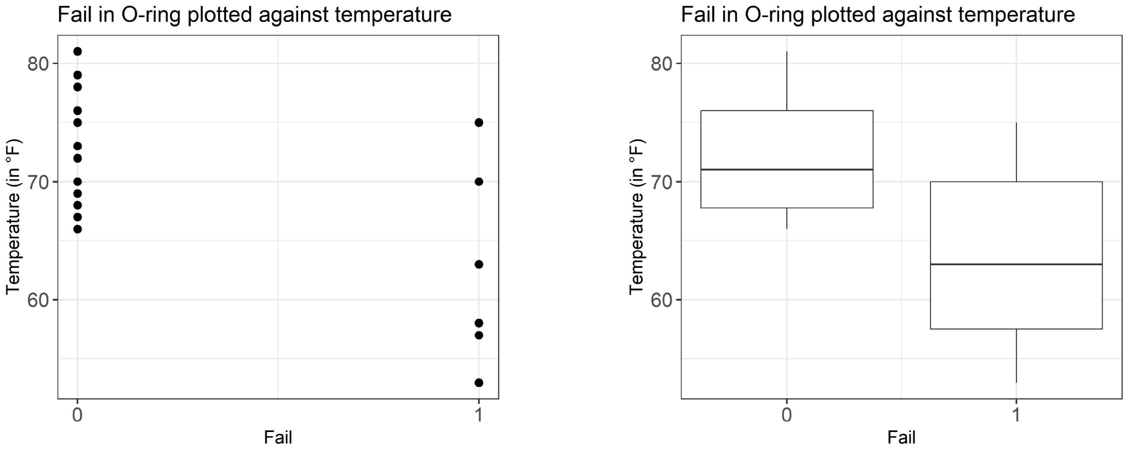
6.1 Challenger disaster

The Challenger disaster refers to the tragic incident that occurred the 28th of January 1986, when the Challenger space shuttle exploded only 73 seconds after launch, killing all of its seven crew members. The cause was later determined to be the result of a failure in the rubber O-rings, which was likely caused by the cold weather on the day of the launch ('Space Shuttle *Challenger* disaster' 2023).

Test data collected before the launch did reveal that there might be a increased risk of O-ring failure in low temperatures. This data set is openly available, for instance in the `R`-package `alr4` Weisberg (2014) by the name `Challeng`. The data set consists of 23 observations and 7 covariates. However, we consider only a subset of the data set containing the continuous covariate `temp`, which is the air temperature at launch, and a covariate `fail` denoting the number of O-rings that failed, which we encode binary with 0 indicating no fails, and 1 indicating at least one O-ring failed. See Figure 6.1 for a plot of these two variables against one another.

We fit a logistic regression model with `fail` as the response. The linear predictor consists only of an intercept and the covariate `temp`, $\boldsymbol{Z}$, meaning

$$\eta_i = \beta_0 + \gamma z_i.$$



(a) Datapoints

(b) Boxplot

Figure 6.1: Plotting the occurrence of one more O-ring failure during testing against the temperature.

	Regular	DSP	ESP
p-value	0.003582	0.004010	0.003621

Table 6.1: The p-values from each of the three discussed methods for the hypothesis test posed in equation 6.1.

We propose to test the hypothesis

$$H_0 : \gamma \geq 0 \quad \text{v.s.} \quad H_1 : \gamma < 0, \tag{6.1}$$

meaning we investigate if low temperatures increase the probability of an O-ring failure.

We run the three score tests, regular score test, effective score test, and double saddlepoint approximated score test, and the resulting p-values are given in table 6.1. The three score tests all indicate a significant association. Of the three computed p-values, the double saddlepoint approximated score test is the most conservative, with the effective score test giving a p-value closer to a regular score test, and the regular score test giving the lowest, or most extreme, p-value. This aligns with what we saw in subsection 5.3.2, with double saddlepoint approximated score test being the most conservative, and effective score test being closer to the regular score test than the double saddlepoint approximated score test.

6.2 Investigating relations between neurons in the brain

In this chapter, we look at a large data set with an imbalanced response that is fitted with a logistic regression model. The data set consists of observations of so-called spikes, or signs of activity, in different neurons in the brain of mice. In neurological research, the brain can be thought of as a network where the neurons act as nodes that interact with each other.

An important part of understanding the brain is understanding how the neurons interact, and which neurons influence other neurons' behavior. However, to understand *how* they influence each other we must first determine which neurons seem to be related. In this example, we use score testing to determine which neurons that are predictive for the behavior of our response neuron, and which that are not. The data set can be found openly available at Laboratory (2023). The package `Matrix` (Bates et al., 2023) was used in the pre-processing of the data set.

6.2.1 Setup

The data set consists of observations over 81 minutes for 35 different neurons V_j, for $j = 1, \ldots, 35$. The observations are made into a binary data set with encoding

$$V_{t,j} = \begin{cases} 1 & \text{if neuron } V_j \text{ spiked in time interval } t, \\ 0 & \text{if neuron } V_j \text{ did not spike in time interval } t, \end{cases}$$

where each time interval t is 2ms long.

Remark 6.2.1. Note that this setup will yield a discrete score statistic with unit step length. As noted earlier, in remark 3.2.2, using a continuous saddlepoint approximation on a discrete random variable will yield mid-p-values.

To investigate the relationship between neurons V_j and V_i, we assume a logistic regression model with $V_{t,j} \sim \text{Bernoulli}(p_{t,j})$, and with canonical link function

$$\text{logit}(p_{t,j}) = \eta_{t,j} = \beta_0 + \omega_{j,i} V_{t-1,i}.$$

for $t = 2, \ldots, n$. The hypothesis test we are interested in is

$$H_0 : \omega_{j,i} \leq 0 \quad \text{v.s.} \quad H_1 : \omega_{j,i} > 0.$$

This is a right-tail test, which means that we are testing whether a spike in neuron V_i at time $t - 1$ increased the probability of a spike in neuron V_j at time t. A left-tail test would correspond to testing if a spike in neuron V_i at time $t - 1$ decreased the probability of a spike in neuron V_j at time t.

The full data set consists of approximately 2.5 million observations. However, since the observations in this data set are time-dependent, and independent observations is an underlying assumption in generalized linear models, we consider only every 10th observation. The data set we work with therefore has a sample size of $n = 243360$.

Testing for neuron 1

We consider neuron 1, $V_{t,1}$, as our response, and consider a simple linear predictor consisting only of an intercept and the neuron we are interested in testing, meaning

$$\eta_{t,1} = \beta_0 + \omega_{1,i} V_{t-1,i}.$$

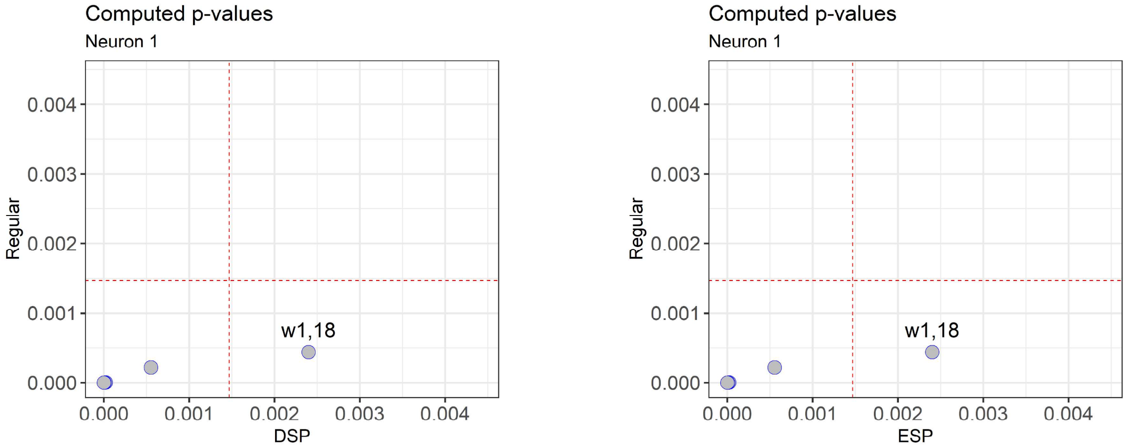
There were 1646 spikes in neuron 1, which is a frequency of about 0.68%.

We perform 34 hypothesis tests,

$$H_0 : \omega_{1,i} \leq 0 \quad \text{v.s.} \quad H_1 : \omega_{1,i} > 0,$$

with $i = 2, \ldots, 35$. Since we want an overall significance level $\alpha = 0.05$ we use a Bonferroni correction (see subsection 2.1.2) and evaluate each test at significance level $\alpha_i = 0.05/34$.

In Figure 6.2, we see a plot of the smallest p-values from the double saddlepoint approximated and effective score tests plotted against the regular score test. The p-values situated in the lower right quadrant of the plot signify coefficients that would be categorized as non-significant by a double saddlepoint or effective score test, meaning they keep H_0, but where the regular score test would reject the null-hypothesis. From the plots we see that this only occurred with one coefficient, namely $\omega_{1,18}$.



(a) Regular score test compared to double saddlepoint method.

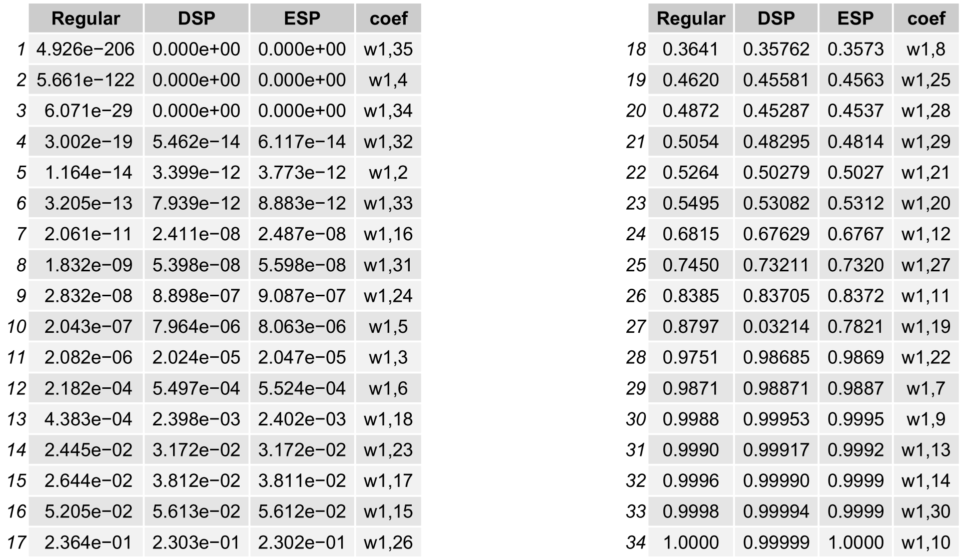
(b) Regular score test compared to effective score test.

Figure 6.2: A plot, enhanced in the tail, showing the p-values of each coefficient $\omega_{1,i}$ for $i = 2, \ldots, 35$.

In Figure 6.3, we see an overview of all p-values computed by each of the tests, along with the coefficient that was tested. Row 13 contains the coefficient $\omega_{1,18}$, which is the only coefficient where the tests conclude differently with respect to rejecting or keeping the null hypothesis strictly based on the decided significance level. This is also the cut-off point for significant covariates.

We see also that the three smallest p-values are 0 for both the double saddlepoint and effective score test. This could indicate some rounding errors in the algorithm for very small values.

Note also row 27, with coefficient $\omega_{1,19}$. Here, the double saddlepoint approximated score test yielded a p-value of 0.03214, whereas the other two tests clearly agreed that this observation should have a p-value around 0.8. This point is also clearly visible in Figure



	Regular	DSP	ESP	coef
1	4.926e−206	0.000e+00	0.000e+00	w1,35
2	5.661e−122	0.000e+00	0.000e+00	w1,4
3	6.071e−29	0.000e+00	0.000e+00	w1,34
4	3.002e−19	5.462e−14	6.117e−14	w1,32
5	1.164e−14	3.399e−12	3.773e−12	w1,2
6	3.205e−13	7.939e−12	8.883e−12	w1,33
7	2.061e−11	2.411e−08	2.487e−08	w1,16
8	1.832e−09	5.398e−08	5.598e−08	w1,31
9	2.832e−08	8.898e−07	9.087e−07	w1,24
10	2.043e−07	7.964e−06	8.063e−06	w1,5
11	2.082e−06	2.024e−05	2.047e−05	w1,3
12	2.182e−04	5.497e−04	5.524e−04	w1,6
13	4.383e−04	2.398e−03	2.402e−03	w1,18
14	2.445e−02	3.172e−02	3.172e−02	w1,23
15	2.644e−02	3.812e−02	3.811e−02	w1,17
16	5.205e−02	5.613e−02	5.612e−02	w1,15
17	2.364e−01	2.303e−01	2.302e−01	w1,26
18	0.3641	0.35762	0.3573	w1,8
19	0.4620	0.45581	0.4563	w1,25
20	0.4872	0.45287	0.4537	w1,28
21	0.5054	0.48295	0.4814	w1,29
22	0.5264	0.50279	0.5027	w1,21
23	0.5495	0.53082	0.5312	w1,20
24	0.6815	0.67629	0.6767	w1,12
25	0.7450	0.73211	0.7320	w1,27
26	0.8385	0.83705	0.8372	w1,11
27	0.8797	0.03214	0.7821	w1,19
28	0.9751	0.98685	0.9869	w1,22
29	0.9871	0.98871	0.9887	w1,7
30	0.9988	0.99953	0.9995	w1,9
31	0.9990	0.99917	0.9992	w1,13
32	0.9996	0.99990	0.9999	w1,14
33	0.9998	0.99994	0.9999	w1,30
34	1.0000	0.99999	1.0000	w1,10

Figure 6.3: A table showing all the p-values computed by each score test, sorted according to the regular score test.

6.4, where the p-values of the effective score test and double saddlepoint approximated score test are plotted against each other. Quite interestingly, neuron 19 is the neuron with the lowest number of positive responses. This neuron only had 202 spikes, or 0.083% responses that are 1. It is plausible that this imbalance is related to why the double saddlepoint algorithm produced such a low p-value. However, as we have no reason to believe that the regular score should indicate a p-value in the wrong tail, we can reasonably suspect that the p-value computed by the double saddlepoint approximated score test is an algorithmic error, and not close to a true p-value.

6.2.2 Discussion

The example discussed here is quite brief, and to perform a complete analysis we would have to repeat what we did for neuron 1 above for the remaining 35 neurons, as $\omega_{i,j} \neq \omega_{j,i}$. However, based on what we saw from this example we can make some general observations.

First and foremost, we see that the different methods did not yield vastly different p-values. In particular, sorting the different p-values from smallest to largest with respect to the different methods would yield the same order for all coefficients where the null-hypothesis were rejected. On the other hand, we see also that for coefficient $\omega_{1,19}$ the effective score test and regular score test disagree, indicating that the need for reference methods is still there. We remember that neuron 19 was the neuron with the least spikes, in other words the most imbalanced covariate, and this could be the reason the effective score and regular score test are so different for this coefficient. Hence, the usefulness of the methods lays not necessarily in them as alternatives to a regular score test, but rather as complementing

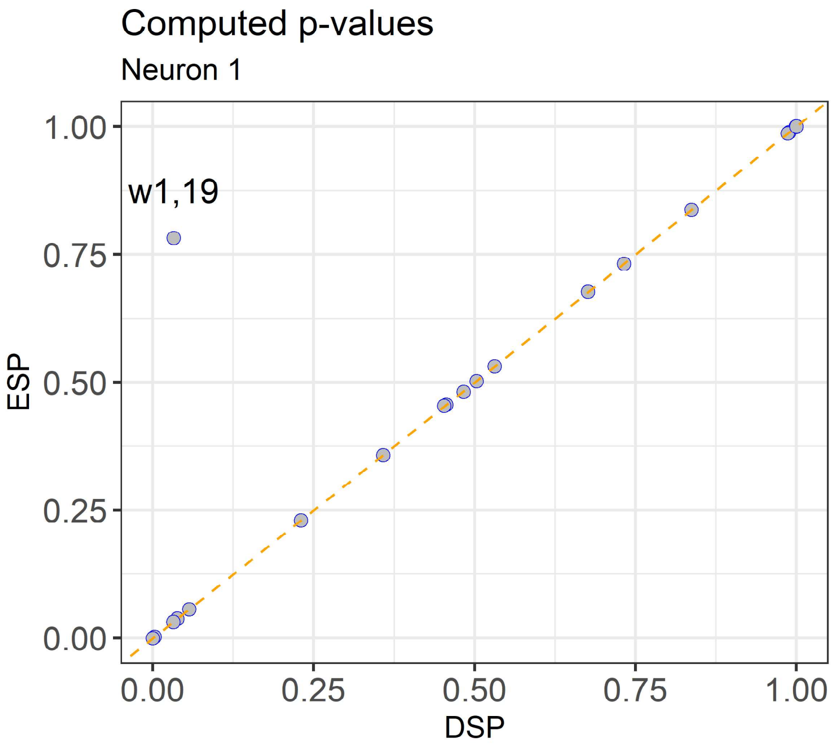


Figure 6.4: A plot of the computed p-values from the double saddlepoint and effective score test.

methods.

We note also that both the double saddlepoint approximated score test and the effective score test are time consuming methods, and it would take much time to compute every p-value three times. Therefore, for situations where we are interested in computing p-values for multiple different hypothesis test, in order to also reduce the number of demanding computations needed we can opt to only compute p-values “close” to the desired significance level with one of the alternative methods. This could be done by first computing all p-values of the test with a regular score test. Then, for p-values such that $p \in [\alpha - d, \alpha + d]$ for some chosen significance level α and distance d, the p-values are computed again using either double saddlepoint approximated score test or effective score test, or both. This would also dodge the problem concerning the singularity near the center of the distribution, by simply choosing d such that $\alpha + d << 0.5$.

Chapter 7

Final discussion

This thesis aimed to present and compare two alternative methods for computing p-values of score tests in the presence of nuisance parameters, that would ensure better control over type I error even with small sample sizes or data sets with imbalanced response. The two alternative methods, namely the double saddlepoint approximated score test and the effective score test, both utilized saddlepoint approximation to estimate the distribution of the score. The distinction between them lies in how they handle nuisance parameters. The double saddlepoint approximation computed the conditional distribution of the score, thus it is a form of conditional inference. The effective score test transforms the score statistic to a decorrelated effective score where the p-value is computed using unconditional inference.

Furthermore, we have shown that the saddlepoint distribution of both the regular score and effective score can be easily implemented in terms of its maximum likelihood estimates. The simulations in section 5.1, show that the approximation in terms of maximum likelihood estimates aligned with the method employing cumulant generating functions for the effective score test. The MLE-method is easier to implement for different models, as we do not have to compute the unique cumulant generating function for our specific model. However, we saw also that the effective score test using CGF-method was slightly faster, thus making both implementations for effective score test good candidates. For the double saddlepoint approximated score test, the theoretical calculations in chapter 3 indicates that they should be equivalent, but we saw from the simulations in section 5.1 and 5.2, that the MLE-method appear to be both easier to implement and more stable for large number of covariates, thus making it preferable over the CGF-method.

For computing p-values, the examples discussed in chapter 5 indicated that the double saddlepoint approximated score test controlled the type I error better than the regular score test, and that the effective score test also controlled the type I error better, but relied more on the sample size of the data than a double saddlepoint approximated score test. However, these alternative tests have some drawbacks. They are computationally more demanding and less robust near the singularity for p-values close to 0.5. These challenges might be mitigated through improved implementations. Nonetheless, in situations where multiple p-values are computed for numerous hypothesis tests, as shown in section 6.2, it is possible to reduce the computational burden by employing the alternative methods only for p-values ”close” to the desired significance level as discussed in subsection 6.2.2.

While this thesis predominantly focuses on p-values, it is widely acknowledged among statisticians that blindly relying on p-values may not be advisable, even if the test adequately controls type I error. However, the findings presented here can still be a valuable tool. Firstly, p-values continue to be widely used in various fields that employ statistical tests. Thus, the need for methods that are able to approximate p-values with high accuracy when other methods fail is an addition to the tools of statistical inference. Secondly, the double saddlepoint approximated score test and the effective score test can enhance the reliability of other methods, or work as a complementing quality check of the methods we would commonly rely on. As noted by Pierce and Peters (1992), “the application of the higher order methods frequently serves more to verify [first order asymptotic methods] than to provide a substantial improvement.[...] practical work ordinarily does not require extremely precise calculation of p-values and confidence limits, and an important role of higher order asymptotics is to improve one feel’s for the adequacy of standard first-order methods.” In other words, the effective score test and double saddlepoint approximated score test give additional verification and ensure higher confidence in the accuracy of our methods.

7.1 Further work

Several aspects were not covered in this thesis, but could be pursued for future research. Since the saddlepoint distribution is typically asymmetric, there is no general scheme presented here for computing two-sided tests. This becomes more problematic when considering continuity-corrected p-values, as discussed both in Butler (2007) and in Johnsen et al. (2023), but this topic has been omitted here due to simplicity.

Furthermore, the *power* of a test, meaning to what degree the test is able to correctly reject H_0 when the null-hypothesis is false, has not been addressed. Within the set of level α tests, power becomes an important measure for comparing the effectiveness of different tests. Hence, it would be interesting to compare the power of a double saddlepoint approximated score test against an effective score test.

Lastly, it is worth noting that when employing a saddlepoint approximation based on maximum likelihood estimates, the “score test” appears more as a likelihood ratio test, as both a model under the null hypothesis as well as a full model need to be fitted. However, we have shown in section 3.3 that the computed p-values do correspond to a score test. To expand on this idea, Jensen (1992) introduces the concept of *the modified signed likelihood statistic*, which appears very similar to the ideas discussed in this thesis. This can be seen by considering chapter 3 as well as lemma 2.1 in Jensen (1992). Comparing a test based on the modified signed likelihood statistic with the double saddlepoint approximated score test and effective score tests presented in this thesis would therefore be an intriguing avenue for future exploration.

Bibliography

Bates, D., Maechler, M., & Jagan, M. (2023). *Matrix: Sparse and dense matrix classes and methods* [R package version 1.5-4.1]. https://CRAN.R-project.org/package=Matrix

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press.

Butler, R. W. (2007). *Saddlepoint approximations with applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511619083

Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd). Duxbury Press.

Choi, S., Hall, W. J., & Schick, A. (1996). Asymptotically uniformly most powerful tests in parametric and semiparametric models [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *24*(2), 841–861. Retrieved April 19, 2023, from https://www.jstor.org/stable/2242678

Daniels, H. E. (1954). Saddlepoint approximations in statistics [Publisher: Institute of Mathematical Statistics]. *The Annals of Mathematical Statistics*, *25*(4), 631–650. Retrieved September 5, 2022, from https://www.jstor.org/stable/2236650

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *American Journal of Human Genetics*, *101*(1), 37–49. https://doi.org/10.1016/j.ajhg.2017.05.014

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in r*. Springer New York. https://doi.org/10.1007/978-1-4419-0118-7

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34333-9

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6082]. *Statistics in Medicine*, *33*(11), 1946–1978. https://doi.org/10.1002/sim.6082

Hall, W. J., & Mathiason, D. J. (1990). On large-sample estimation and testing in parametric models [Publisher: [Wiley, International Statistical Institute (ISI)]]. *International Statistical Review / Revue Internationale de Statistique*, *58*(1), 77–97. https://doi.org/10.2307/1403475

Härdle, W. K., & Simar, L. (2015). *Applied multivariate statistical analysis* (4th ed. 2015). Springer Berlin Heidelberg : Imprint: Springer. https://doi.org/10.1007/978-3-662-45171-7

Hemerik, J., Goeman, J. J., & Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(3), 841–864. https://doi.org/10.1111/rssb.12369

Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, *79*(4), 693–703. https://doi.org/10.2307/2337225

Johnsen, P. V., Bakke, Ø., Bjørnland, T., DeWan, A. T., & Langaas, M. (2023). Saddlepoint approximations to score test statistics in logistic regression for analyzing genome-wide association studies [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9746]. *Statistics in Medicine*, *n/a*. https://doi.org/10.1002/sim.9746

Karr, A. F. (1993). *Probability*. Springer New York. https://doi.org/10.1007/978-1-4612-0891-4

Laboratory, I. B. (2023). Data release - brainwide map - q4 2022 [Artwork Size: 12507703 Bytes Publisher: figshare], 12507703 Bytes. https://doi.org/10.6084/M9.FIGSHARE.21400815

Lindsay, B. (1982). Conditional score functions: Some optimality results [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, *69*(3), 503–512. https://doi.org/10.2307/2335985

Lindsey, J. K. (1996). *Parametric statistical inference*. Oxford University Press.

Lugannani, R., & Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables [Publisher: Applied Probability Trust]. *Advances in Applied Probability*, *12*(2), 475–490. https://doi.org/10.2307/1426607

Marohn, F. (2002). *A COMMENT ON LOCALLY MOST POWERFUL TESTS IN THE PRESENCE OF NUISANCE PARAMETERS* [Taylor & francis online]. https://doi.org/10.1081/STA-120002852

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman; Hall.

Pierce, D. A., & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families [Publisher: [Royal Statistical Society, Wiley]]. *Journal of the Royal Statistical Society. Series B (Methodological)*, *54*(3), 701–737. Retrieved June 21, 2023, from https://www.jstor.org/stable/2345853

Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions [Publisher: Applied Probability Trust]. *Journal of Applied Probability*, *24*(4), 875–887. https://doi.org/10.2307/3214212

Smyth, G. K. (2003). Pearson's goodness of fit statistic as a score test statistic. In *Institute of mathematical statistics lecture notes - monograph series* (pp. 115–126). Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215091138

Space shuttle *Challenger* disaster [Page Version ID: 1159378975]. (2023, June 9). In *Wikipedia*. Retrieved June 14, 2023, from https://en.wikipedia.org/w/index.php?title=Space_Shuttle_Challenger_disaster&oldid=1159378975

Team, R. (2021). *RStudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion [Publisher: Institute of Mathematical Statistics and Bernoulli Society]. *Electronic Journal of Statistics*, *8*(1), 817–840. https://doi.org/10.1214/14-EJS909

Waterman, R. P., & Lindsay, B. G. (1996). Projected score methods for approximating conditional scores [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, *83*(1), 1–13. Retrieved September 6, 2022, from https://www.jstor.org/stable/2337428

Weisberg, S. (2014). *Applied linear regression* (4th). Wiley. http://z.umn.edu/alr4ed

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (2nd ed. 2016). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-24277-4

Appendix A

Code used to simulate data sets

Below is the code used to generate simulated data sets for the examples in chapter 5. The seed was set to be 248.

A.1 Example that compared the implementations

Code for generating data sets used in the discussion of different implementations of the saddlepoint approximation (section 5.1).

```
# Generate the nuisance covariates and parameter
X = matrix(c(rep(1,n),rnorm(n,0,0.25),runif(n,-1,0.5),
            rexp(n,1),rgamma(n,1,1),runif(n)),
            ncol = 6, byrow = F)
colnames(X) = c("intercept", "x1", "x2", "x3", "x4", "x5")
beta = c(-1,0.2,1,0.5,0.1,-0.3)


# Generate the parameter of interest and covariates
gamma = 0
Z = rgamma(n,1,3)

# compute mu
mu = exp(X%*%beta + gamma*Z)/(1 + exp(X%*%beta + gamma*Z))

#generate samples
y = rbinom(n,1,mu)
```

A.2 Example that explored different number of nuisance covariates

Code for generating data sets used to investigate how the number of nuisance covariates affected the implementations of the alternative score tests (section 5.2).

```
# Testing with p = 9 + intercept
if (exm == 1){
  # Generate the nuisance covariates and parameter
  X = matrix(c(rep(1,n),rnorm(n,0,0.25),runif(n,-1,0.5),
             rexp(n,1),rgamma(n,1,1),runif(n),
             rbinom(n,1,0.25), rexp(n,1),runif(n,-2,2),
             rpois(n,0.5)), ncol = 10, byrow = F)
  colnames(X) = c("intercept", "x1", "x2", "x3", "x4", "x5",
                  "x6","x7", "x8", "x9")
  beta = c(-3,0.2,1,0.5,0.1,-0.3,1,1,0.5,1)


  # Generate the parameter of interest and covariates
  gamma = 0
  Z = rgamma(n,1,3)

}

# Testing with p = 2 + intercept
else if (exm == 2){
  # Generate the nuisance covariates and parameter
  X = matrix(c(rep(1,n),rnorm(n,0,2),rpois(n,0.5)),
             ncol = 3, byrow = F)
  colnames(X) = c("intercept", "x1", "x2")
  beta = c(-3,1,1)


  # Generate the parameter of interest and covariates
  gamma = 0
  Z = rgamma(n,1,3)

}
# compute mu
mu = exp(X%*%beta + gamma*Z)/(1 + exp(X%*%beta + gamma*Z))

#generate samples
y = rbinom(n,1,mu)
```

A.3 Poisson regression with imbalanced response

Below is the code for generating the imbalanced data sets with a Poisson distributed response, used in the example of subsection 5.3.1.

```
# Generate the nuisance covariates and parameter
X = matrix(c(rep(1,n),rnorm(n,0,0.25),runif(n,-1,0.5),
            rexp(n,1),rgamma(n,1,1),runif(n)),
            ncol = 6, byrow = F)
colnames(X) = c("intercept", "x1", "x2", "x3", "x4", "x5")
beta = c(-2.3,0.2,1,0.2,0.1,-0.3)


# Generate the parameter of interest and covariates
gamma = 0
Z = rgamma(n,1,3)

# compute mu
mu = exp(X %*% beta + gamma*Z)

# generate samples
y = rpois(n,mu)
```

A.4 Logistic regression example

Below is the code for generating the small data sets with a Bernoulli distributed response, used in the example of subsection 5.3.2.

```
# Generate the nuisance covariates and parameter
X = matrix(c(rep(1,n),rnorm(n,0,0.25),runif(n,-1,0.5),
            rexp(n,1),rgamma(n,1,1),runif(n)),
            ncol = 6, byrow = F)
colnames(X) = c("intercept", "x1", "x2", "x3", "x4", "x5")
beta = c(-1,0.2,1,0.5,0.1,-0.3)


# Generate the parameter of interest and covariates
gamma = 0
Z = rgamma(n,1,3)

# compute mu
mu = exp(X%*%beta + gamma*Z)/(1 + exp(X%*%beta + gamma*Z))

#generate samples
y = rbinom(n,1,mu)
```

Appendix B

Code for implementations of the different score tests

Below, the R code used to perform the different score tests is included. The code to compute the cumulant generating function, and the twice derivative of the cumulant generating function is needed for both the effective score test and the double saddlepoint approximated score test, and is included below.

```
# CGF is the cumulant generating function
# s: variable of the function
# x: matrix of covariates
# mu: mean vector
# family: either "binomial" or "poisson"
CGF <- function(s, X, mu, family = "binomial"){
  if (family == "binomial"){
    return(sum(log(1 - mu + mu*exp(X %*% s)))-sum(s * mu %*% X))
  }
  else{
    exp_term <- exp(X %*% s)
    return(sum(mu * (exp_term - 1)) - sum(s * mu %*% X))
  }
}

# CGF.D2 is the twice derivative of the CGF wrt s
CGF.D2 <- function(s, X, mu, family = "binomial"){
  if (family == "binomial"){
    emsx = exp(-X%*%s)
    muexp = mu * (1 - mu) * emsx / (((1 - mu) * emsx + mu)^2)
    return(t(as.numeric(muexp) * X) %*% X)
  }
  else{
    emsx = exp(X%*%s)
    return(t (as.numeric(mu * emsx) * X) %*% X)
  }
}
```

B.1 Score test

This is a manual implementation of a score test. It is also possible to use the function `glm.scoretest` from the package `statmod` (Dunn and Smyth, 2018).

```
# X: matrix-argument with nuisance covariates (including intercept)
# Z: vector-argument with covaraite we want to test
# y: vector of responses
# fam: the exponential family of Y, either "binomial" or "poisson"

score.test.int <- function(y, X, Z, fam = "binomial", left.tail = T){

  # Fit model under H0
  mod0 = glm(y ~ . - 1, data = as.data.frame(cbind(y,X)), family = fam)

  # Get fitted values for mu
  mu.hat = mod0$fitted.values

  # Weights
  if (fam == "binomial"){
    W = mu.hat*(1-mu.hat)
  }
  else {
    W = mu.hat
  }

  # Compute observed score
  u <- t(Z) %*% (y - mu.hat)

  # Compute covariance matrix
  I.1 <- t(Z) %*% (Z * W)
  I.12 <- t(Z) %*% (X * W)
  I.2 <- solve(t(X) %*% (X * W))

  I <- I.1 - I.12 %*% I.2 %*% t(I.12)

  # Return p-value
  pnorm(u,mean = 0, sd = sqrt(I), lower.tail = left.tail)
}
```

B.2 Double saddlepoint approximated score test

In the code below the double saddlepoint approximated score test is implemented using both the cumulant generating function and maximum likelihood estimates.

```
# X: matrix-argument with nuisance covariates (including intercept)
# Z: vector-argument with covaraite we want to test
# y: vector of responses
# fam: the exponential family of Y, either "binomial" or "poisson"

double.sadpnt.int <- function(y, X, Z, fam = "binomial", CGF = F,
                              left.tail = T, init = 0){

  # Fit model under H0
  mod0 <- glm(y ~ . - 1, family = fam, data = as.data.frame(cbind(y,X)))

  if (CGF){

    # Get fitted values for mu
    mu.hat = mod0$fitted.values


    # compute the score
    u = t(Z)%*%(y-mu.hat)

    x = cbind(Z,X)
    nx = dim(x)[2]
    x2 = x[,2:nx]

    # Estimate the saddlepoint
    s.hat = optim(rep(init,nx),
                function(s)CGF(s,x,mu.hat, family = fam)-u*s[1])$par

    # Compute w.hat and u.hat according to Butler p.12
    w.hat = sign(s.hat[1])*sqrt(2*(-CGF(s.hat,x,mu.hat, family = fam)+
                s.hat[1]*u))
    u.hat = s.hat[1]*sqrt(det(CGF.D2(s.hat,x,mu.hat, family = fam))/
                    det(CGF.D2(rep(0,nx-1),x2,mu.hat, family = fam)))

    # Compute left-tail p-value according to Butler p.12
    p.val = pnorm(w.hat) + dnorm(w.hat)*((1/w.hat)-(1/u.hat))

    # Return p-value in either left or right tail
    if (left.tail){
      return(p.val)
    }
    else{
      return(1 - p.val)
    }
  }

  else {
    # Fit model under H1
    mod1 <- glm(y ~ . - 1, family = fam, data = as.data.frame(cbind(y,X,Z)))

    # Compute covariance matrix for restricted and full model
    I0.inv = vcov(mod0)
    I1.inv = vcov(mod1)

    # Compute the log-likelihood for restricted and full model
    ll0 = as.numeric(logLik(mod0))
    ll1 = as.numeric(logLik(mod1))

    # Get the MLE estimate for the parameter of interest
    coefZ.hat = as.numeric(coef(mod1)["Z"])

    # Compute w.hat and u.hat according to Butler p.170
    w.hat = sign(coefZ.hat)*sqrt(2*ll1-2*ll0)
    u.hat = coefZ.hat*sqrt((1/det(I1.inv))/(1/det(I0.inv)))

    # Compute left-tail p-value according to Butler p.113
    p.val = pnorm(w.hat) + dnorm(w.hat)*((1/w.hat)-(1/u.hat))

    # Return p-value in either left or right tail
    if (left.tail){
      return(p.val)
    }
    else{
      return(1 - p.val)
    }
  }
}
```

B.3 Effective score test

In the code below the effective score test is implemented using both the cumulant generating function and maximum likelihood estimates.

```
# X: matrix-argument with nuisance covariates (including intercept)
# Z: vector-argument with covaraite we want to test
# y: vector of responses
# fam: the exponential family of Y, either "binomial" or "poisson"
# mu: fitted values for mu under H0
```

```
effectiveZ <- function(X, Z, mu, family = "binomial"){

  # Weights
  if (family == "binomial"){
    W = mu*(1-mu)
  }
  else {
    W = mu
  }

  # Compute Z*
  M = solve(t(X * W) %*% X) %*% t(X * W) %*% Z
  Z.star = Z - X %*% M
  # Return Z*
  return(Z.star)
}



effective.sadpnt.int <- function(y, X, Z, fam = "binomial", CGF = F,
                                left.tail = T){

  # Model under H0
  mod0 <- glm(y ~ . -1, family = fam, data = as.data.frame(cbind(y,X)))

  # Get fitted values for mu
  mu.hat = mod0$fitted.values

  # Compute Z*
  Z.star = effectiveZ(X, Z, mu.hat, family = fam)
  u.e = t(Z.star)%*%(y-mu.hat)
  # can also use u = t(Z)%*%(y-mu.hat)

  if (CGF){
    # Estimate the saddlepoint
    s.hat = optimize(function(s)CGF(s,Z.star,mu.hat,family = fam)-u.e*s,
                    c(-10,10),tol=1e-10)$minimum

    # Compute w.hat and u.hat according to Butler p.12
    w.hat = sign(s.hat)*sqrt(2*(s.hat*u.e-
                            CGF(s.hat,Z.star,mu.hat,family = fam)))
    u.hat = s.hat*sqrt(CGF.D2(s.hat, Z.star, mu.hat, family = fam))

    # Compute left-tail p-value according to Butler p.12
    p.val = pnorm(w.hat) + dnorm(w.hat)*((1/w.hat)-(1/u.hat))

    # Return p-value in either left or right tail
    if (left.tail){
      return(p.val)
    }
    else{
      return(1 - p.val)
    }
  }

  else {

    # Weights
    if (fam == "binomial"){
      theta = log(mu.hat/(1-mu.hat))
    }
    else {
      theta = log(mu.hat)
    }

    # Modified model under H1 (correct to remove intercept?)
    mod1 = glm(y ~ -1 + Z.star + offset(theta), family = fam,
                data = as.data.frame(cbind(y,Z.star)))

    # Compute covariance matrix for full model
    I1 = as.numeric(vcov(mod1))

    # Compute the log-likelihood for restricted and full model
    ll0 = as.numeric(logLik(mod0))
    ll1 = as.numeric(logLik(mod1))

    # Get the MLE estimate for the parameter of interest
    coefZ.hat = as.numeric(mod1$coefficients)

    # Compute w.hat and u.hat according to Butler p.170
    w.hat = sign(coefZ.hat)*sqrt(2*ll1-2*ll0)
    u.hat = coefZ.hat*sqrt(1/I1) # Hvorfor er det 1/ her?

    # Compute left-tail p-value according to Butler p.12
    p.val = pnorm(w.hat) + dnorm(w.hat)*((1/w.hat)-(1/u.hat))

    # Return p-value in either left or right tail
    if (left.tail){
      return(p.val)
    }
    else{
      return(1 - p.val)
    }
  }
}
```

NTNU

Norwegian University of
Science and Technology