A Simulation Study of Goodness-of-Fit Tests for Binary Regression with Applications to Norwegian Intensive Care Registry Data

Ellisif Nygaard

Supervisor: Geir Drage Berentsen
Øystein A. Haaland

Department of Mathematics
University of Bergen

This dissertation is submitted in partial fulfillment of the requirements for the degree of

Master of Science in Statistics (Data Analysis)

The Faculty of Mathematics and
Natural Sciences

January 2019

Acknowledgements

I would like to thank my advisers Geir Drage Berentsen and Øystein Ariansen Haaland for introducing me to this topic and their guidance throughout the process of writing this thesis. I am also grateful to the Department of Heart Disease at Haukeland University Hospital for valuable experience, and their patience and generosity. In addition, I would like to thank the Norwegian Intensive Care Registry (NIR) for providing an unique data set.

Many thanks to friends and family for their indispensable cheers of encouragement. I am also grateful for the proofreading and feedback received from Ranveig Nygaard. And I am grateful that this process is finally over.

Abstract

When using statistical methods to fit a model, the consensus is that it is possible to represent a complex reality in the form of a simpler model. It is helpful to systematically measure a model's ability to capture the underlying system which controls the data generation in the population being examined.

One of the possible tools we can apply to evaluate model adequacy is goodness-of-fit (GOF) tests. Summary GOF statistics are computed for a specific fitted model, then attributed an asymptotic distribution, and finally the null hypothesis that the model fits the data adequately is tested. A great challenge, when the model is a binary regression model and it has one or several continuous covariates, is to verify which asymptotic distributions the GOF statistics in fact have (Hosmer et al., 1997).

In this thesis, we will evaluate the validity of the distributions of some established GOF test statistics mentioned in the literature. We have chosen so-called *global* GOF tests, where user input is not necessary. Tests demanding user input, such as the Hosmer-Lemeshow test, have been shown to have some considerable disadvantages. Hosmer et al. (1997) states that number of groups (which are determined by user discretion) can influence whether the GOF test rejects the model fit or not.

Binary regression models present a specific set of challenges with regards to GOF measures, especially in situations where at least one covariate is continuous. There appears to be no broad general agreement on which GOF statistics are reliable options when fitting such models. This thesis aims to extend the current knowledge in this area. A modified version of one of the statistics is introduced. The GOF tests studied are later applied in a data analysis on real data set from the Norwegian Intensive Care Registry (NIR).

An exploration was performed in the attempt to suggest a suitable tool to evaluate the discrepancies between the estimated logistic probabilities and the outcome variable, and how different GOF tests will behave for different categories of discrepancies.

Table of contents

Symbols

Roman Symbols

H_0	The null hypothesis
M_0	The true logistic model
$\hat{S}_{st}$	The standardised USS statistic
X^2_{st}	The standardised Pearson chi-square statistic

Greek Symbols

φ_1, φ_2	The shape parameters of Stukel's generalised model

Acronyms / Abbreviations

GOF	Goodness-of-fit
IMT	Information matrix test
LPM	Linear probability model
LRT	Likelihood ratio test
ML	Maximum likelihood
NIR	The Norwegian Intensive Care Registry
PRD	Predicted risk of death
USS	Unweighted sum-of-squares

Chapter 1

Introduction to Binary Regression

1.1 The Classical Linear Regression Model

Let the data $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, consist of n observations of the continuous response variable y and the k covariates $x_1, \ldots, x_k$. The covariates can be continuous or categorical. In Fahrmeir et al. (2013), the classical linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to be independent and identically normally distributed with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. The model is linear in the parameters $\beta_0, \ldots, \beta_k$, whereas the covariates can be non-linear expressions.

The following quantity, which represents the influence the covariates have on the model, is referred to as a *linear predictor*:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}, \tag{1.1}$$

where $\boldsymbol{x}_i^\mathsf{T} = [1 \;\; x_{i1} \;\; x_{i2} \;\; \ldots \;\; x_{ik}]$ and $\boldsymbol{\beta} = [\beta_0 \;\; \beta_1 \;\; \ldots \;\; \beta_k]^\mathsf{T}$. This can be expressed in vector form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ and} \tag{1.2}$$

$$E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \tag{1.3}$$

where $\boldsymbol{y} = [y_1 \ y_2 \ \dots \ y_n]^\mathsf{T}$, and $\boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \dots \ \mu_n]^\mathsf{T}$. The matrix $\boldsymbol{X}$, which is often called the *design matrix*, is defined as

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\mathsf{T} \\ \boldsymbol{x}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{x}_n^\mathsf{T} \end{bmatrix},$$

and $\boldsymbol{X\beta}$ is the linear component of the classical linear regression model.

In this setting, the ordinary least squares (OLS) estimate and the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ are identical, and given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}. \tag{1.4}$$

Once the model parameters have been fitted to the data in $\boldsymbol{X}$, the linear combinations of $\boldsymbol{\beta}$ and the rows of the design matrix comprise the estimated linear predictors:

$$\hat{\eta}_i = \boldsymbol{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}. \tag{1.5}$$

Due to the model assumptions, the estimated linear predictors are suitable estimators for $E(y_i) = E(y_i|x_{i1},\ldots,x_{ik}) = E(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$. Hence $\hat{\eta}_i$ is used to predict y_i, i.e. $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The classical linear regression is applied in prediction, forecasting, and when quantifying the strength of linear relationships.

1.2 Generalised Linear Models (GLMs)

Classical linear regression models can be generalised in order to permit response variables whose errors are not normally distributed. GLMs are extensions of linear models which facilitate modelling non-normal response variables. A GLM consists of three components:

1) The random component,

2) the systematic component, and

3) the link function.

The response variables $Y_1,\ldots,Y_n$ are the *random component*. A standard assumption is that the response variables are random and independent, but not identically distributed. They each have a distribution in canonical form from the same exponential family. In some cases, the observations of $Y_1,\dots,Y_n$ are binary, taking on values such as "0" and "1", or "success" and "failure".

The *systematic component* is the function of the covariates $x_1,\dots,x_k$ which is related to the expected value of $Y_1,\dots,Y_n$. Just as in classical linear regression, the function is called the *linear predictor* and takes the usual form:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}, \tag{1.6}$$

where, as before, $\boldsymbol{x}_i^{\mathsf{T}} = [1 \;\; x_{i1} \;\; x_{i2} \;\; \dots \;\; x_{ik}]$ and $\boldsymbol{\beta} = [\beta_0 \;\; \beta_1 \;\; \dots \;\; \beta_k]^{\mathsf{T}}$.

Finally, the *link function* connects the random component and the systematic component by stating that

$$g(\mu_i) = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}, \tag{1.7}$$

where $\mu_i = E(Y_i)$ and $g(\cdot)$ is the link function. In a GLM framework, g is a differentiable and monotone function; i.e. its first derivative does not change sign (Dobson, 2008). The inverse link function, $g^{-1}(\cdot)$, also called the *mean function*, is such that $g^{-1}(\eta_i) = \mu_i$. In models where the relationship $\mu_i = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}$ is assumed, i.e. $g(\mu_i) = \mu_i$, the link function is called the *identity link*.

However, μ_i is often non-linearly related to the linear predictor. Several of the distributions a response variables may have, impose restrictions on the mean. For example, when μ_i cannot be negative, which is the case with count data, the link function $g(\mu_i) = \log(\mu_i)$ may be suitable. This link function is called the *log link*.

In some cases, such as when the response variables are Bernoulli distributed, μ_i must be restricted to the interval $[0,1]$. A common procedure is to choose a probability density function, referred to as the *tolerance distribution*, and subsequently use the corresponding cumulative distribution function (CDF) to model the mean. Thus, the link function is derived from the CDF.

If the standard normal distribution is the chosen tolerance distribution, for example, the mean would be modelled as follows:

$$\mu_i = \Phi(\eta_i) = g^{-1}(\eta_i), \tag{1.8}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. As a result, the link function is $g(\mu_i) = \Phi^{-1}(\eta_i)$, which is known as the *probit link*. GLMs where $Y_1, Y_2, \dots, Y_n$ are Bernoulli distributed, and link functions such as the probit link are appropriate, are covered in the following section.

1.3 Binary Regression Models

Binary response variables, also referred to as dichotomous responses, are commonplace in statistical analysis. This type of categorical response takes on the values 0 ("failure") or 1 ("success") to indicate the occurrence of a particular characteristic or event. Whether a tumour is malignant or benign, and whether a customer is loyal or chooses a competitor, are examples of responses one may wish to model.

The expected value of a binary variable Y (which is Bernoulli distributed) is given by

$$E(Y) = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = P(Y = 1).$$

Hence in the case of binary response variables, *the expectation is also a probability*. In order to underline this fact, we define π to be equal to the probability of success, i.e. $\pi = P(Y = 1) = E(Y)$, and use the following notation:

$$\pi_i = \pi(\eta_i) = P(Y_i = 1 \mid \boldsymbol{x}_i) = g^{-1}(\eta_i), \tag{1.9}$$

for GLMs where $Y_1, \dots, Y_n$ are dichotomous.

Observations with identical rows in the design matrix, can be grouped into N distinct subgroups called *covariate patterns*. If the data can be aggregated in this manner, we define the responses $Y_1, \dots, Y_N$ as the number of "successes" with probability π_j among n_j "trials" in covariate pattern j, i.e. $Y_j \sim Bin(n_j, \pi_j)$, where $j = 1, \dots, N$.

Often when including a continuous variable in in one's model (or when multiple covariates are multicategorical), the number of covariate patterns is equal to n. According to Hosmer (2013), this is the most common number of covariate patterns in practice when there is at least one continuous covariate in the model. This thesis will only consider aspects of binary regression related to ungrouped Bernoulli responses, i.e. only cases where $Y_i \sim Bin(n_i, \pi_i)$ where $n_i = 1$, $E(Y_i) = \pi_i$, $Var(Y_i) = \pi_i(1 - \pi_i)$, and $i = 1, \dots, n$.

1.3.1 Link Functions and Their Corresponding Models

A GLM with binary responses and identity link function is called a *linear probability model* (LPM) (Agresti, 2013). This model, where $\pi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, allows the probabilities π_i to take on any value on the real line. The LPM offers a simple interpretation of the covariate effects, β_j, $j = 1, \ldots, k$, but it is often inappropriate to limit the range of covariate values $\boldsymbol{x}_i$ so that $0 \leq \pi_i \leq 1$. Agresti (2013) also stated that the maximum likelihood (ML) estimation of multiple covariate effects could be adversely affected due to the non-constant variance of y.

Another disadvantage of the LPM is the assumption of a linear relationship between π_i and $\boldsymbol{x}_i^\top \boldsymbol{\beta}$. This assumption implies that a fixed change in $\boldsymbol{x}_i$ has the same effect on π_i regardless of its initial values, which is unrealistic and counter-intuitive in many settings. In many cases, the relationship between π_i and the linear predictor is better captured by a sigmoid (S-shaped) curve.

The aforementioned shortcomings of the LPM justify considering non-linear link functions when modelling π_i. The most prevalent link functions are summarized in Table 1.1. In principle, any link function g, where g^{-1} is monotonically increasing and maps $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ to $[0,1]$, can be used. As mentioned in the previous section, a common choice of g is a function such that g^{-1} equals a cumulative distribution function (CDF).

The *probit model* is a GLM which uses the previously mentioned probit link, where $g(\pi_i) = \Phi^{-1}(\eta_i)$. The symmetric S-shape of $\Phi(\eta_i)$ lends itself well to describing π_i in some situations. The probit model was originally used to describe binary responses in toxicology studies, more specifically dose response data resulting from bioassays (Dobson, 2008).

Bliss (1934) suggested transforming the success probability π_i into so-called "probits" using the inverse Normal CDF Φ^{-1}. This method facilitated linear regression despite the S-shaped relationship between the dosage level of a toxic agent and the proportion killed in a set of organisms exposed to said dosage (Bliss, 1934). The probit link has later been applied in a wide range of disciplines, such as social sciences and biological sciences (Dobson, 2008).

Table 1.1 Three common link functions for GLMs with binary responses. The rightmost column lists their respective mean functions.

Link	*Tolerance distribution*	$g(\pi_i) = \eta_i$	$\pi_i = g^{-1}(\eta_i)$
Logit	Logistic distribution	$\log\left(\frac{\pi_i}{1-\pi_1}\right)$	$\frac{e^{\eta_i}}{1+e^{\eta_i}}$
Probit	Normal distribution	$\Phi^{-1}(\pi_i)$	$\Phi(\eta_i)$
Complementary log-log	Extreme value distribution	$\log\left[-\log\left(1-\pi_i\right)\right]$	$1 - e^{-e^{\eta_i}}$

Another link function which is used when dealing with dose response data and other dichotomous responses is the *complementary log-log link.* Fahrmeir et al. (2013) state that the GLM using this link function, the *complementary log-log model*, is useful in more specific applications. The tolerance distribution used for modelling π_i is the extreme value distribution, resulting in the link $\log\left[-\log\left(1-\pi_i\right)\right]=\eta_i$.

The mean function of the complementary log-log model is asymmetric. In cases where the true functional form of π_i deviates considerably from a sigmoid which is symmetric about the point where $\pi_i = 0.5$, the complementary log-log link may be an appropriate choice. The asymmetric mean function is an important feature which distinguishes this model from the probit model and the very popular logit model.

The most popular link function for binary responses is the *logit link*:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i) = \eta_i, \tag{1.10}$$

which is based on the logistic distribution. This link is also referred to as the *log odds* transformation. A GLM with a binary random component and the logit link is called a *logistic regression model.* The s-shaped mean function, $\pi_i = {}^{e^{\eta_i}}/_{(1+e^{\eta_i})}$, is the well-known standard logistic function — a function which has several useful mathematical properties.

The standard logistic function is symmetric about $(0, \frac{1}{2})$, i.e. $\pi(\eta_i) = 1-\pi(-\eta_i)$, and its first derivative is $\pi'(\eta_i) = \pi(\eta_i)(1-\pi(\eta_i))$. These convenient features is a central reason for choosing the logistic model when considering GLMs whose link functions are derived from cumulative distribution functions (Hosmer, 2013).

Another leading reason to favour the logistic regression model is the interpretability of the covariate effects $\beta_0, \beta_1, \ldots, \beta_k$. However, when prediction of the response variable is regarded as more helpful than meaningful parameter estimates, Hosmer (2013) recommends considering the probit, log-log, or complementary log-log link functions in addition to $logit(\pi_i)$. These alternative GLMs may produce better estimates of the outcome (or success) probability, π_i, than the logistic regression model. The interpretability of logistic regression models are covered in the following subsection.

1.3.2 Interpretation of logistic regression models

Assuming that the linear predictor we are dealing is of the usual form $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$, a covariate effect β_j, $j = 1, \ldots, k$, is the change in the log-odds of success caused by a one-unit increase in the covariate x_{ij}. This is evident when looking at the difference between

$logit(\pi_i)$ evaluated at $x_{ij}+1$ and $logit(\pi_i)$ evaluated at x_{ij}.

Consider, for simplicity, the following model with $k=2$ covariates:

$$\log\left(\frac{\pi\left(\beta_0+\beta_1 x_{i1}+\beta_k\left(x_{i2}+1\right)\right)}{1-\pi\left(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2}\right)}\right)-\log\left(\frac{\pi\left(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2}\right)}{1-\pi\left(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2}\right)}\right)=$$

$$\beta_0+\beta_1 x_{i1}+\beta_2(x_{i2}+1)-\beta_0-\beta_1 x_{i1}-\beta_2 x_{i2} = \beta_2.$$

Hence the former equality may be rewritten:

$$\log\left(\frac{\pi(\beta_0+\beta_1 x_{i1}+\beta_k(x_{i2}+1))/[1-\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})]}{\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})/[1-\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})]}\right)=\beta_2,$$

and exponentiating both sides results in the following:

$$\frac{\pi(\beta_0+\beta_1 x_{i1}+\beta_k(x_{i2}+1))/[1-\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})]}{\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})/[1-\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})]}=e^{\beta_2} \implies$$

$$\frac{\pi(\beta_0+\beta_1 x_{i1}+\beta_k(x_{i2}+1))}{1-\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})}=e^{\beta_2}\frac{\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})}{1-\pi(\beta_0+\beta_1 x_{i1}+\beta_k x_{i2})}.$$

This means that a one-unit increase of the covariate x_{ij} produces a multiplicative change of the odds of success, where e^{β_j} is the multiplicative change factor. If β_j is positive, a one-unit increase in x_{ij} causes the odds of success to increase; if β_j is negative, the one-unit increase causes the odds of success to decrease. In the case where there is no relationship between x_{ij} and π_i, β_j equals zero and the odds of success remains unaffected by increasing x_{ij} to $x_{ij}+1$. Hosmer (2013) has an entire chapter devoted to the interpretation of fitted logistic models.

1.3.3 Maximum Likelihood Estimation

This section gives a very brief mention of the method of estimation which provides a basis for a large proportion of binary data analysis methods and processes – maximum likelihood (ML). It is a large subject area in its own right, and its application to binary regression models is covered in detail by Agresti (2013) and Hosmer (2013). The maximum likelihood estimates (MLEs) are the values of the parameters of a statistical model which maximize the likelihood, or log-likelihood, function of the model.

In this thesis, where the focus is on binary regression models, we have that $y_i \sim$

Bernoulli(π_i) and $f(y_i) = \pi_t^{y_i}(1-\pi_i)^{1-y_i}$. Hence the joint likelihood function of $y_1, y_2, \ldots, y_n$ is defined as

$$L(\boldsymbol{\beta};\boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_t^{y_i}(1-\pi_i)^{1-y_i}, \tag{1.11}$$

and the log-likelihood function is

$$\ell(\boldsymbol{\beta};\boldsymbol{y}) = \log\left(L(\boldsymbol{\beta};\boldsymbol{y})\right) = \sum_{i=1}^{n}\left[y_i \log \pi_i + (1-y_i)\log(1-\pi_1)\right]. \tag{1.12}$$

The MLEs of $\boldsymbol{\beta}$ are the values

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\beta}_0 & \hat{\beta}_1 & \ldots & \hat{\beta}_k\end{bmatrix}^{\mathsf{T}}.$$

which maximize (1.11) and (1.12).

When fitting a logistic regression model, it is possible to evaluate the linear predictors $\hat{\eta}_i$, once $\hat{\boldsymbol{\beta}}$ is estimated. These $\hat{\eta}_i$ are also called *sample logits*. Finally, we get the estimates of the probabilities $\hat{\pi}_i$ by evaluating ${e^{\hat{\eta}_i}}/{(1+e^{\hat{\eta}_i})}$ at the sample logits $\hat{\eta}_i = \boldsymbol{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}$.

Chapter 2

Goodness-of-Fit Tests and Their Statistics

In this chapter, we will present the goodness-of-fit tests compared in our simulation studies. Supplementary information on their implementation in `R` is presented in Chapter 3.

2.1 The Standardized Pearson Test

The classic Pearson chi-squared statistic is frequently used when a GLM has has less than n covariate patterns, and is defined as

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{2.1}$$

in cases where the data is grouped into N covariate patterns and n_i is the number of observations in each of those subgroups (Hosmer, 2013). This statistic is based on the difference between the observed response variables and the fitted probabilities of the model in question.

This thesis addresses the case where $n_i = 1$ and $i = 1, \ldots, N = n$, which is a common occurrence when at least one covariate is continuous. Hence for the remainder of this text, the classic Pearson chi-squared statistic equals

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}\ . \tag{2.2}$$

When performing a classic Pearson chi-squared test, the main assumption is that the statistic X^2 is approximately chi-squared distributed with $n-p-1$ degrees of freedom when the null hypothesis (H_0) that the model that produced the fitted probabilities $\hat{\pi}_i$ is correctly specified. As noted by Dobson (2008) and Hosmer (2013), however, this assumption does not hold when $n_i = 1$. Thus, using this test on ungrouped data will produce incorrect p-values.

A set of approximations of the asymptotic expectation and variance of the classic Pearson chi-square statistic was introduced by McCullagh (1985). These moments are conditional on the estimated parameters $\hat{\boldsymbol{\beta}}$ and their estimates are quite complicated to compute (Hosmer et al., 1997). A few years later, Osius and Rojek (1992) showed that in the special case of binary data, the conditional and unconditional moments of X^2 are asymptotically equivalent, and presented a much more painless way of computing the a large sample approximations of the moments.

Osius and Rojek (1992) stated that when H_0 holds, X^2 has an asymptotic normal distribution, and can be standardised such that it approximates the standard Normal distribution. The estimation of the standardising moments, the expectation and variance of X^2, is described in detail in Hosmer (2013) and Hosmer et al. (1997).

In Hosmer (2013), the estimated mean equals $n-k-1$, where k is the number of covariates and $k+1$ is the number of parameters. In the Appendix of Hosmer et al. (1997), on the other hand, the estimator equals n. We will use the Osius and Rojek estimation method described by Hosmer (2013), which is a more recent publication. In this method, the estimator of the variance of X^2 is the residual sum-of-squares, denoted RSS_P, resulting from the regression of the artificial response $c_i = {}^{(1-2\hat{\pi}_i)}/_{(\hat{\pi}_i(1-\hat{\pi}_i))}$ on the design matrix $\boldsymbol{X}$ with weights $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$. Recall that $y_i \sim Bernoulli(\pi_i)$ and that $Var(y_i) = \pi_i(1-\pi_i)$. Hence the maximum likelihood estimate of the variance of y_i is very a influential component of the standardised Pearson test.

Finally, when the estimates of the standardising moments have been computed, the standardised Pearson statistic can be evaluated:

$$X_{st}^2 = \frac{X^2 - (n-k-1)}{\sqrt{RSS_P}}, \tag{2.3}$$

which simply is a standardised version of the classic Pearson chi-squared statistic X^2. When H_0 is true, X_{st}^2 is approximately $N(0,1)$. It is recommended to obtain the p-value using a two-tailed test (Osius and Rojek, 1992).

It is worth noting that for small samples, Hosmer et al. (1997) advises using expressions involving the estimated moments to firstly, scale X^2, and secondly, calculating a constant denoted τ. Subsequently, the p-value is computed using the chi-square distribution with τ degrees of freedom. This approach was not chosen due to the prevalence of Osius and Rojek's two-tailed z-test in many different R-packages.

2.2 Unweighted Sum of Squares Test

The unweighted sum-of-squares (USS) statistic,

$$S = \sum_{i=1}^{n} (y_i - n_i\hat{\pi}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2 , \tag{2.4}$$

was proposed by Copas (1989). Befitting its name, the unweighted sum-of-squares statistic, does not have a denominator which influences its value as seen in (Section 2.1). The statistic was later used to test the overall model adequacy of logistic regression models and compared to other goodness-of-fit statistics by Hosmer et al. (1997).

In this article, the unweighted sum-of-squares test was performed by standardizing S in a similar manner as in Section 2.1, and subsequently computing the p-value using the standard normal distribution. Both Hosmer et al. (1997) and Hosmer (2013) state that under H_0,

$$\hat{S}_{st} = \frac{S - \hat{\mu}_S}{\hat{\sigma}_S} \xrightarrow{d} N(0,1), \tag{2.5}$$

where $\hat{\mu}_S$ and $\hat{\sigma}_S^2$ are respectively the estimates of the asymptotic expectation and variance of the USS statistic. We will refer to $\hat{S}_{st}$ as the *standardised USS statistic*. Similarly to X_{st}^2, the p-value of $\hat{S}_{st}$ is obtained using a two-tailed z-test (Hosmer, 2013).

The estimator of the asymptotic moment μ_S used when computing $\hat{S}_{st}$ is defined as

$$\hat{\mu}_{\hat{S}} = \sum_{i=1}^{n} \hat{\pi}_i (1 - \hat{\pi}_i), \tag{2.6}$$

according to both aforementioned publications. The larger the estimated variance, the smaller the the numerator in (2.6). The approach used for estimating the asymptotic variance of the USS statistic is similar to the one described for the asymptotic variance of X_{st}^2 in the previous section.

The estimation of σ_S^2 was done by regressing $d_i = 1 - 2\hat{\pi}_i$ on the design matrix $\boldsymbol{X}$ with weights $v_i = \hat{\pi}_i (1 - \hat{\pi}_i),\ i = 1, \ldots, n$ (Hosmer, 2013). The residual sum-of-squares from that artificial regression, denoted $\sqrt{RSS_U}$, is the estimate of $\hat{\sigma}_S^2$.

2.3 Stukel's Test

A class of models in which asymmetry of the mean function's curve, or probability curve, $\pi(\eta_i)$, is allowed was introduced by Stukel (1988). This class provides an alternative to the standard logistic regression model, where $\pi_i = \pi(\eta_i) = {}^{e^{\eta_i}}/_{(1+e^{\eta_i})}$ and the probability curve has the symmetry property $1 - \pi(\eta_i) = \pi(-\eta_i)$ about $\eta_i = 0$.

The standard logistic model, where the probability π_i is modelled using the logistic function, comes with several restrictions. Its mean function, $\pi(\eta_i)$, has the aforementioned symmetry property, and first derivative $\pi'(\eta_i) = \pi(\eta_i)(1 - \pi(\eta_i))$. These restrictions make the model less suitable for certain types of data whose true probability curves do not have the same functional form as the logistic function. This may be because the probability curve in question is asymmetric, has a different first derivative (i.e. steepness), and/or has a different tolerance distribution than the logistic distribution. However, even when this is the case, the standard logistic model can serve as a framework for encompassing a wider variety of data, which is what was done by Stukel (1988).

Stukel (1988) introduced a generalised model which permits a more extensive range of shapes of probability curves. The standard logistic model was generalized by adding two additional parameters, φ_1 and φ_2, and proposing a new general model form. The general form of Stukel's model is

$$\pi_{\boldsymbol{\varphi}}(\eta_i) = \frac{e^{h_{\boldsymbol{\varphi}}(\eta_i)}}{1 + e^{h_{\boldsymbol{\varphi}}(\eta_i)}}, \tag{2.7}$$

or, equivalently,

$$logit(\pi_i) = h_{\boldsymbol{\varphi}}(\eta_i), \tag{2.8}$$

where $h_{\boldsymbol{\varphi}}$ are strictly increasing functions defined as follows:

For $\eta_i \geq 0 \Leftrightarrow \pi_i \geq \frac{1}{2}$:

$$h_{\boldsymbol{\varphi}}(\eta_i) = \begin{cases} \frac{1}{\varphi_1}\left(e^{\varphi_1|\eta_i|} - 1\right), & \varphi_1 > 0 \\ \eta_i, & \varphi_1 = 0 \\ -\frac{1}{\varphi_1} log(1 - \varphi_1|\eta_i|), & \varphi_1 < 0, \end{cases} \tag{2.9}$$

and for $\eta_i \leq 0 \Leftrightarrow \pi_i \leq \frac{1}{2}$:

$$h_{\boldsymbol{\varphi}}(\eta_i) = \begin{cases} -\frac{1}{\varphi_2}\left(e^{\varphi_2|\eta_i|}-1\right), & \varphi_2 > 0 \\ \eta_i, & \varphi_2 = 0 \\ \frac{1}{\varphi_2} log\left(1-\varphi_2|\eta_i|\right), & \varphi_2 < 0. \end{cases} \tag{2.10}$$

In this framework, the standard logistic model is a special case of Stukel's generalized model, occurring when $\varphi_1 = \varphi_2 = 0$.

Since φ_1 and φ_2 regulate the presence of asymmetry and how heavy the tails are in the probability curve $\pi_{\boldsymbol{\varphi}}(\eta_i)$, it follows that they are shape parameters. When $\varphi_1 \neq \varphi_2$, the curve is asymmetric, whereas when $\varphi_1 = \varphi_2$, it is symmetric. The upper tail is controlled by φ_1, and the lower tail is controlled by φ_2. When examining (2.9) and (2.10), one can see that:

1) when $\varphi_1 > 0$, (2.9) is exponential (with a relatively large positive $\frac{d}{d\eta_i}h_{\boldsymbol{\varphi}}$),

2) when $\varphi_1 < 0$, (2.9) is logarithmic (with a relatively small positive $\frac{d}{d\eta_i}h_{\boldsymbol{\varphi}}$),

3) when $\varphi_2 > 0$, (2.10) is exponential (with a relatively large positive $\frac{d}{d\eta_i}h_{\boldsymbol{\varphi}}$), and

4) when $\varphi_2 < 0$, (2.10) is logarithmic (with a relatively large positive $\frac{d}{d\eta_i}h_{\boldsymbol{\varphi}}$).

It follows that when a shape parameter is positive, it causes the h function controlled by the parameter to increase much more rapidly. This makes its respective tail shorter, i.e. steeper, than compared to the standard logistic model. Conversely, when the shape parameter is negative, its respective tail is longer, i.e. less steep, than compared to the corresponding tail of the standard logistic model where $\boldsymbol{\varphi} = (0,0)$. The greater the $|\varphi_1|$, or $|\varphi_2|$, the more pronounced the effect on the heaviness of the tail.

Stukel (1988) supplied values of $\boldsymbol{\varphi}$ where the corresponding mean functions $\pi_{\boldsymbol{\varphi}}(\eta_i)$ approximate some well known tolerance distributions. Stukel's model approximates the probit model when $\boldsymbol{\varphi} \approx (0.165, 0.165)$. This means that $\pi_{\boldsymbol{\varphi}}(\eta_i)$ approximates the standard Normal CDF.

When $\boldsymbol{\varphi} \approx (0.62, -0.037)$, $\pi_{\boldsymbol{\varphi}}(\eta_i)$ is approximately the minimum extreme value distribution's CDF, which gives us the complementary log-log model. The values approximating the maximum extreme value distribution (the log-log model) and the standard Laplace distribution are also provided. Hence it is possible to test whether other link functions than the logit link are more appropriate when analysing data.

Stukel (1988) advised that the maximum likelihood estimates of the covariate effects $\boldsymbol{\beta}$ and the shape parameters $\boldsymbol{\varphi}$ should be computed using an Newton-Raphson-like procedure called the delta algorithm. This algorithm is described in detail by Jørgensen (1984). The variance of the estimated $\hat{\boldsymbol{\varphi}}$

Stukel (1988) stated that one could evaluate the fit of the standard logistic model by testing whether $\boldsymbol{\varphi} = (0,0)$ using a score test. The score test of the null hypothesis that φ_1 and φ_2 are equal to 0 (or other specific values) can be calculated using statistical software where the specified model is defined as

$$logit(\pi_i) = \eta_i + \varphi_1 z_{1,i} + \varphi_2 z_{2,i}\text{ , where} \tag{2.11}$$

$$z_{1,i} = \frac{1}{2}\hat{\eta}_i^2 I(\hat{\eta}_i \geq 0)\text{ , and} \tag{2.12}$$

$$z_{2,i} = -\frac{1}{2}\hat{\eta}_i^2 I(\hat{\eta}_i < 0), i = 1,\ldots,n. \tag{2.13}$$

Evaluating how well a specified standard logistic model fits the data can be done by using score tests. Stukel (1988) provides equations for the the score vector, the asymptotic mean and variance-covariance matrix, and their asymptotic chi-squared distribution under $H_0 : \boldsymbol{\varphi} = (0,0)$. The score statistic, evaluated at $\boldsymbol{\varphi} = (0,0)$ and the maximum likelihood estimates fitted under the standard logistic regression, has an asymptotic $\chi^2(2)$ distribution.

In this thesis, the aforementioned score test is referred to as *Stukel's score test*. Stukel (1988) also recommended performing a likelihood ratio test (LRT). One may use a likelihood ratio test (LRT) to compare the nested models $logit(\pi_i) = \eta_i$ and $logit(\pi_i) = \eta_i + \varphi_1 z_{1,i} + \varphi_2 z_{2,i}$ (Hosmer, 2013). In the following chapters, this LRT is referred to as *Stukel's LRT*. The computation of the p-values, and the introduction of a modified version of Stukel's LRT, is covered in Chapter 3.

For large sample sizes, score tests are asymptotically equivalent to likelihood ratio tests (LRTs) i terms of distribution when H_0 is true (Yan, 2009). It is therefore possible that Stukel's score test and LRT perform similarly for very large n.

2.4 The Information Matrix Test

The information matrix test (IMT) was proposed by White (1982) as a test for model misspecification when applying maximum likelihood estimation techniques. It is based on a theorem stating that the Hessian form and the outer product form of the information matrix (denoted respectively by $-A(\boldsymbol{\beta})$ and $B(\boldsymbol{\beta})$) are equivalent when the model is correctly specified. Specifically, for element (i,j) in these matrices, we have that

$$\{-A(\boldsymbol{\beta})\}_{i,j} := -E\left\{\frac{\partial^2 \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_i\partial\beta_j}\right\} = E\left\{\frac{\partial \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_i}\cdot\frac{\partial \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_j}\right) =: \{B(\boldsymbol{\beta})\}_{i,j} \tag{2.14}$$

where $\ell(\boldsymbol{\beta};Y_t) = \log f(Y_t;\boldsymbol{\beta})$, $i,j = 1,2,\ldots,p$, and the expectations are taken with respect to the true probability density (or mass) function, f. The model is misspecified if this equality fails to hold, i.e. when $A(\boldsymbol{\beta}) + B(\boldsymbol{\beta})$ does not equal the $p \times p$ null matrix $\mathbf{0}_{p\times p}$ (White, 1982). The main focus of this thesis is when $Y_t \sim Bernoulli(\pi_t)$ and the logit link is used. Hence $f(y_i) = \pi_t^{y_t}(1-\pi_i)^{1-y_t}$ in our case.

White (1982) specified the following matrices

$$\{A_n(\boldsymbol{Y};\boldsymbol{\beta})\}_{i,j} = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_i\partial\beta_j}, \tag{2.15}$$

$$\{B_n(\boldsymbol{Y};\boldsymbol{\beta})\}_{i,j} = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_i}\cdot\frac{\partial \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_j},\ i,j = 1,2,\ldots,p, \tag{2.16}$$

and used $A_n(\boldsymbol{y};\hat{\boldsymbol{\beta}}) + B_n(\boldsymbol{y};\hat{\boldsymbol{\beta}})$ as a gauge of model misspecification ($\boldsymbol{y}$ is the $n \times 1$ vector of observations of $\boldsymbol{Y}$). A test statistic for the IMT was obtained by looking at the asymptotic distribution of the elements of $\sqrt{n}(A_n(\boldsymbol{y};\hat{\boldsymbol{\beta}}) + B_n(\boldsymbol{y};\hat{\boldsymbol{\beta}}))$.

Due to the fact that $A_n(\boldsymbol{Y};\boldsymbol{\beta}) + B_n(\boldsymbol{Y};\boldsymbol{\beta})$ is symmetric, at least $p^2 - p(p+1)/2$ of its elements are superfluous and unnecessary to consider. The $q \leq p(p+1)/2$ non-redundant elements, referred to as "indicators of interest", are placed in a $q \times 1$ vector denoted by $D_n(\boldsymbol{Y};\boldsymbol{\beta})$. This vector of indicators is defined as

$$D_n(\boldsymbol{Y};\boldsymbol{\beta}) = \frac{1}{n}\sum_{t=1}^{n} d(Y_t,\boldsymbol{\beta}), \tag{2.17}$$

where $d(Y_t,\boldsymbol{\beta})$ is a $q \times 1$ vector with typical element

$$d_r(Y_t,\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_i\partial\beta_j} + \frac{\partial \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_i}\cdot\frac{\partial \ell(\boldsymbol{\beta};Y_t)}{\partial\beta_j} \tag{2.18}$$

for rows $r = 1,\ldots,q$, and $i = 1,\ldots,p$, $j = i,\ldots,p$ (unless $q < p(p+1)/2$, in which case some subset of (i,j) is omitted).

If the model is specified correctly, then

$$\sqrt{n}\, D_n(\boldsymbol{Y};\hat{\boldsymbol{\beta}}) \overset{A}{\sim} MVN\big(\mathbf{0}_q, V(\boldsymbol{\beta})\big), \tag{2.19}$$

where $MVN\big(\mathbf{0}_q, V(\boldsymbol{\beta})\big)$ is the multivariate Normal distribution with mean vector $\mathbf{0}_q$ and asymptotic variance-covariance $V(\boldsymbol{\beta})$ (White, 1982). The mean vector is the $q \times 1$ null vector, i.e. it has q components, each of which is 0. The asymptotic covariance matrix is defined by

$$V(\boldsymbol{\beta}) = E\big\{w(Y_t;\boldsymbol{\beta})w(Y_t;\boldsymbol{\beta})^{\mathsf{T}}\big\}, \tag{2.20}$$

where $w(Y_t;\boldsymbol{\beta})$ is a $q \times 1$ vector defined by

$$w(Y_t;\boldsymbol{\beta}) = d(Y_t,\boldsymbol{\beta}) \;-\; \nabla D(\boldsymbol{\beta})A(\boldsymbol{\beta})^{-1}\nabla\ell(\boldsymbol{\beta};Y_t)^{\mathsf{T}} \tag{2.21}$$

and

$$\nabla D(\boldsymbol{\beta}) = E\left\{\frac{\partial d(Y_t,\boldsymbol{\beta})}{\partial \beta_i}\right\}, \tag{2.22}$$

$$\tag{2.23}$$

$$\nabla\ell(\boldsymbol{\beta};Y_t) = \left\{\frac{\partial \ell(\boldsymbol{\beta};Y_t)}{\partial \beta_i}\right\}, \tag{2.24}$$

are, respectively, the $q \times p$ and $1 \times p$ Jacobian matrices with $i = 1,\ldots,p$.

Given the assumptions listed in White (1982) and any consistent estimator for $V(\boldsymbol{\beta})$, denoted by $\hat{V}_n(\hat{\boldsymbol{\beta}})$, the information matrix test statistic

$$\mathscr{I}_n = nD_n(\boldsymbol{Y};\hat{\boldsymbol{\beta}})^{\mathsf{T}}\hat{V}_n(\hat{\boldsymbol{\beta}})^{-1}D_n(\boldsymbol{Y};\hat{\boldsymbol{\beta}}) \tag{2.25}$$

has an asymptotic $\chi^2(q)$ distribution when the model is correctly specified (under H_0). The null hypothesis that the model is correctly specified is rejected when one computes $\mathscr{I}_n$ and it exceeds the critical value of the $\chi^2(q)$ distribution for a given significance level.

Several covariance matrix estimators have been proposed. White (1982) suggested a consistent estimator involving the Jacobian matrix of $D_n(\boldsymbol{Y};\hat{\boldsymbol{\beta}})$, which involves third derivatives of the log-likelihood functions of the random variables Y_t. Dealing with analytical third derivatives can make White's test statistic inconvenient to compute, as noted by White (1982) and Orme (1990).

Orme (1988) used an asymptotically efficient maximum likelihood estimator of $V(\boldsymbol{\beta})$ (as recommended by Davidson and Mackinnon (1984)) and presented a calculation procedure for IMT statistics specific to binary data models. This is the procedure used to perform the IMT in this thesis. Here follows a condensed outline of Orme's calculation procedure for two IMT statistics under the logistic regression model.

The ML estimator is obtained by replacing $\boldsymbol{\beta}$ by the MLEs $\hat{\boldsymbol{\beta}}$ in the expression produced by calculating $V(\boldsymbol{\beta})$ under the null hypothesis. When this particular estimator is plugged in (2.25), the resulting IMT statistic is the explained sum-of-squares from a specific artificial linear regression with no intercept term. In the special case of the logistic regression model, $\hat{\boldsymbol{r}}$ is regressed on $\boldsymbol{W}^* = (\boldsymbol{X}^*, \boldsymbol{Z}^*)$, where $\hat{\boldsymbol{r}}$ is a $n \times 1$ vector with typical element

$$\hat{r}_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}, \; i = 1,2,\ldots,n, \tag{2.26}$$

$\boldsymbol{X}^*$ is a $n \times p$ matrix with rows

$$\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}\,\boldsymbol{x}_i^{\mathsf{T}}, \; i = 1,2,\ldots,n, \tag{2.27}$$

and $\boldsymbol{Z}^*$ is a $n \times p(p+1)/2$ matrix with rows

$$\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}\,(1-2\hat{\pi}_i)\boldsymbol{z}_i^{\mathsf{T}}, \; i = 1,2,\ldots,n, \tag{2.28}$$

where $\boldsymbol{z}_i = vech(\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}})$, is the half-vectorization of the symmetric matrix $\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}$.

The explained sum-of-squares from the above regression gives us the statistic we will refer to as the *IMT1 statistic*. In addition, an alternative statistic can be obtained by dividing $IMT1$ by $\hat{\boldsymbol{r}}^{\mathsf{T}}\hat{\boldsymbol{r}}/n$. This statistic is referred to as the *IMT2 statistic*. The $IMT1$ and $IMT2$ are asymptotically equivalent, and under H_0 their asymptotic distribution is $\chi^2\left(k(k+1)/2\right)$.

Chapter 3

Significance Level Study

Two simulation studies were carried out in an effort to better understand how the overall goodness-of-fit (GOF) tests perform in different scenarios. In statistics, the methods with well established properties are the ones we trust the most when modelling real life data. However, these properties are often not feasible (or indeed possible) to determine analytically, and analytical results may necessitate assumptions which are usually violated in practice. The following study was the first of two simulations studies to be carried out, and in an effort to estimate the empirical significance levels of the GOF statistics.

The sampling distribution of a GOF test statistic, for example, is needed in order to determine whether the significance level of the test is equal to the *nominal significance level*, denoted by α. Likewise when trying to answer what *power* the test has when testing the null hypothesis H_0 of model adequacy. We use simulations studies to approximate the true sampling distribution of the GOF test statistics under a variety of conditions, such as sample size and form of linear predictor. By doing this we gain a better understanding of how the GOF tests behave in terms of significance level and power in predetermined situations.

In this chapter, we examine test performance when the correct logistic model had been fitted, and later in Chapter 4 we will cover the tests' ability to recognise when the fitted logistic model is incorrect (i.e. the power of the global GOF tests). Both studies were structured in a similar manner to the simulations in Hosmer et al. (1997) and employed some of the same covariate distributions and models. Hosmer and Hjort (2002) also used a similar set-up. The configurations of covariate distributions and logistic models from Hosmer et al. (1997) produce a wide variety of distributions of π_i-s, hence there was no apparent reason to refrain from using them. Three additional set-ups were added in order to include situations where the true probabilities π_i were highly left skewed, moderately left skewed and moderately right skewed.

In all our simulations, the performance of the test were evaluated based on $R = 1000$ replications. Each situation, or model, was investigated at the $\alpha = 0.05$ significance level with these three different sample sizes: $n = 100$, $n = 500$ and $n = 1000$. Hosmer et al. (1997) used 500 replications so there was an initial expectation that our results would not be identical, even in the parts where the study design is the same. Samples of size $n = 1000$ were not considered by Hosmer et al. (1997), but were included in these two studies due to the prevalence of data sets where $n \geq 1000$ and since the added computational burden was minimal. All of the simulations and computation were implemented in R.

This chapter also presents an exploration of a possibly new method. First it is posited that the euclidean distances between the estimated logistic probabilities and the observed response variable may be modelled by the Weibull distribution.

Then the possibility of

Then this was incorporated into a possible

development of a method where

the early stages of

3.1 The Goodness-of-Fit Statistics and Their Implementation

The overall GOF tests used in both the significance level study and the power study were:

1) the standardised Pearson test,
2) the unweighted sum-of-squares (USS) test,
3) Stukel's score test,
4) Stukel's likelihood ratio test (LRT), and
5) the information matrix test (IMT).

Two different Stukel's LRT statistics were included in the study, and the two different IMT statistics mentioned in Section 2.4 also. The remaining three tests had one statistic each.

The standardised Pearson chi-square statistic, X^2_{st}, was obtained by performing artificial regression to estimate the Osius and Rojek large sample normal approximation as described by Hosmer (2013). This choice of this standardisation method is due to its availability and ease of computation, even though Hosmer et al. (1997) stated that using estimates of McCullagh's moments, and scaling the statistic using a chi-square distribution, lead to better small sample performance.

The `R` function used to compute X^2_{st} was adapted from a function provided on the website accompanying the textbook by Bilder and Loughin (2014). The estimated mean in the function was changed to n minus the number of parameters in the model. Similarly to X^2_{st}, the standardised USS statistic $\hat{S}_{st}$ was also computed using artificial regression as outlined by Hosmer (2013) (see Section 2.1). The `R` function producing the p-value of the USS test was written specifically for this study.

Stukel's score test statistics and p-values were computed by the `R` function `stukel()` from the `logisticDx` package. The function follows the score test procedure described by Stukel (1988). The estimated variance-covariance matrix produced was singular in some instances. This occurred in cases where almost all of the n fitted values $\hat{\pi}_i$ were either greater than or less than 0.5, and caused the computation of the statistic and its p-value to fail. This was also the case in many circumstances where all of the $\hat{\pi}_i$ were either greater than or less than 0.5 (or, equivalently, all the $\hat{\eta}_i$ were either positive or negative).

The study reports the results from some of the situations where replications failed to produce a p-value. If the percentage of failed replications in a particular situation was less than or equal to 25%, the result based on the successful replications was included and marked with an asterisk in its respective table. However, if more than 25% of the 1000 replications failed, the result were not included.

Stukel's LRT was implemented by using the `anova()` function to compare two nested models, $logit(\pi_i) = \eta_i$ and $logit(\pi_i) = \eta_i + \varphi_1 z_{1,i} + \varphi_2 z_{2,i}$, which were fitted using the `glm()` function in `R`. Hence errors caused by trying to invert singular matrices were avoided. The resulting statistic is referred to as *Stukel's LRT1*.

Examination of a few test simulations revealed that the number of observations such that $z_{1,i} \neq 0$ was very low in some situations (less than 5 out of 500 observations in some example cases). Similarly, there were very few observations such that $z_{2,i} \neq 0$ in other examples. In these cases, the `glm()` function often returns `NA` as the estimated coefficient of the variable with very few non-zero values. As a result, the subsequent LRT involves comparing the null (logit) model with the generalized Stukel model where the variable with the `NA` coefficient is excluded, i.e. only one of the shape parameters is included in the more complex Stukel generalization.

A modified version of the algorithm computing Stukel's LRT statistic, referred to as *Stukel's LRT2*, was introduced. This was motivated by the question of whether the estimation of the shape parameter could be adversely affected if the number of observations where the corresponding variable was non-zero, was barely high enough to avoid `Na` coefficients, but still relatively low. A constraint requiring a minimum percentage of non-zero observed values of $z_{1,i}$ and $z_{2,i}$ was introduced.

It was decided that if less than 10% of the $\hat{\eta}_i$ resulted in non-zero $z_{1,i}$ then z_1 would be excluded from the alternative model and `anova()` would compare the null model to $logit(\pi_i) = \eta_i + \varphi_2 z_{2,i}$. Similarly, if less than 10% of the $\hat{\eta}_i$ resulted in non-zero $z_{2,i}$ then z_2 was not included and the alternative model used was $logit(\pi_i) = \eta_i + \varphi_1 z_{1,i}$. The statistic produced by this alternative version will be referred to as Stukel's LRT 2 statistic. Different constraints on the number or relative percentage of non-zero $z_{1,i}$ and $z_{2,i}$ was of interest to investigate, but not feasible due to time constraints.

When only one additional variable was included in the alternative model, the `anova()` function produced a p-value using the $\chi^2(1)$ distribution. When both variables were added in the alternative model, the `anova()` used the $\chi^2(2)$ distribution.

As mentioned in Section 2.4, there are two asymptotically equivalent versions of the IMT statistic available. These were both included in the simulation studies. At the time leading up to the simulations, the IMT was not found in any readily available `R`-packages. The `R` function computing the $IMT1$ and $IMT2$ statistics were therefore developed in accordance with the estimation procedure for logit models presented in Orme (1988) specifically for this study.

3.2 A Weibull-based behaviour indicator

An attempt was made to lay the foundation for a method with which one could predict a GOF test's performance in terms of rejection region, or significance level, possibly with an accompanying visual indication. To be able to assess your fitted model with a tool that provided a visualisation of discrepancy between the observed values of y_i and the sample logits $\hat{\eta}_i$, and additionally provided an indication of how specific GOF tests will behave in this setting, could be useful.

The two-parameter Weibull distribution was used as a model for the euclidean distance between y_i and $\hat{\pi}_i$, $i = 1, \ldots, n$. The significance level study simulations served as an explorative vehicle to study the behaviour of the Weibull distributions fitted to the aforementioned euclidean distances. There was an anticipation that certain patterns of fitted parameters could potentially offer a new procedure for gauging a GOF test statistic's performance.

For each observation $i = 1, \ldots, n$, let d_i denote the euclidean distance between the observed response variables and the fitted logistic probabilities, given by

$$d_i = \sqrt{(y_i - \hat{\pi}_i)^2} = \left| y_i - \hat{\pi}_i)^2 \right| . \tag{3.1}$$

During early stages of the study, simulated examples of such d_i where plotted as histograms and found to be similar in shape to a Weibull probability density function (PDF). An example of this is shown in Figure 3.1.

There are several ways one can check whether the Weibull distribution is a reasonable distribution for the distances d_i. One way of assessing whether a distribution is appropriate is by inspecting a Weibull probability plot as described in Devore and Berk (2012). Figure 3.2 contains an example of such a plot. This assessment of the plausibility of the Weibull distribution is not rigorous, and was not intended to be so due to the exploratory nature of this part of the thesis.

We hypothesized that $d_i \sim \mathit{Weibull}(a, b)$, where $a > 0$ and $b > 0$ are the shape and scale parameter, respectively. Different values of these two parameters can be combined to produce a variety of different distributional shapes (Devore and Berk, 2012). Because of this versatility, the Weibull distribution may be a viable alternative for modelling d_i. A more comprehensive description of the Weibull distribution can be found in Lai (2013).

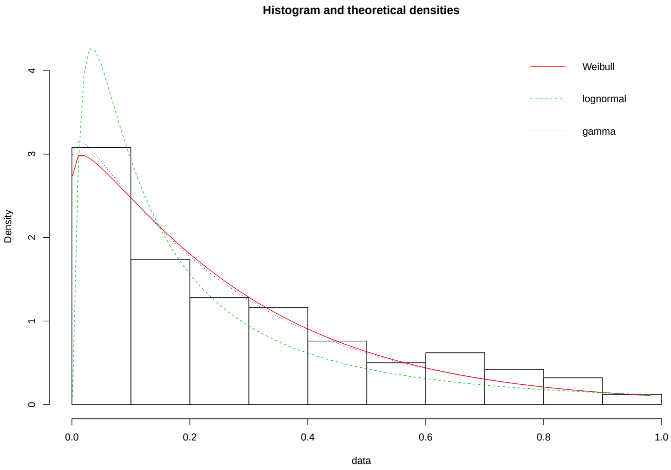


Fig. 3.1 A histogram of d_i based on a simulated example of a fitted logistic regression model, accompanied the PDFs of the Weibull, lognormal, and gamma distributions fitted to d_i, $i = 1, \ldots, n$.

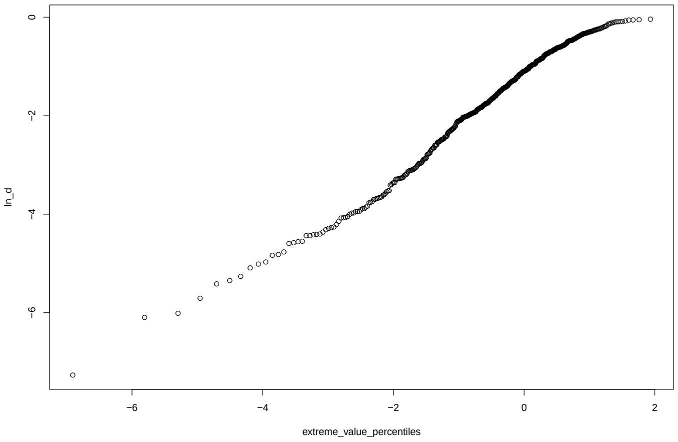


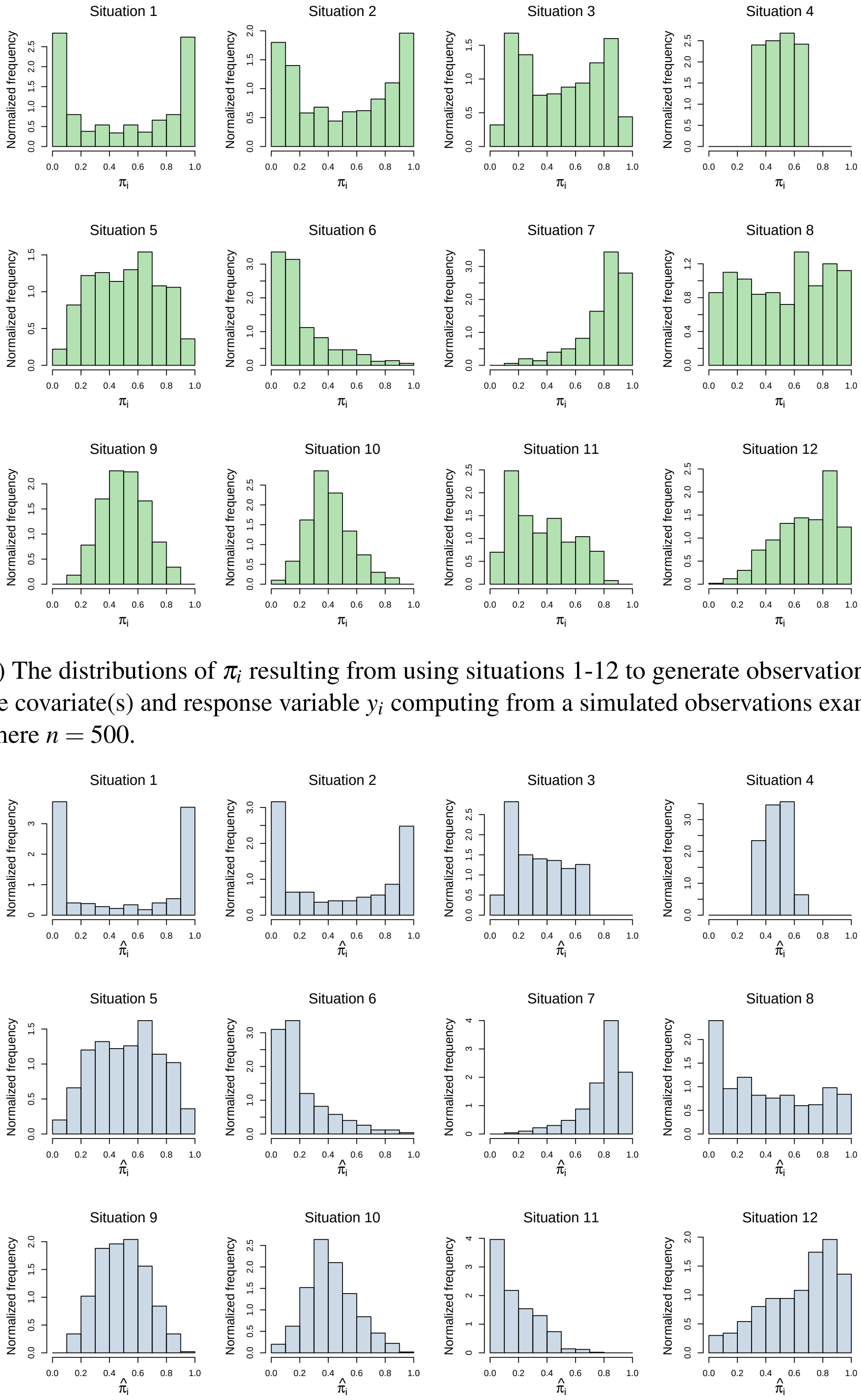
Fig. 3.2 Weibull probability plot

Therefore, in addition to the GOF statistics mentioned in section 3.1, estimates of the two parameters of the Weibull distribution were also computed during the simulations. The manner in which they were implemented and studied is described in sections 3.5.

3.3 Significance Level Study

It is useful to know how the tests behave when the fitted logit model is correct. In this chapter, the null hypothesis H_0 refers to the statement that a the logit model one has specified is correct. Ideally, a particular test should reject the true null hypothesis $100\times\alpha\%$ of the time. In other words, a tests' probability of making a *type I error* is supposed to equal the significance level α. If this is the case for many different situations, one can assume that the test statistic is typically neither too conservative nor too anti-conservative (i.e. having a unjustifiably large rejection region). Therefore, it is promising if the GOF statistics' empirical significance levels, i.e. the mean rejection rate of the true H_0, were to be approximately equal to α. The empirical significance level of a GOF test, denoted by $\hat{\alpha}$, was approximated by the rate of replications where the GOF test statistic incorrectly rejected H_0.

As previously stated, some of the situations (the choice of covariates, their distributions and coefficient values) coincide with the ones used by Hosmer et al. (1997) to compose the null hypotheses. One difference, however, is that this study added situations which produce negatively skewed distributions of the true probabilities π_i. Hosmer et al. (1997) had 9 different situations whereas this study included 12. They are indexed by $s = 1, 2, \ldots, 12$ and summarised in Table 3.1.



(a) The distributions of π_i resulting from using situations 1-12 to generate observations of the covariate(s) and response variable y_i computing from a simulated observations example where $n = 500$.

(b) The distributions of $\hat{\boldsymbol{\pi}}_j$ $(j = 1, \ldots, 5)$ from simulated example data sets where $n = 500$.

Fig. 3.3 The respective distributions of π_i and $\hat{\pi}_i$, $i = 1, \ldots, n$, produced by the logistic models in situations 1-12 with simulated example data sets where $n = 500$.

The true models corresponding to the 12 situations are all logistic regression models. In situations 1-7, there is one single covariate, while the remaining situations have either two or three covariates. Table 3.1 lists the distribution of the covariate(s) and the true model coefficients for each situation s. The table also includes a summary of the true logistic probabilities $\boldsymbol{\pi}_s$, where

$$\boldsymbol{\pi}_s = \left[\pi_{s1} \;\; \pi_{s2} \;\; \ldots \;\; \pi_{sn}\right]^{\mathsf{T}}, \tag{3.2}$$

are computed using the mean function $\pi_{si} = \pi(\eta_{si}) = {}^{e^{\eta_{si}}}\!/\!{}_{(1+e^{\eta_{si}})}$ and $n = 500$ generated observations of the covariate(s) with the covariate distribution(s) listed in Table 3.1. The summary is comprised of the smallest value of π_{si}, the first, second and third quartiles, and the largest value of π_i, denoted $\pi_{(1)}$, Q_1, Q_2, Q_3 and $\pi_{(n)}$, respectively. The histograms in Figure 3.3a show the distributions of these 12 sets of π_{si}.

In this significance level study, $R = 1000$ sets of $\boldsymbol{\pi}_s$ are computed directly from R generated data set of sample size n for every situation s, in the same way as described above. These $\boldsymbol{\pi}_s$ are used as a parameter of the `rbinom()` function to generate corresponding sets of response variables $\boldsymbol{y}_s = \left[y_{s1} \;\; y_{s2} \;\; \cdots \;\; y_{sn}\right]^{\mathsf{T}}$. Correctly specified models are subsequently fitted to their respective data sets consisting of the simulated observations of their covariate(s) and the corresponding $\boldsymbol{y}_s$. The GOF test statistics are applied to the resulting fitted $\hat{\boldsymbol{\pi}}_s = \left[\hat{\pi}_{s1} \;\; \hat{\pi}_{s2} \;\; \ldots \;\; \hat{\pi}_{sn}\right]^{\mathsf{T}}$, and the observed response variables $\boldsymbol{y}_s$.

Figure 3.3b presents histograms which show the distributions of the estimated probabilities $\hat{\pi}_{si}$, for each of the 12 situations included in this simulation study, for the same

Table 3.1 Situations used to examine the test statistics' rate of true null hypothesis rejection

Situation	**Covariate distribution**	**Logistic coefficients**	**Distributional characteristics of the logistic probabilities** π_i $(n = 500)$				
			$\pi_{(1)}$	Q_1	Q_2	Q_3	$\pi_{(n)}$
1	$U(-6,6)$	$\beta_0 = 0,\ \beta_1 = 0.8$	0.008	0.085	0.516	0.913	0.992
2	$U(-4.5,4.5)$	$\beta_0 = 0,\ \beta_1 = 0.8$	0.027	0.144	0.512	0.853	0.973
3	$U(-3,3)$	$\beta_0 = 0,\ \beta_1 = 0.8$	0.083	0.234	0.508	0.764	0.916
4	$U(-1,1)$	$\beta_0 = 0,\ \beta_1 = 0.8$	0.310	0.402	0.503	0.597	0.690
5	$\mathscr{N}(0,1.5)$	$\beta_0 = 0,\ \beta_1 = 0.8$	0.034	0.322	0.523	0.699	0.970
6	$\chi^2(4)$	$\beta_0 = -3.2,\ \beta_1 = 0.42$	0.040	0.147	0.372	0.756	1.000
7	30×Beta$(18,2)$	$\beta_0 = -12,\ \beta_1 = 0.5$	0.119	0.720	0.849	0.905	0.950
8	3 Independent $U(-6,6)$	$\beta_0 = 0,\ \beta_1 = \beta_2 = \beta_3 = 0.8/3$	0.028	0.261	0.554	0.773	0.974
9	3 Independent $\mathscr{N}(0,1.5)$	$\beta_0 = 0,\ \beta_1 = \beta_2 = \beta_3 = 0.8/3$	0.136	0.393	0.503	0.620	0.884
10	Independent $\chi^2(4)$ and 30×Beta$(18,2)$	$\beta_0 = -8,\ \beta_1 = 0.42/2,\ \beta_2 = 0.5/2$	0.061	0.313	0.395	0.502	0.881
11	Independent $U(-6,6)$, $\mathscr{N}(0,1.5)$ and $\chi^2(4)$	$\beta_0 = -1.3,\ \beta_1 = \beta_2 = 0.8/3,\ \beta_3 = 0.42/3$	0.037	0.175	0.325	0.528	0.845
12	Independent $U(-6,6)$, $\mathscr{N}(0,1.5)$ and 30×Beta$(18,2)$	$\beta_0 = 0,\ \beta_1 = \beta_2 = 0.8/3,\ \beta_3 = 0.19$	0.047	0.529	0.707	0.854	0.974

$n = 500$ data set used to produce the true probabilities in Figure 3.3a. These histograms give a general impression of how the fitted probabilities are distributed for the settings in the 12 different situations. There are slight variations in these distribution for different samples of the covariate(s) and for the different sample sizes, but Figure 3.3 still provides a very useful overview.

3.4 The results of the significance level study

The percentage of times each of the seven statistics rejected the true null hypotheses are listed in Table 3.2. Row s contains entries equal to $\hat{\alpha}_s \times 100$, where $\hat{\alpha}_s$ is the computed empirical significance level of a particular test statistic in situation s.

The six entries marked with an asterisk are cases where some of the computations of Stukel's score statistic failed due to singularity of the estimated variance-covariance matrix. In situations 4 and 7, the percentage of failed replications was 4.3% and 2.5%, respectively. In the remaining four cases, only 0.3% or less of the replications were unsuccessful in producing a p-value. The six resulting $\hat{\alpha}_s$ are based on only marginally smaller sets of simulated observations than the remaining entries in Table 3.2, so they are still useful. Nevertheless, this difference should be taken into account when comparing Stukel's score test to the other tests.

The standardised Pearson chi-square statistic, X^2_{st}, and the standardised USS statistic, $\hat{S}_{st}$, performed similarly in settings where the logistic probabilities were approximately symmetrically distributed with mostly small and large π_i's (in situations 1 and 2 where the covariates have the $U(-6,6)$ and $U(-4.5,4.5)$ distribution, respectively). In cases where the situations produced π_i which were mostly clustered around 0.5, however, X^2_{st} had very high rejection rates. A less than ideal performance was expected for small sample sizes, but in situation 4, over 90% of the true null hypotheses were rejected for all three sample sizes. The USS test was much more stable overall than the standardised Pearson test.

It is noteworthy that among the seven statistics included in this study, X^2_{st} results in both the lowest $\hat{\alpha}$ and the highest $\hat{\alpha}$ for sample sizes $n = 500$ and $n = 1000$. There appears to be little reason to choose the standardised Pearson test over the USS test with regards to type I errors. When comparing $\hat{S}_{st}$ with the other five statistics, however, the choice is not that obvious.

The USS test had the strongest tendency to produce small rejection regions, i.e. reject the true H_0 less often than desired. Compared to the three Stukel's test statistics, $IMT1$, and

$IMT2$, $\hat{S}_{st}$ yielded more negative $\hat{\alpha}_s$'s which were of a considerable magnitude ($|\hat{\alpha}_s - \alpha| \geq 0.5$) and also had the highest frequency of negative $\hat{\alpha}_s$. This frequency, and the magnitudes, appear to decrease as n gets larger.

Stukel's score test outperformed Stukel's LRT1 and LRT2 in most situations. When $n = 100$, the score statistic had rejection rates that were considerably closer to 5% than the LRT1 and LRT2 statistics in 10 of the 12 situations. In situation 4, the LRT2 statistic had the same $\hat{\alpha}$, whereas in situation 7, LRT2 performed slightly better. These are the two situations which had the highest rate of failed replications that were mentioned earlier in this section. In the other situations where computation failed, Stukel's score test was so much better than the likelihood ratio based statistics that it is unlikely that the smaller bases for estimating $\hat{\alpha}_6$, $\hat{\alpha}_{10}$, $\hat{\alpha}_{11}$, and $\hat{\alpha}_{12}$ had any critical influence in this case. However, as sample size increased, the difference between the score test and the LRTs became smaller.

Compared to the $IMT1$ and $IMT2$ statistics, Stukel's score test does better in most situations when $n = 100$ and when $n = 500$. Among these three tests when $n = 1000$, however, the $IMT1$ has the most $\hat{\alpha}_s$ which are closer to α, though only by a small margin. Furthermore, the $IMT2$ generated $\hat{\alpha}_s$ with markedly better proximities to α than Stukel's score statistic when $n = 1000$.

For all three sample sizes, Stukel's score test was better than the standardised Pearson test in a large proportion of the 12 situations. The USS test had a more comparable performance. When $n = 100$ it achieved fairly similar proximities to α, whereas it had more situations where its $\hat{\alpha}_s$ was closer to α than Stukel's score test when $n = 500$. The opposite was true when $n = 1000$; Stukel's score test was better in a larger number of situations than the USS test.

Stukel's LRT1 and LRT2 statistics had identical results in six of the situations. In the situations 4, 10, 11, and 12, the difference in $\hat{\alpha}_s$ is very small and only present when $n = 100$. The LRT1 and LRT2 statistics do, however, perform very differently in situations 6 and 7. Recall from Figure 3.3a, that these are the situations that produce the two most highly skewed distributions of $\hat{\pi}_i$.

In situation 6, where the distribution of π_i is highly right skewed, LRT1 resulted in

$$\hat{\alpha}_6 - \alpha = 3.4, \text{ when } n = 100.$$

The corresponding result for LRT2 is 1.6, i.e. the difference between LRT2's $\hat{\alpha}_6$ and α is less than $1/2$ of difference between LRT1's $\hat{\alpha}_s$ and α. When $n = 500$, the statistics perform equally well, and this is also roughly the case when $n = 1000$.

Table 3.2 Simulated per cent rejection, $\hat{\alpha}_s \times 100$, at the $\alpha = 0.05$ level using sample sizes of 100, 500, and 1000, with 1000 replications.

		St. Pearson Test			USS Test			Stukel's Score Test			Stukel's LRT1			Stukel's LRT2			*IMT*1			*IMT*2		
Situation	Covariate distribution/ sample size	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000
1	$U(-6,6)$	5.4	3.7	4.8	4.9	4.8	5.3	5.8	4.5	5.0	7.8	5.3	5.5	7.8	5.3	5.5	5.8	5.2	4.6	3.9	5.3	4.5
2	$U(-4.5,4.5)$	4.8	4.0	5.1	4.1	5.1	4.8	5.2	5.0	5.6	7.8	5.6	6.2	7.8	5.6	6.2	4.9	5.0	5.5	3.5	4.7	5.4
3	$U(-3,3)$	11.8	5.3	5.0	5.4	4.6	5.3	4.5	5.8	6.5	6.8	5.9	6.6	6.8	5.9	6.6	4.9	6.0	5.9	4.6	6.0	5.9
4	$U(-1,1)$	93.2	95.2	94.2	5.0	3.7	4.6	6.4*	3.2	5.6	6.6	3.2	5.8	6.4	3.2	5.8	5.5	4.3	6.6	5.5	4.3	6.6
5	$\mathscr{N}(0,1.5)$	8.2	6.2	5.3	2.9	5.9	6.4	4.4	6.1	6.4	6.5	5.8	5.5	6.5	5.8	5.5	4.8	6.5	6.1	4.0	6.5	5.8
6	$\chi^2(4)$	6.7	2.7	3.0	4.4	5.2	4.5	5.0*	4.4	4.8	8.4	5.1	5.2	6.6	4.9	5.3	6.1	4.3	4.9	6.1	4.3	4.9
7	30×Beta$(18,2)$	7.5	4.8	6.5	4.2	3.6	5.7	4.2*	4.2	5.2	7.5	4.2	7.4	5.5	4.2	5.4	5.2	3.4	4.9	4.9	3.4	5.0
8	3 Independent $U(-6,6)$	6.9	6.0	5.5	4.9	5.5	6.0	5.0	6.1	4.9	8.3	5.1	5.7	8.3	5.1	5.7	7.2	6.0	5.1	5.9	5.7	4.6
9	3 Independent $\mathscr{N}(0,1.5)$	55.7	31.3	18.8	4.4	4.8	4.9	4.1	4.8	4.8	7.8	5.3	5.7	7.8	5.3	5.7	6.2	5.9	4.9	6.3	5.8	5.0
10	Independent $\chi^2(4)$ and 30×Beta$(18,2)$	29.8	7.9	5.8	4.8	3.6	4.1	5.3*	4.8	5.4	7.4	5.7	6.2	7.5	5.7	6.2	6.5	6.2	5.8	6.3	6.1	5.7
11	Independent $U(-6,6)$, $\mathscr{N}(0,1.5)$ and $\chi^2(4)$	16.4	5.0	6.4	3.7	4.8	5.5	5.5*	4.9	5.5	6.9	5.9	6.5	6.8	5.9	6.5	5.6	6.4	4.4	5.0	5.8	4.5
12	Independent $U(-6,6)$, $\mathscr{N}(0,1.5)$ and 30×Beta$(18,2)$	14.6	6.7	6.5	4.6	5.6	6.4	4.5*	6.0	5.8	6.4	5.9	6.1	6.3	5.9	6.1	5.7	6.2	5.3	5.2	6.2	5.3

In situation 7, where the distribution of π_i is highly left skewed, LRT1 resulted in

$$\hat{\alpha}_7 - \alpha = 2.5, \text{ when } n = 100, \text{ and}$$
$$\hat{\alpha}_7 - \alpha = 2.4, \text{ when } n = 1000.$$

The equivalent differences for the LRT2 statistics are respectively 0.5 and 0.4, i.e. they are $1/5$ or less than their LRT1 counterparts. When $n = 500$, however, the two statistics reject H_0 at the same rate in this study.

When comparing the information matrix test statistics to each other, it appears that they perform the most unequally for small sample sizes, and that their results converge as n gets larger. This is consistent with the fact that $IMT1$ and $IMT2$ are asymptotically equivalent. If one compares the statistics' $\hat{\alpha}_s$'s for each $s = 1, 2, \dots, 12$, it becomes apparent that $IMT1$ performs better in slightly more situations than $IMT2$ when $n = 100$. In contrast, the results show that $IMT2$ comes closer to a empirical significance level of 5% more frequently than $IMT1$ when $n = 500$ and $n = 1000$.

3.5 Study of the Weibull-based behaviour indicator applied to simulation data

In the previous section, we presented how our seven GOF statistics performed for each situation s from Table 3.1, $s = 1, \dots, 12$. During the $R = 1000$ replications, where we computed the p-values of the GOF statistics, we also calculated the distances $d_{i,s}$, $i = 1, \dots, n$, and subsequently fitted a Weibull distribution to $d_{i,s}$. First the maximum likelihood (ML) estimated parameters were computed by the `fitdist()` function. Values of $\hat{\pi}_i$ which were less than 1×10^{-8} and greater than $1 - 1 \times 10^{-8}$ were excluded in order to ensure successful computation

The estimates where then saved in a $R \times 2$ table. Finally, each column of this table was averaged and resulted in the two empirical shape and scale parameters, denoted $\hat{a}_s$ and $\hat{b}_s$ respectively. The resulting $\hat{\boldsymbol{a}} = [\hat{a}_1 \;\; \hat{a}_2 \;\; \dots \;\; \hat{a}_{12}]^\mathsf{T}$ and $\hat{\boldsymbol{b}} = \left[\hat{b}_1 \;\; \hat{b}_2 \;\; \dots \;\; \hat{b}_{12}\right]^\mathsf{T}$ are summarized in Table 3.3. For each n-value there are 12 fitted Weibull distributions corresponding to the 12 situations listed in in Table 3.1.

In addition to looking at $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$, we were also interested in what the 12 fitted Weibull PDFs look like. How asymmetric are they? Which values are they centred around, and how are they dispersed? Plotting the PDFs for $n = 1000$ in Figure 3.4 gives us a visual impression

Table 3.3 The shape and scale parameters of the fitted Weibull distributions

	Shape parameter, $\hat{a}_s$			Scale parameter, $\hat{b}_s$		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
Situation 1	0.78584	0.79323	0.79630	0.17334	0.17740	0.17841
Situation 2	1.05705	1.05560	1.06238	0.26188	0.26654	0.26871
Situation 3	1.60394	1.59228	1.59393	0.37921	0.38714	0.38836
Situation 4	5.26034	4.88266	4.85821	0.50969	0.51703	0.51789
Situation 5	1.92510	1.94289	1.94553	0.42338	0.43546	0.43703
Situation 6	1.14757	1.13990	1.14204	0.27301	0.27660	0.27799
Situation 7	1.21477	1.22247	1.22730	0.29087	0.29970	0.30174
Situation 8	1.36765	1.42546	1.42028	0.34867	0.36818	0.36852
Situation 9	3.02649	3.27066	3.32633	0.48199	0.49883	0.50118
Situation 10	2.59270	2.65876	2.68150	0.46621	0.48010	0.48233
Situation 11	1.65364	1.72877	1.73836	0.39196	0.41126	0.41378
Situation 12	1.53588	1.59353	1.59848	0.37450	0.39323	0.39517

Table 3.4 Parameters and measures of the fitted Weibull distributions, $n = 1000$

	Shape parameter, $\hat{a}_s$	Scale parameter, $\hat{b}_s$	Mode, Mo_s	Median, Md_s	Mean, μ_s	Standard deviation, σ_s	Skewness, $\gamma_{1,s}$
Situation 1	0.7963	0.1784	0.0000	0.1126	0.2028	0.2569	2.8353
Situation 2	1.0624	0.2687	0.0186	0.1903	0.2624	0.2471	1.8267
Situation 3	1.5939	0.3884	0.2091	0.3086	0.3483	0.2237	0.9682
Situation 4	4.8582	0.5179	0.4939	0.4803	0.4747	0.1116	-0.2340
Situation 5	1.9455	0.4370	0.3016	0.3620	0.3875	0.2077	0.6685
Situation 6	1.1420	0.2780	0.0448	0.2017	0.2651	0.2327	1.6391
Situation 7	1.2273	0.3017	0.0764	0.2238	0.2822	0.2312	1.4701
Situation 8	1.4203	0.3685	0.1564	0.2847	0.3351	0.2393	1.1713
Situation 9	3.3263	0.5012	0.4501	0.4489	0.4497	0.1490	0.0707
Situation 10	2.6815	0.4823	0.4053	0.4207	0.4288	0.1723	0.2824
Situation 11	1.7384	0.4138	0.2528	0.3351	0.3687	0.2187	0.8308
Situation 12	1.5985	0.3952	0.2137	0.3142	0.3543	0.2269	0.9635

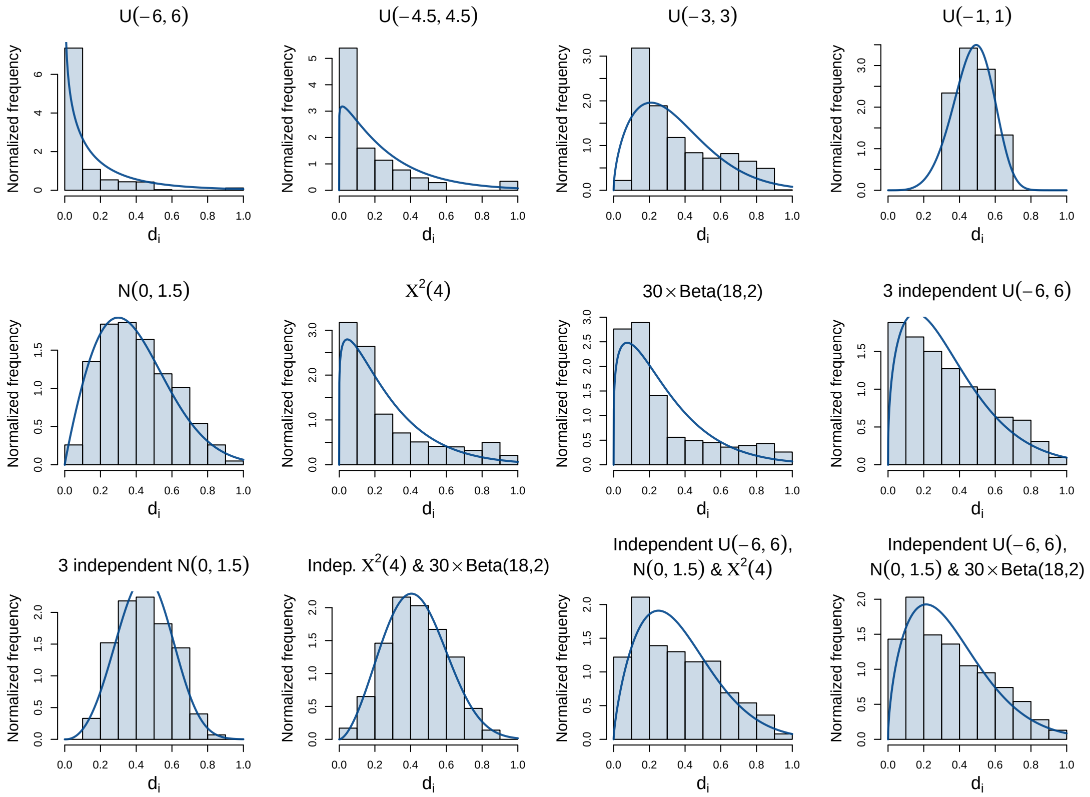


Fig. 3.4 The PDFs of the $Weibull(\hat{a}_s, \hat{b}_s)$ distributions ($n = 1000$) accompanied by their own illustrative example histogram of d_i (resulting from 12 generated data examples adhering to the set-ups in Table 3.1)

of the centre measures and dispersion, and thus how the $d_{i,s}$'s are distributed. The values of such measures are also useful and are included in Table 3.4. Each row in Table 3.4 shows the $\hat{a}_s$ and $\hat{b}_s$ belonging to situation s when $n = 1000$, along with the mode, median, mean, variance, standard deviation and skewness of the *Weibull*$(\hat{a}_s, \hat{b}_s)$ distribution.These measures are denoted Mo_s, Md_s, μ_s, σ_s^2, σ_s and γ_1 respectively.

The formulas for the expectation and variance of a random variable, say T, with the two-parameter *Weibull*$(\hat{a}_s, \hat{b}_s)$ distribution are

$$\mu_s = E(T) = \hat{b}_s \Gamma\left(1 + \frac{1}{\hat{a}_s}\right), \text{ and} \tag{3.3}$$

$$\sigma_s^2 = Var(T) = \hat{b}_s^2 \left\{ \Gamma\left(1 + \frac{2}{\hat{a}_s}\right) - \left[\Gamma\left(1 + \frac{1}{\hat{a}_s}\right)\right]^2 \right\}, \tag{3.4}$$

(Devore and Berk, 2012). The mode and median of the same distribution are defined as

$$Mo_s = \hat{b}_s \left(\frac{\hat{a}_s - 1}{\hat{a}_s}\right)^{\frac{1}{\hat{a}_s}}, \text{ for } \hat{a}_s > 1 \text{ (0 otherwise), and} \tag{3.5}$$

$$Md_s = \hat{b}_s (\log 2)^{\frac{1}{\hat{a}_s}}, \tag{3.6}$$

according to Lai (2013).

A distribution's lack of symmetry is often measured by its skewness. The chosen skewness measure in this text is the standardised third central moment and denoted by γ_1. A random variable T which is *Weibull*$(\hat{a}_s, \hat{b}_s)$ distributed has the following skewness:

$$\gamma_{1,s} = E\left[\left(\frac{T - \mu_s}{\sigma_s}\right)^3\right] = \frac{E\left[(T - \mu_s)^3\right]}{(\sigma_s^2)^{3/2}} \tag{3.7}$$

$$= \frac{\Gamma(1 + \frac{3}{\hat{a}_s}) - 3\Gamma(1 + \frac{2}{\hat{a}_s})\Gamma(1 + \frac{1}{\hat{a}_s}) + 2\left[\Gamma(1 + \frac{1}{\hat{a}_s})\right]^3}{\left(\Gamma(1 + \frac{2}{\hat{a}_s}) - \left[\Gamma(1 + \frac{1}{\hat{a}_s})\right]^2\right)^{3/2}}, \tag{3.8}$$

(McCool, 2012). The larger the $\gamma_{1,s}$, the larger the degree to which the distribution strays from symmetry.

So now we have a visual representation of the Weibull PDFs and a summary where $\hat{a}_s$ and $\hat{b}_s$ are accompanied by their corresponding μ_s, σ_s , etc. We wish to study if one can predict how the GOF test statistics will perform based on this information. After inspecting the graphs of the Weibull PDFs in Figure 3.4, one can argue that they come in three main categories with regards to skewness: (1) close to symmetric distributions; (2) moderately right skewed distributions; and (3) highly right skewed distributions. There might, however, be potentially informative groupings of the distributions that are harder to spot. For this reason, principal component analyses (PCA) were carried out.

3.5.1 Principal Component Analysis

The principal component analyses were performed with the `prcomp()` function following the recommendations of James et al. (2013). In order to emphasize whether a GOF test statistic performs conservatively, anti-conservatively, or as desired, the following variable was introduced:

$$\varDelta_s = (\hat{\alpha}_s - \alpha) \times 100, \tag{3.9}$$

where $\hat{\alpha}_s$ is the empirical significance level of a GOF test statistic when $n = 1000$ in situation s, $s = 1, 2, \ldots, 12$. Negative $\varDelta_s$ of substantial magnitude indicate that the GOF test statistic in question is conservative.

Conversely, considerably large positive values mean that the statistic is too intolerant or anti-conservative. There appears to be no well-established standard by which to interpret the distance between $\hat{\alpha}_s$ and the chosen significance level α. Nevertheless, in this text, the GOF test statistic will be regarded as having an acceptable rejection region if $|\varDelta_s| \leq 0.5$.

For each GOF test, a dataset was formed by merging the observed values of $\varDelta_s$ and the columns of Table 3.4. Seven PCAs were carried out using seven 12×10 data matrices in `R`, and the first two principal components (PCs) were used to produce a loading vector plot (also called loading plot). A biplot in PCA can reveal clustering of observations. and correlations of the variables of a dataset (Gabriel, 1971). A loading plot is essentially a biplot, but without the visualization of the PC scores that describe the observations. In our case, the observations are the 12 different situations, and the variables are $\varDelta_s$, $\hat{a}_s$, $\hat{b}_s$, Mo_s, Md_s, μ_s, σ_s^2, σ_s and $\gamma_{1,s}$. Loading plots were used instead of biplots since the initial focus was on uncovering possible correlations between the GOF tests' $\varDelta_s$ and the variables.

The PCA loading vector plot provides us with a visual overview of the variables listed in Table 3.4 and the variable $\varDelta_s$ which expresses how close $\hat{\alpha}_s$ of a specific test is to α. It does not, however, determine the true values of the correlation coefficients between the variables included in the PCA or give us an explicit answer to whether the $Weibull(\hat{a}_s, \hat{b}_s)$ distributions' parameters, and its measures of dispersion and central tendencies, can be used to predict how a particular test performs in a particular type of situation s. A loading plot is merely a useful tool for extracting the most prominent or interesting relationships between variables from a data structure.

Figure 3.5 contains one loading plot for each GOF test statistic where the first two principal components of the PCA have been plotted. According to Gabriel (1971), the cosine of the angle between two vectors pertaining to two variables/columns of a data set is an approximation of the correlation coefficient of those two variables.

In the following text, the variables $\hat{b}_s$, Mo_s, Md_s, and μ_s will be referred to as the *percentile-related variables*. The scale parameter $\hat{b}_s$ approximates the 63.2th percentile of the fitted $Weibull(\hat{a}_s, \hat{b}_s)$ distribution, the median Md_s equals the 50th percentile and Mo_s and μ_s equal percentiles not far from the median.

The Standardised Pearson Test

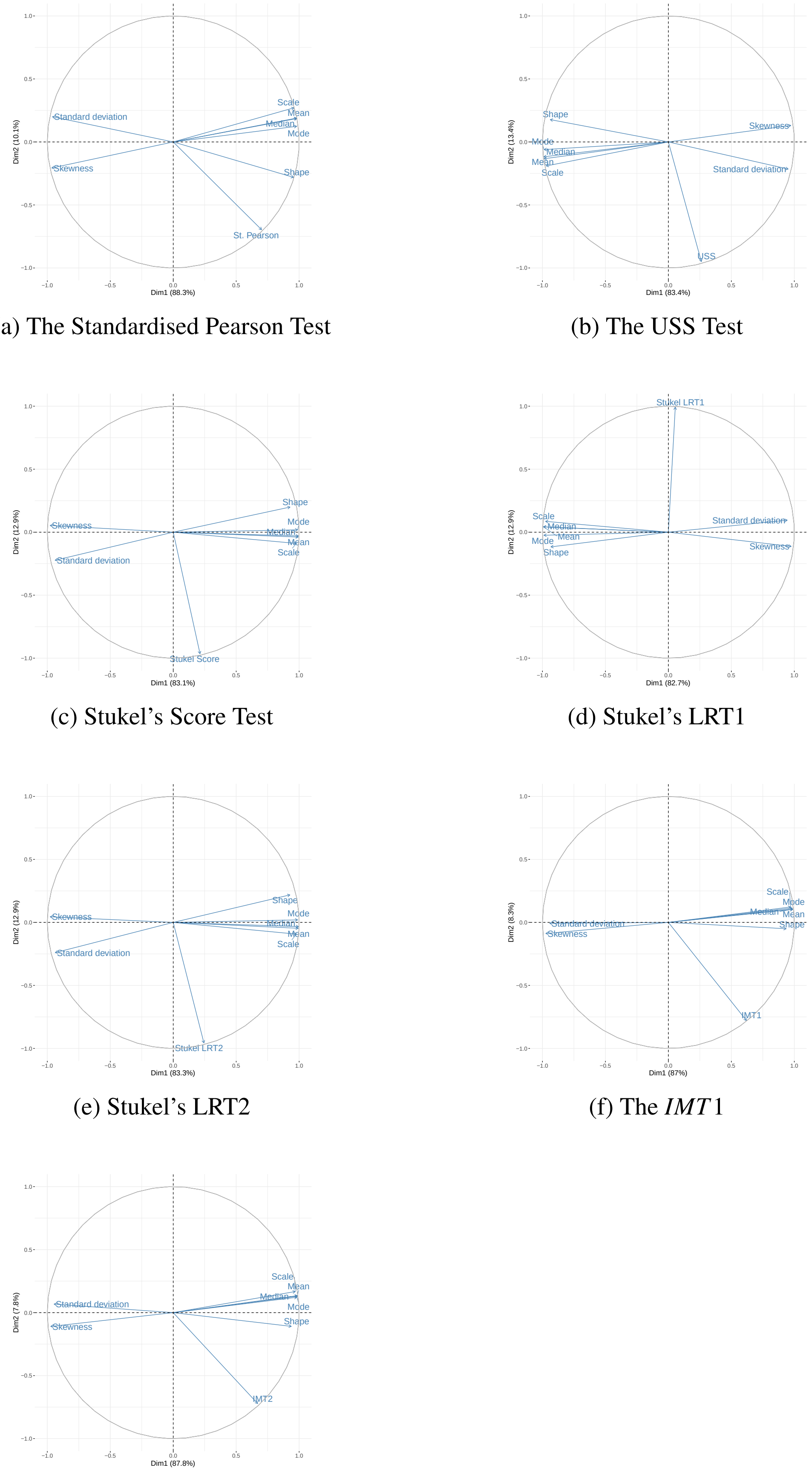
The standardised Pearson test's performance, Δ_s, has the strongest positive correlation with the shape parameter. The test is also positively correlated with the percentile-related parameters (Mo_s, μ_s, $\hat{b}_s$ and Md_s), but their vectors form angles with the Δ_s vector that are roughly twice as large. There are only two parameters which are negatively correlated with the standardised Pearson test's Δ_s, namely σ_s and $\gamma_{1,s}$. The standard deviation σ_s has the strongest negative correlation to Δ_s, with an angle between which is approximately 147.0°. The angle between Δ_s and $\gamma_{1,s}$ is approximately 123.1°.

The USS Test

Performing a PCA on the data set with the USS test's Δ_s, results in the loading plot of the first two principal components shown in Figure 3.5b. This plot shows a pattern of vectors similar to Figure 3.5a. The percentile-related variables are clustered together, and pointing in the opposite direction of $\gamma_{1,s}$. The shape parameter $\hat{a}_s$ is pointing in the same horizontal direction as the percentile-related cluster, but in the opposite direction of σ_s.

The relationships between Δ_s and the other variables, however, are the opposite of what we see in Figure 3.5a (although not the exact opposites). The variable with the strongest positive correlation to Δ_s is σ_s. The skewness $\gamma_{1,s}$ is also positively correlated, but only slightly. The rest of the parameters are negatively correlated with Δ_s, where $\hat{a}_s$ forms the largest angle (approximately 116°) with Δ_s. As seen in Figure 3.5a, $\hat{a}_s$ is isolated from the cluster of the percentile-related variables.

Stukel's Score Test



(a) The Standardised Pearson Test

(b) The USS Test

(c) Stukel's Score Test

(d) Stukel's LRT1

(e) Stukel's LRT2

(f) The $IMT1$

(g) The $IMT2$

Fig. 3.5 The loading vector plots of the first two PCs from the PCAs performed on the data sets comprised by Δ_s and their corresponding rows in Table 3.4. Each variable has a vector which represents its PC loadings. The Δ_s vectors are labelled by the name of their respective test.

The first two principal components of the PCA performed with Stukel's Score test's $\varDelta_s$ are plotted in Figure 3.5c. The vector pattern looks like the USS loading plot has been mirrored about the vertical axis, or the standardised Pearson loading plot mirrored about the horizontal axis. Despite these pattern similarities, $\varDelta_s$ appears to have weaker correlations with the other variables than what we see in Figure 3.5a and Figure 3.5b. This observation is supported by the size of the angles between the vectors.

Stukel's score test's performance $\varDelta_s$ has the strongest positive correlation with $\hat{b}_s$, closely followed by μ_s, Md and Mo. The only parameter with a negative correlation with $\varDelta_s$ in this plot is $\gamma_{1,s}$. The angle between their vectors is approximately 105.5°, hence the correlation does not appear to be very strong.

There are two variables whose loading vectors are nearly perpendicular to the $\varDelta_s$ loading vector, namely $\hat{a}_s$ and σ_s. The angles between the $\varDelta_s$ vector and the vectors of $\hat{a}_s$ and σ_s, are approximately equal to 89.7° and 89.1°, respectively. Hence these two variables might be uncorrelated with the $\varDelta_s$ of Stukel's Score test, $s = 1, 2, \ldots, 12$.

Stukel's LRT1

Overall, the strength of the linear relations between the $\varDelta_s$ variable and the other variables appears to be weaker for Stukel's LRT1 dataset compared to the standardised Pearson dataset and the USS dataset.

The strongest positive correlation between the LRT1 test's performance and another variable is the one with σ_s. The angle between their vectors, however, is approximately 81.1°. This is much closer to 90° than the less than 28.4° angle between the vectors for the standardised Pearson test's $\varDelta_s$ and the shape parameter, for example.

Stukel's LRT1 test's $\varDelta_s$ has the strongest negative correlation with $\hat{a}_s$. The vectors belonging to $\varDelta_s$ and $\hat{a}_s$ in Figure 3.5d form an obtuse angle approximately equal to 100.3°. Thus, the PCA indicates that Stukel's LRT1 is not particularly strongly correlated with any of the Weibull parameters. The five remaining parameters, $\gamma_{1,s}$, $\hat{b}_s$, Mo, and finally the two overlapping μ_s and Md, are only a few degrees away from being perpendicular to the $\varDelta_s$ vector.

Stukel's LRT2

The vector pattern in Stukel's LRT2 loading plot in Figure 3.5e is fairly similar to its LRT1 counterpart, but indicates slightly different linear relations between Stukel's LRT2 $\varDelta_s$ and the other variables.

According to the first two principal components, $\hat{b}_s$ is the variable with the strongest positive correlation with Δ_s. The vectors of Mo_s, Md_s and μ_s also form acute angles with the Δ_s vector, but they are slightly closer to 90°. Hence percentile-related variables have the vectors which make the smallest angles between themselves and the LRT2 test performance vector.

The shape parameter $\hat{a}_s$ is clearly separated from the four percentile-related variables, as indicated by the approximately 89.0° angle between its vector and the LRT2 test's Δ_s vector. The standard deviation σ_s, which has the strongest positive correlation to Δ_s in the PCA performed on Stukel's LRT1 dataset, has an angle approximately equal to 89.929° between its own vector and Stukel's LRT2 Δ_s vector. This might indicate that the modification of Stukel's LRT1 test causes the standard deviation to be less positively correlated with the accuracy of $\hat{\alpha}$ belonging to Stukel's LRT2 test.

In terms of negative correlation with Δ_s, the only variable with this characteristic in Stukel's LRT2 loading plot is $\gamma_{1,s}$. Its vector forms an angle with the Δ_s vector which is approximately equal to 106.9°. This suggests that there is a slightly stronger relationship than the one between Stukel's LRT1 Δ_s and $\hat{a}_s$.

IMT1

According to the first two principal components plotted in Figure 3.5f, the $IMT1$'s performance Δ_s has the strongest positive correlation with $\hat{a}_s$. Similarly to all the other tests, the shape parameter $\hat{a}_s$ is isolated from the percentile-related cluster of variables, but they all point towards the same vertical edge of the plot. Similarly to the $\hat{a}_s$ vector, the vectors of $\hat{b}_s$, Mo_s, Md_s and μ_s also form angles less than 90° with the $IMT1$'s vector, but they are noticeably larger.

The $IMT1$ Δ_s has the strongest negative linear relation with σ_s, closely followed by $\gamma_{1,s}$. None of the Weibull-related vectors are close to being perpendicular to the Δ_s vector, i.e. their correlations with the performance of the $IMT1$ test statistic are fairly strong according to this PCA.

IMT2

The loading vector plot Figure 3.5f is fairly similar to its $IMT1$ counterpart, but there are some noticeable differences. The $IMT2$ Δ_s is also the most positively correlated with $\hat{a}_s$, but the relation is slightly stronger than the one between the $IMT1$ Δ_s and $\hat{a}_s$. The percentile-related variables are also positively correlated with the $IMT1$ Δ_s, but to a lesser degree than $\hat{a}_s$.

Another similarity is that the $IMT2$ Δ_s vector has the strongest negative correlation with σ_s variable. However, both $\gamma_{1,s}$ and σ_s form larger obtuse angles with the Δ_s variable in this vector loading plot compared to the $IMT2$ plot.

Summary

Given the above information, there are two sets of variables which emerge as potential predictors of Δ_s if we focus on the two variables with the strongest correlation with Δ_s in each PCA. The first set consists of the shape parameter $\hat{a}_s$ and the standard deviation σ_s. The the second set consists of the scale parameter $\hat{b}_s$ and skewness $\gamma_{1,s}$.

The shape parameter and standard deviation set had the strongest correlation with Δ_s for the following GOF tests: the Standardised Pearson test, the USS test, Stukel's LRT1, $IMT1$ and $IMT2$. The scale parameter and skewness set had the strongest correlation with Δ_s for Stukel's Score test and Stukel's LRT2. Note that the tests with the weakest correlations between Δ_s and the Weibull-related variables are the three Stukel's tests. Hence this text will be limited to investigating the set with $\hat{a}_s$ and σ_s and its prospective use in predicting the behaviour of the Standardised Pearson test, the USS test, Stukel's LRT1, $IMT1$ and $IMT2$.

3.5.2 The Weibull shape parameter and standard deviation as potential indicators of empirical significance levels

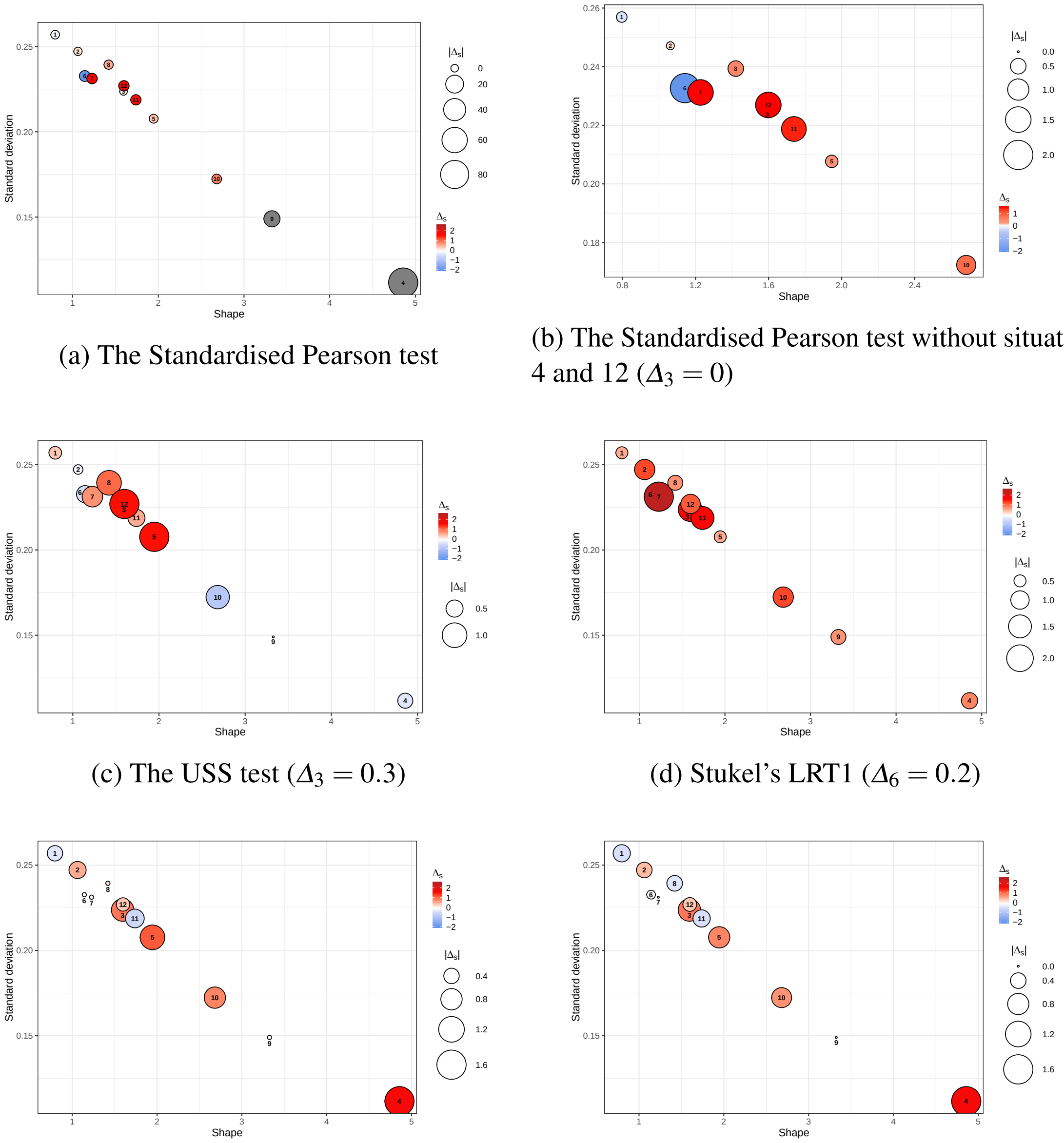
The standardised Pearson test

As mentioned in section 3.4, the standardised Pearson chi-square statistic resulted in some very high empirical rejection rates of the true null hypothesis. This can be seen in Figure 3.6a, where the circles representing $|\Delta_4|$ and $|\Delta_9|$ are much larger than the circles belonging to the remaining situations. Both extreme values for $\hat{\alpha}_s$ coincide with the two largest pairs of $\hat{a}_s$ and σ_s values, which belong to situations 4 and 9.

Conservative $\hat{\alpha}_s$ occur only when the estimated shape parameter is relatively small and standard deviation is large, but the more precise $\Delta_2 = 0.1$ is also in this area.

The 4 largest circles (excluding situation 4 and 9) coincide with $\hat{a}_s \in [1.1420, 1.7384]$ and $\sigma_s \in [0.2187, 0.2327]$, but $\Delta_3 = 0$ and is also in this area of the heat map. Situation 3 and 12 have values for $\hat{a}_s$ and σ_s that are very close to each other, yet $\Delta_{12} = 1.5$ which is relatively larger than Δ_3. This might be explained, however, by the fact that situation 12 produces parameters and measures that are all greater than the ones produced by situation 3.

Despite being very close to each other in Figure 3.6b, Δ_6 and Δ_7 have opposite signs, i.e. the test is conservative in situation 6, but too quick to reject in situation 7. Examining these two situations reveals that situation 6 produces values smaller than situation 7 for



(a) The Standardised Pearson test

(b) The Standardised Pearson test without situation 4 and 12 ($\Delta_3 = 0$)

(c) The USS test ($\Delta_3 = 0.3$)

(d) Stukel's LRT1 ($\Delta_6 = 0.2$)

(e) The $IMT1$

(f) The $IMT2$

Fig. 3.6 Heat maps of five GOF tests' observed Δ_s at different values of $\hat{a}_s$ and σ_s. The colour of the circles indicates the value of Δ_s, i.e. if the test's $\hat{\alpha}$ is smaller or larger than α. The circle sizes indicate how much distance there is between $\hat{\alpha}$ and α, i.e. $|\Delta_s|$.

almost all parameters and measures listed in Table 3.4, except for σ_s and $\gamma_{1,s}$. Furthermore, the measures of central tendency, Mo_s, Md_s and μ_s , are further apart in situation 6 than in situation 7. These particularities might explain why Δ_6 and Δ_7 have opposite signs, but this study is not comprehensive enough to determine this with certainty.

The USS test

One can see in Figure 3.6c, that the occurrences of negative Δ_s are more dispersed for the USS test than for the standardised Pearson test.

Similarly to the standardised Pearson test, the Δ_s belonging to situations 6 and 7 have opposite signs, but the difference is not as pronounced. Another similarity is the difference between Δ_3 and Δ_{12} despite adjacent $\hat{a}_s$ and σ_s values. Furthermore, the USS test heat map also exhibits a cluster of large Δ_s in approximately the same area of the plot as the standardised Pearson test heat map. It is worth noting that at least half of the situations used in this simulation study produce fitted Weibull distributions with shape parameters and standard deviations that fall within this area of the heat maps' coordinate plane. So the fact that larger values of Δ_s are clustered together in this region might be incidental.

In this heat map, the four largest red circles (i.e. the four most anti-conservative situations) are contained in the area of the plot where $\hat{a}_s \in [1.2272, 1.9455]$ and $\sigma_s \in [0.2077, 0.2393]$. Situations 3 and 11 are also in this area, but Δ_3 and Δ_{11} are equal to 0.3 and 0.5, respectively.

Stukel's LRT1

The six situations with the largest values of Δ_s for Stukel's LRT1 are situations 7, 3, 11, 2, 10, and 12 (in descending order). Five of these six are found in the same area of the heat map as the clusters mentioned for the standardised Pearson test and the USS test.

As seen in Figures 3.6b and 3.6c, $|\Delta_{10}|$ is considerable large, but has a lower standard deviation and higher shape parameter which separates it from the main cluster of large Δ_s's. This is also the case in the heat maps belonging to $IMT1$ and $IMT2$.

IMT1 and IMT2

Figures 3.6e and 3.6f show larger Δ_4 than the USS test and Stukel's LRT1, suggesting that $IMT1$ and $IMT2$ might be more sensitive to large values of $\hat{a}_s$ and small values of σ_s. Situation 4 is also the only situation with $\gamma_{1,s} < 0$. The circles where $\sigma_s \in [0.225, 0.257]$ are all of an acceptable size, i.e. $|\Delta_s| \leq 0.5$. All of the negative Δ_s have $\hat{a}_s < 1.75$, except for Δ_9 in the $IMT1$ heat map. Note that situation 9 has a much lower σ_s and also the $\gamma_{1,s}$ which is the closest to 0.

The four situations with the largest $|\Delta_s|$ are situations 3, 4, 5, and 10. Their circles do not fall into the upper left corner of the heat maps. They have standard deviations which are less than 1.75. Situations 11 and 9 are also placed in this area of heat map. For $IMT1$, $\Delta_{11} = -0.6$, whereas for $IMT2$ $\Delta_{11} = 0.5$. As mentioned previously, situation 9 has the skewness $\gamma_{1,s}$ which is the closest to 0, so that might be what sets it apart from situation 3, 4, 5, and 10.

3.5.3 Future work

It was beyond the scope of this thesis to study this potential behaviour indicator further. If time had allowed, we could have performed an additional principal component analysis, similar to the one in Section 3.5, where the variables would be the different measures and parameters of the fitted Weibull distribution and the estimated power of the tests for sample sizes equal to 1000.

It would be interesting to see if the same variables were correlated with high power, as the variables that were correlated with the differences delta between the empirical and the nominal significance level. Perhaps the variables which are the most highly correlated are different to the shape parameter and the standard deviation of the fitted Weibull distribution. If that were the case, then one could look at two separate sets of viable Weibull parameters and dispersion measurements to gauge how the goodness of fit test statistic will behave in different settings.

If a visualisation tool could be made based on fitted Weibull distribution parameters and measurements that quickly and intuitively could advise the user of how the goodness of the tests potentially may behave, this would be a useful addition when performing data analysis. It could potentially be time-saving and a more readable alternative to studying literature describing the behaviours of the tests in written form.

However, an extensive amount of additional work is necessary to see if this may be accomplished. A much more extensive range of distributions of fitted logistic probabilities and larger samples of such distributions should be included in the study if this were to be examined again. 12 estimated Weibull distributions for every sample size is unlikely an optimal amount. This task is sadly to broad and comprehensive to be included in the subject area of this thesis, and would hence be an interesting topic to cover in future work.

Chapter 4

Power Study

The study in this chapter was the second of two simulations studies to be conducted. In addition to having appropriate rates of rejection when the fitted logistic model is correct, it is also desirable for GOF tests to recognise when the null hypothesis is in fact false. A *type II error* occurs when a test fails to reject H_0 when H_0 is false (Devore and Berk, 2012). If a GOF test fails to reject H_0 (the null hypothesis that the specified model is true) when the specified model is missing one or multiple effects present in the true model, it has committed a type II error. Likewise, the failure to discern that the specifications of an incorrect link function $g(\pi_i)$ is also classified as a type II error.

The *power* of a test is the probability of not committing a type II error, hence it is a central quality when evaluating a GOF test statistic. We need to know how often a test detects that the model claimed to be true in H_0 deviates from the true model (when this is in fact the case).

4.1 Types of departure

A discrepancy between the true model and a specified model will be referred to as a departure from the true model. This part of the simulation study examined the ability of the test statistics to recognize four specific types of departures from the correct model. These departure types are denoted by *D1*, *D2*, *D3*, and *D4*, and each of them consists of several models with varying severity of misspecification in order to gauge the GOF tests' sensitivity. The types of departures addressed in the study were:

D1: the omission of the quadratic term from a linear predictor with one continuous covariate,

D2: the omission of the log term from a linear predictor with one continuous covariate,

D3: the omission of the main effect of a binary covariate and its interaction with a continuous covariate, and

D4: the selection of an incorrect link function.

Algorithms which produced the observed rejection rates of the GOF statistics were implemented in R. A general scheme of these algorithms is provided in the next subsection and is followed by more in-depth descriptions that are specific to each of the four departure types.

4.2 A General Outline of the Power Study Design

This section outlines the simulation procedure used to estimate the power of the GOF statistics. The situations and structure used to examine the power bears a close resemblance to that of Hosmer et al. (1997). As indicated in Section 3.1, there were cases where the computation of Stukel's score statistic failed. They were dealt with using the same approach as in the significance level study.

For each departure type and sample size, the following sequence of steps were taken:

(i) ***The true models and their respective covariate effects were specified.***
A set of κ true models, $M_{T_1}, M_{T_2}, \ldots, M_{T_\kappa}$, with true covariate effects, $\boldsymbol{\beta}_{T_1}, \boldsymbol{\beta}_{T_2}, \ldots, \boldsymbol{\beta}_{T_\kappa}$, were specified. The constant κ is a positive integer, and $\boldsymbol{\beta}_{T_j} = \begin{bmatrix}\beta_{j0} & \beta_{j1} & \ldots & \beta_{jk}\end{bmatrix}^\mathsf{T}$, $j = 1, \ldots, \kappa$. Recall from Section 1.1 that $\boldsymbol{\beta}_{T_1}, \boldsymbol{\beta}_{T_2}, \ldots, \boldsymbol{\beta}_{T_\kappa}$ are $(k+1) \times 1$ vectors and that k is the number of parameters in a GLM. For *D1*, *D2*, and *D3*, the only component separating these κ models is $\boldsymbol{\beta}_{T_j}$, $j = 1, \ldots, \kappa$. These $\boldsymbol{\beta}_{T_j}$ were chosen so that the effect of the characteristic in focus ranged from low to substantial, whilst attempting to keep the distributions of the π_i's produced as similar as reasonably possible. For *D4*, the linear predictors were identical, i.e. $\boldsymbol{\beta}_{T_1} = \boldsymbol{\beta}_{T_2} = \ldots = \boldsymbol{\beta}_{T_\kappa}$, and $M_{T_1}, M_{T_2}, \ldots, M_{T_\kappa}$ only differed by the choice of link functions g.

(ii) ***Observations of the covariate(s) and the binary responses were simulated.***
The following procedure was carried out once for each sample size n:

a) A generated sample of the covariate(s) was used to define a $n \times (k+1)$ design matrix $\boldsymbol{X}_{n\times(k+1)}$.

b) For every M_{T_j}, $\boldsymbol{X}_{n\times k+1}$ was multiplied by $\boldsymbol{\beta}_{T_j}$ to produce a $n \times 1$ vector consisting of the model's linear predictors, namely $\boldsymbol{\eta}_j = \begin{bmatrix}\eta_{j1} & \eta_{j2} & \dots & \eta_{jn}\end{bmatrix}^{\mathsf{T}} = \boldsymbol{X}_{n\times k+1}\boldsymbol{\beta}_{T_j}$. This resulted in κ vectors – one for each true model.

c) For every M_{T_j}, its corresponding $\boldsymbol{\eta}_j$ was used to compute a $n \times 1$ vector of logistic probabilities, namely $\boldsymbol{\pi}_{T_j} = \begin{bmatrix}\pi_{j1} & \pi_{j2} & \dots & \pi_{jn}\end{bmatrix}^{\mathsf{T}}$, where $\pi_{ji} = g^{-1}(\eta_{ji})$, $i = 1,\dots,n$. For *D1*, *D2*, and *D3*, $g^{-1}(\eta_{ji}) = {}^{e^{\eta_{ji}}}/_{(1+e^{\eta_{ji}})}\ \forall\ j = 1,\dots,\kappa$. For *D4*, g^{-1} was different for each M_{T_j}. After these computations there were κ vectors of true probabilities.

d) The resulting $\boldsymbol{\pi}_{T_1},\dots,\boldsymbol{\pi}_{T_\kappa}$ served as input when generating κ vectors of the response variable, denoted $\boldsymbol{y}_j = \begin{bmatrix}y_{j1} & y_{j2} & \dots & y_{jn}\end{bmatrix}^{\mathsf{T}}$. These vectors of dichotomous outcomes were reproducibly generated in `R` using the `rbinom()` function. Each generated $\boldsymbol{y}_{T_j}$ was based on the exact same design matrix, $\boldsymbol{X}_{n\times k+1}$.

(iii) ***The incorrect models were specified.***

At this stage, models with the predetermined type of departure were specified. These κ models are denoted by $M_1, M_2, \dots, M_\kappa$. For *D4*, the components of M_j were equal to that of M_{T_j} except for the link function g. For *D1*, *D2*, and *D3*, M_j were deliberately misspecified by defining a simpler systematic component than in M_{T_j}, i.e. at least one of the parameters present in $\boldsymbol{\beta}_{T_j}$ were omitted (M_{T_j} and M_j are nested models). The design matrix of the incorrect models, denoted by $\boldsymbol{X}$, was equal to $\boldsymbol{X}_{n\times k+1}$, but without the column(s) corresponding to the omitted parameter(s).

(iv) ***The incorrect models were fitted.***

The computed $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_\kappa$ and $\boldsymbol{X}$ could then be used to obtain estimates of the covariate effects for the incorrect models. These estimates, denoted by $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \dots, \hat{\boldsymbol{\beta}}_\kappa$, were calculated using the `glm()` function in `R`. Once the values of $\hat{\boldsymbol{\beta}}_j$ were fitted, it was possible to find the estimated linear predictors, $\hat{\boldsymbol{\eta}}_j$, and subsequently the estimated logistic probabilities, $\hat{\boldsymbol{\pi}}_j$. For every M_j,

$$\hat{\boldsymbol{\eta}}_j = \begin{bmatrix}\hat{\eta}_{j1} & \hat{\eta}_{j2} & \dots & \hat{\eta}_{jn}\end{bmatrix}^{\mathsf{T}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_j \text{ , and}$$
$$\hat{\boldsymbol{\pi}}_j = \begin{bmatrix}g^{-1}(\hat{\eta}_{j1}) & g^{-1}(\hat{\eta}_{j2}) & \dots & g^{-1}(\hat{\eta}_{jn})\end{bmatrix}^{\mathsf{T}} = \begin{bmatrix}\hat{\pi}_{j1} & \hat{\pi}_{j2} & \dots & \hat{\pi}_{jn}\end{bmatrix}^{\mathsf{T}} .$$

At this point it was finally possible to apply the GOF tests in order to test the null hypothesis that the misspecified models, M_j, were the true models of the underlying mechanisms that produce the outcomes/response variables y_j.

(v) ***The GOF tests were applied to the fitted incorrect models.***

The seven GOF statistics were calculated for each of the fitted incorrect models obtained in the previous step. For every M_j and test statistic, a p-value was computed (except for a few cases where the computation failed) and the null hypotheses that M_j is an adequate model was rejected for p-values below α. This produced κ separate vectors, $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_\kappa$, with seven elements indicating whether the seven statistics rejected H_0 or not (the value 1 indicated rejection and 0 indicated failure to reject).

(vi) ***Steps (ii), (iv), and (v) were repeated until R=1000 replications had been performed.***

This resulted in R sets of κ vectors, $\boldsymbol{r}_j$, from which we can estimate the power of the GOF tests for the departure type and sample size in question.

(vii) ***The empirical power of the GOF statistics were calculated.***

After R replications, the rejection rate of each statistic was calculated for every misspecified model, M_j. For every model $j = 1, \ldots, \kappa$, the R separate $\boldsymbol{r}_j$ vectors, which were computed during step **(v)**, were summed and consequently divided by R to provide us with the proportion of replications where the statistics rejected M_j. In the cases where M_j had at least one replication where Stukel's score statistic failed to compute, the total number of rejections was divided by the number of successful replications.

4.3 Details of the power study design

This section gives supplemental details to the brief overview given in Section 4.2 about the power study design. Information on the choices of M_{T_j}, $\boldsymbol{\beta}_{T_j}$, M_j, and $\boldsymbol{\beta}_j$ for the departure types are provided. In addition, a figure containing the distributions of both $\boldsymbol{\pi}_{T_j}$ and $\hat{\boldsymbol{\pi}}_j$ ($j = 1, \ldots, \kappa$) from a simulated data set, where $n = 500$, is presented for each departure type. These figures are included to provide an approximate view of how $\boldsymbol{\pi}_{T_j}$ and $\hat{\boldsymbol{\pi}}_j$ were distributed for the replications.

In *D1*, *D2*, and *D3*, defining κ vectors of $\boldsymbol{\beta}_{T_j}$ was done by following an approach similar to that of Hosmer et al. (1997). Starting with $\boldsymbol{\beta}_{T_1}$, and subsequently increasing the size of the element(s) of interest when defining the remaining $\boldsymbol{\beta}_{T_j}$-s (while keeping the rest of the parameters equal for all j), would have undesirable consequences.

In Section 3.4 we found that $\hat{\alpha}$ of one statistic could range from significantly less than α (conservative) to significantly larger than α depending on how the π_i-s were distributed. Since the GOF tests did not perform uniformly, it was important to try to keep $\boldsymbol{\pi}_{T_1}, \ldots, \boldsymbol{\pi}_{T_5}$ as similarly distributed as possible. At the the same time, the parameter(s) which were to be omitted in the misspecified models, had to increase in size for each M_{T_j}, $j = 1, \ldots, \kappa$, otherwise the study would be uninformative.

4.3.1 Departure type *D1*: Omission of a quadratic term

The set-up used to examine the power of the statistics when a quadratic term was omitted from a logistic regression model is from Hosmer et al. (1997). The number of models, κ, was equal to 5. For all the true models, M_{T_j}, the linear predictor was defined as

$$\eta_{ji} = \beta_{j0} + \beta_{j1}x_i + \beta_{j2}x_i^2, \tag{4.1}$$

where x_i was $U(-3,3)$ distributed ($i = 1, \ldots, n$). The simulated observations of x_i were generated using the `runif()` function and were contained in $\boldsymbol{X}_{n\times k+1}$. The $\boldsymbol{\pi}_{T_j}$, which were computed using $\boldsymbol{X}_{n\times 3}\boldsymbol{\beta}_{T_j}$ and the inverse logit link g^{-1}, were used as input for the `rbinom()` function when simulating the response variables $\boldsymbol{y}_j$.

When specifying M_j, the effect of the quadratic term, β_{j2}, was omitted from the linear predictor and the third column of $\boldsymbol{X}_{n\times 3}$ was omitted from $\boldsymbol{X}$. Consequently, for all the misspecified models, the estimated linear predictor was defined as

$$\hat{\eta}_{ji} = \hat{\beta}_{j0} + \hat{\beta}_{j1}x_i, \tag{4.2}$$

just as in Hosmer et al. (1997).

The five $\boldsymbol{\beta}_{T_j}$ were determined by using the same logic as in Hosmer et al. (1997). The resulting $\boldsymbol{\pi}_{T_1}, \ldots, \boldsymbol{\pi}_{T_5}$ had distributions of reasonable similarity. The effect of the quadratic term, β_{j2}, increased in size for each $j = 1, \ldots, 5$. This was achieved by defining the following equations:

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=3} = 0.95 = \pi_{ji}, \tag{4.3}$$

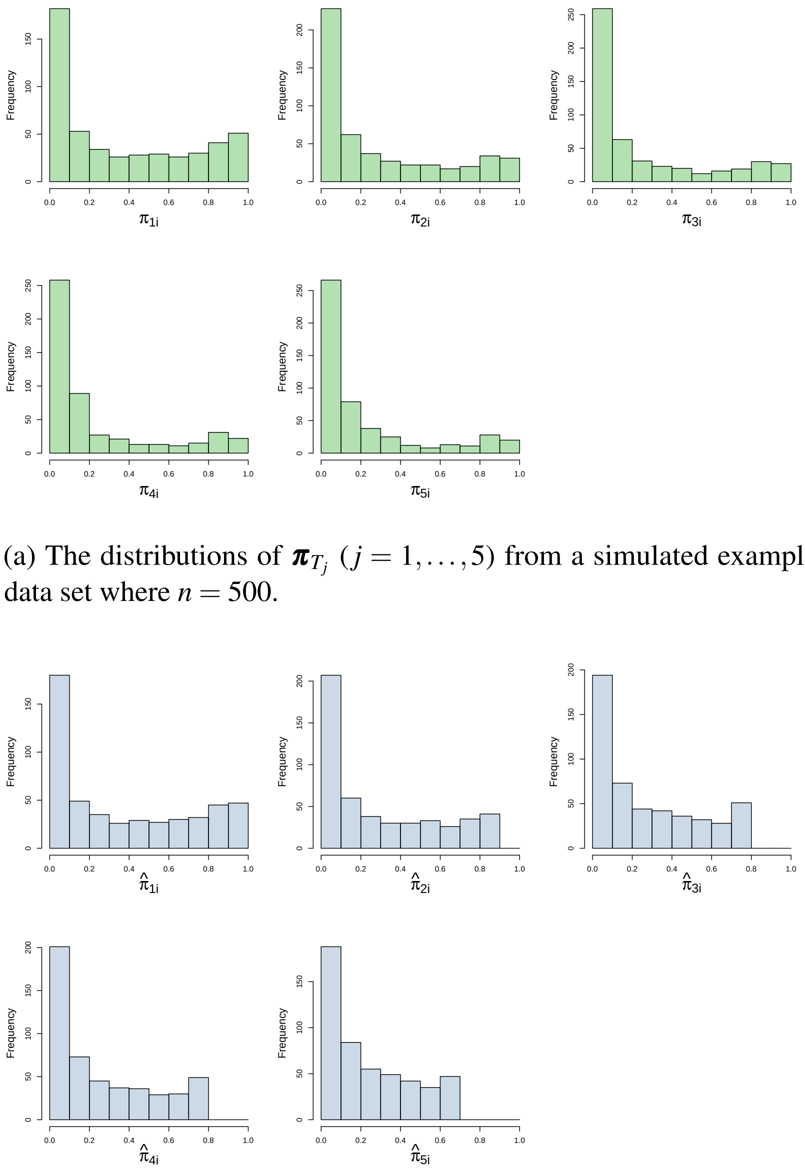
$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=-1.5} = 0.05 = \pi_{ji}\text{ , and} \tag{4.4}$$

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=-3} = I = \pi_{ji}, \tag{4.5}$$

where I equals 0.01, 0.05, 0.1, 0.2, and 0.4 for $M_{T_1}, \ldots, M_{T_5}$, respectively.

For every I, (4.3), (4.4), and (4.5) were solved simultaneously for $\boldsymbol{\beta}_{T_j}$. This produced five $\boldsymbol{\beta}_{T_j}$ (and hence five models) where the quadratic term became increasingly more influential.

The larger the difference between $g^{-1}(-3)$ and $g^{-1}(-1.5)$, the greater the non-linearity. This difference is controlled by I. Equations (4.3) and (4.4), which are held constant for each j, ensure that the distributions are comparable.



(a) The distributions of $\boldsymbol{\pi}_{T_j}$ ($j = 1,\dots,5$) from a simulated example data set where $n = 500$.

(b) The distributions of $\hat{\boldsymbol{\pi}}_j$ ($j = 1,\dots,5$) from a simulated example data set where $n = 500$.

Fig. 4.1 The distributions of $\boldsymbol{\pi}_{T_j}$ and $\hat{\boldsymbol{\pi}}_j$, respectively, for an example data set with departure type *D1* and $n = 500$.

The distributions of the true probabilities in $M_{T_1},\dots,M_{T_5}$ from a simulated example, where $n = 500$, are shown in Figure 4.1a. Starting by choosing $\boldsymbol{\beta}_{T_1}$ and consequently just increasing the value of β_{j2} would not have had the same results. Several experiments with that approach lead to a histogram of π_{5i} which was very dissimilar to the histogram of π_{1i}. Figure 4.1b shows the distributions of the estimated probabilities in $M_1, \ldots, M_5$ from the same example used to generate Figure 4.1a.

4.3.2 Departure type *D2*: Omission of a log term

This particular type of misspecification, *D2*, was not examined by Hosmer et al. (1997), but developed exclusively for this study. In *D2*, there were $\kappa = 6$ logistic regression models. The linear predictor for M_{T_j} was defined as

$$\eta_{ji} = \beta_{j0} + \beta_{j1} x_i + \beta_{j2} \log(x_i), \tag{4.6}$$

where $\beta_{j2} > 0$, and x_i was $U(1,51)$ distributed ($i = 1, \ldots, n$) and generated using the `runif()` function. The continuous explanatory variable x_i was assigned this distribution in an effort to mimic variables one might encounter in practice, such as the SAPS II score in Chapter 6. Just as in Section 4.3.1, $g^{-1}(\eta_{ji}) = e^{\eta_{ji}}/(1+e^{\eta_{ji}}) = \pi_{ij}$, and these π_{ij} were used as parameters by `rbinom()` when generating the y_{ji} for M_{T_j}.

The lack of fit was due to the omission of the effect of the log term. The parameter β_{j2} was not included in linear predictors belonging to $M_1, \ldots, M_6$, and the third column of $\boldsymbol{X}_{n \times k+1}$ was omitted from $\boldsymbol{X}$. As a result, the estimated linear predictor was defined as

$$\hat{\eta}_{ji} = \hat{\beta}_{j0} + \hat{\beta}_{j1} x_i, \tag{4.7}$$

for $j = 1, \ldots, 6$.

The process of defining the vectors $\boldsymbol{\beta}_{T_j}$ was based on the approach described in the previous section. The effect of the log term, β_{j2}, ranged from approximately 0.11 when $j = 1$, to approximately 1.73 when $j = 6$. The value of β_{j1} was kept in the interval $[0.13, 0.27]$ $\forall$ j.

This was achieved by defining the following equations:

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=1} = 0.001 = \pi_{ji}, \tag{4.8}$$

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=51} = 0.999 = \pi_{ji}\text{ , and} \tag{4.9}$$

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=26-J} = 0.5 = \pi_{ji}, \tag{4.10}$$

where J equals 0.5, 1.0, 2.0, 4.0, 6.0, and 10.0 for $M_{T_1},\dots,M_{T_6}$, respectively. Simultaneously solving (4.8), (4.9), and (4.10) for $\boldsymbol{\beta}_{T_j}$, and doing so once for every J, produced six $\boldsymbol{\beta}_{T_j}$ in which β_{j2} becomes increasingly large.

We will now account for the reasoning behind the choice of the equations defined above. The functional form of η_{ji} was an instrumental characteristic in the process. The slope of η_{ji} in this setting,

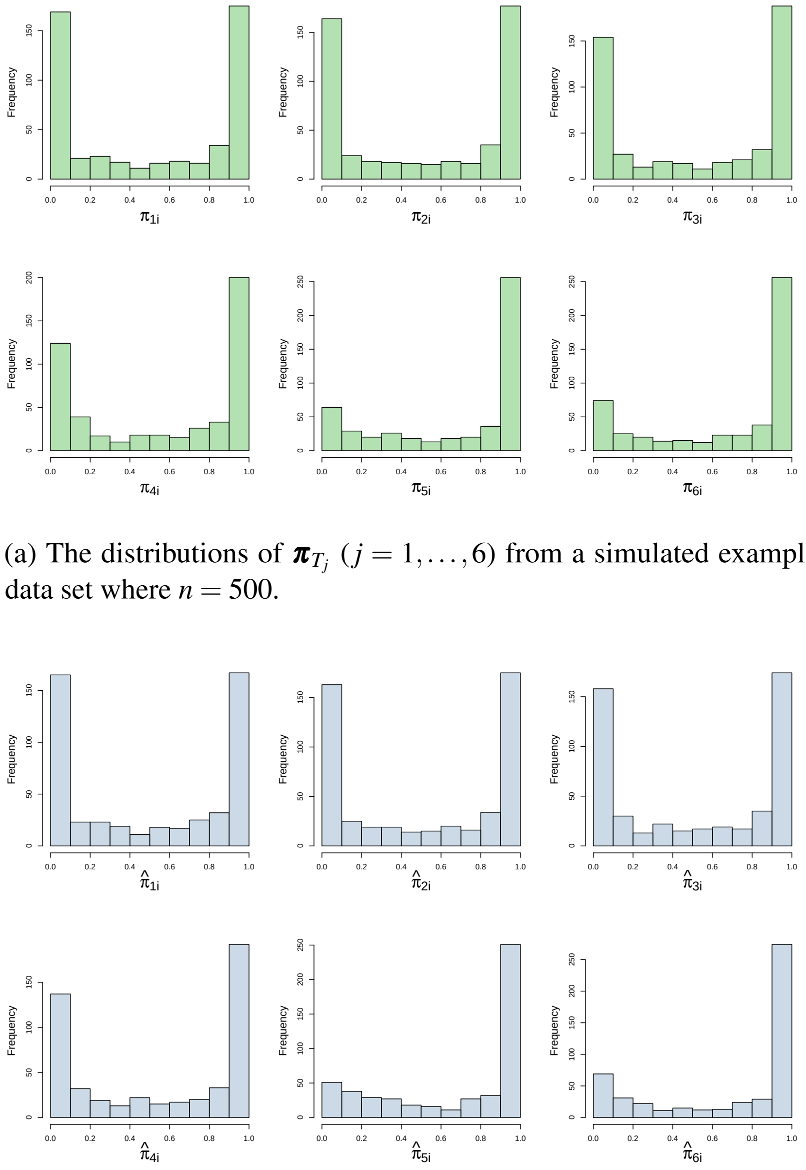
$$\frac{d}{dx_i}\eta_{ji} = \beta_{j1} + \frac{\beta_{j2}}{x_i}, \tag{4.11}$$

decreases monotonically with $x_i \in [1,51]$. Hence η_{ji} increases the most rapidly for small values of x_i (i.e. values such that η_{ji} is less than 0).

If viewed as a function of β_{j2}, $\frac{d}{dx_i}\eta_{ji}$ increases monotonically with $\beta_{j2} > 0$. The larger the β_{j2}, the more rapidly η_{ji} will reach 0. Hence the x_i for which $\eta_{ji} = 0$ should become increasingly closer to 1 for each $j = 1,\dots,6$. For this reason, it was decided that $\eta_{ji}\big|_{x_i=26-J}$ should equal 0 (where $J > 0$), such that an increase in J produces a larger β_{j2}. This was the basis for (4.10). The equations (4.8) and (4.9) were chosen in order to keep the frequency of $\pi_{ji} \leq 0.2$ and frequency of $\pi_{ji} \geq 0.8$ somewhat constant for each J (or M_{T_j}).

The distributions of the true probabilities in $M_{T_1},\dots,M_{T_6}$ from a simulated example, where $n = 500$, are shown in Figure 4.2a. The distributions of the resulting $\boldsymbol{\pi}_{T_1},\dots,\boldsymbol{\pi}_{T_5}$ were reasonably similar, but the frequency of large π_{ji} did become substantially higher for every model $j = 1,\dots,6$. Figure 4.2b shows the distributions of the estimated probabilities in $M_1,\dots,M_5$ using the aforementioned example. The $\hat{\pi}_{ji}$ appear to be distributed quite similarly to π_{ji}, except for models M_{T_4} and M_4, where $\hat{\pi}_{ji}$ has slightly more observations in the interval $[0,0.1]$ than that of π_{ji}.

When inspecting the distributions of $\hat{\boldsymbol{\pi}}_j$, it becomes evident that $\hat{\boldsymbol{\pi}}_1$ bears resemblance to the true probabilities in Situation 1 in the significance level study (see Figure 4.3b). This is also the case for $\hat{\boldsymbol{\pi}}_2$, $\hat{\boldsymbol{\pi}}_3$, and $\hat{\boldsymbol{\pi}}_4$. One could also argue that among the situations covered in the significance level study, Situation 7 generates the probability distribution most similar to that of $\hat{\boldsymbol{\pi}}_5$ and $\hat{\boldsymbol{\pi}}_6$. In Section 3.4, we saw that several statistics had rather dissimilar $\hat{\alpha}_1$ and $\hat{\alpha}_7$, especially for sample sizes $n = 100$ and $n = 500$. It is therefore expected that the results of the power study might show interesting patterns for these statistics.



(a) The distributions of $\boldsymbol{\pi}_{T_j}$ $(j = 1,\ldots,6)$ from a simulated example data set where $n = 500$.

(b) The distributions of $\hat{\boldsymbol{\pi}}_j$ $(j = 1,\ldots,6)$ from a simulated example data set where $n = 500$.

Fig. 4.2 The distributions of $\boldsymbol{\pi}_{T_j}$ and $\hat{\boldsymbol{\pi}}_j$, respectively, for an example data set with departure type *D2* and $n = 500$.

4.3.3 Departure type *D3*: Omission of the main effect of a binary covariate and its interaction with a continuous covariate.

The chosen set-up for *D3* is more or less identical to the one proposed by Hosmer et al. (1997) to examine this type of misspecification. However, an additional model (M_{T_5}) was added to the four models which were detailed in Hosmer et al. (1997). Hence there are $\kappa = 5$ models in this set-up. The linear predictor in M_{T_j} was defined as

$$\eta_{ji} = \beta_{j0} + \beta_{j1}x_i + \beta_{j2}d_i + \beta_{j3}x_i d_i, \tag{4.12}$$

where x_i was $U(-3,3)$ distributed (as in *D1*) and d had the $Bernoulli(1/2)$ distribution $(i = 1,\ldots,n)$. When generating observations of x_i and d, the `runif()` and `rbinom()` functions were used, respectively. These two covariates were independent of each other. The `set.seed()` function was used directly before the random generation of each covariate, and they both had their own unique numeric $n \times 1$ vector which served as input for the preceding `set.seed()`.

The simulated observations of x_i and d, as well as their interaction xd, were contained in $\boldsymbol{X}_{n\times 4}$. Just as for *D1*, $\boldsymbol{y}_j$ was generated by the `rbinom()` function with $\boldsymbol{\pi}_{T_j}$ as input, where $\boldsymbol{\pi}_{T_j}$ was computed using $\boldsymbol{X}_{n\times 4}\boldsymbol{\beta}_{T_j}$ and the inverse log link. The main effect of d, β_{j2}, and the effect of its interaction with x_i, β_{j3}, were omitted when specifying M_j. Consequently, for all misspecified models M_j, the estimated linear predictor was defined as

$$\hat{\eta}_{ji} = \hat{\beta}_{j0} + \hat{\beta}_{j1}x_i, \tag{4.13}$$

and the third and fourth column of $\boldsymbol{X}_{n\times 4}$ were omitted from $\boldsymbol{X}$.

The $\boldsymbol{\beta}_{T_j}$ were determined using the same approach as described in Section 4.3.1. The resulting $\boldsymbol{\pi}_{T_j}$ had distributions which were similar to a certain extent (see example in Figure 4.3). There is a considerable difference between M_1 and M_4 in regards to the amount of observations i where $\pi_{ji} > 0.5$. This should be kept in mind when analysing the results of the power study.

The main effect of the binary variable, β_{j2}, increased from approximately 0.27 up to approximately 2.17 for $j = 1,\ldots,5$, whereas β_{j3} increased from approximately 0.090 to approximately 0.72. This was achieved by defining the following equations:

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=-3,\, d_i=0} = 0.1 = \pi_{ji}, \tag{4.14}$$

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=-3,\, d_i=1} = 0.1 = \pi_{ji}, \tag{4.15}$$

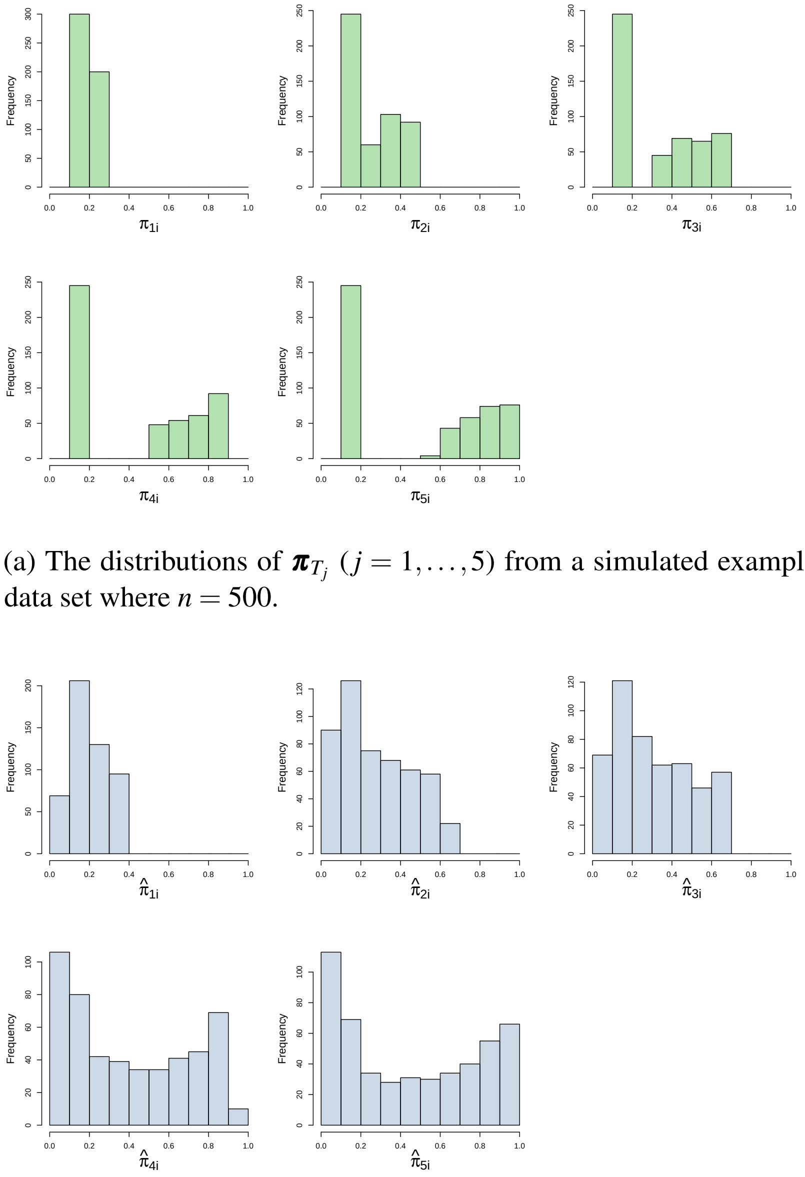
$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=3,\, d_i=0} = 0.2 = \pi_{ji} \text{ , and} \tag{4.16}$$

$$g^{-1}\Big\{\eta_{ji}(x_i)\Big\}\Big|_{x_i=3,\, d_i=1} = 0.2 + K = \pi_{ji}, \tag{4.17}$$

where K equals respectively 0.1, 0.3, 0.5, 0.7, and 0.75 for $M_{T_1},\ldots,M_{T_5}$. The larger the value of π_{ji} in (4.17) is compared to that of (4.16), which is controlled by K, the greater β_{j2} and β_{j3} must be. For every K, (4.14), (4.15), (4.16), and (4.17) were solved simultaneously for $\boldsymbol{\beta}_{T_j}$, producing five $\boldsymbol{\beta}_{T_j}$ where β_{j2} and β_{j3} became larger for every j.

The distributions of the true probabilities in $M_{T_1},\ldots,M_{T_5}$ from a simulated example, where $n = 500$, are shown in Figure 4.3a. Equations (4.14), (4.15), and (4.16) ensure that the distribution are reasonably similar. Figure 4.3b shows the distributions of the estimated probabilities in $M_1,\ldots,M_5$ from the same example used to generate Figure 4.1a. There is a large difference between the distributions of $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_5$.

It is expected that the difference in how the estimated probabilities are distributed will cause some of the statistics to perform differently as K increases, and not just as a result of the increasing lack of fit. When comparing the histograms in Figure 4.3b to those in Figure 3.3b, which belong to the situations used in the significance level study, $\hat{\boldsymbol{\pi}}_1$ does not appear to be comparable to any of the situations presented. For $\hat{\boldsymbol{\pi}}_2$ and $\hat{\boldsymbol{\pi}}_3$, on the other hand, their distributions bear resemblance to that of Situation 11. Situation 8 generates a distribution of probabilities which is fairly similar to that of $\hat{\boldsymbol{\pi}}_4$ and $\hat{\boldsymbol{\pi}}_5$. Therefore, the statistics that perform very differently in Situation 8 compared to Situation 11 (such as X^2_{st}), might display interesting behaviours as the lack of fit increases.



(a) The distributions of $\boldsymbol{\pi}_{T_j}$ $(j = 1, \ldots, 5)$ from a simulated example data set where $n = 500$.

(b) The distributions of $\hat{\boldsymbol{\pi}}_j$ $(j = 1, \ldots, 5)$ from a simulated example data set where $n = 500$.

Fig. 4.3 The distributions of $\boldsymbol{\pi}_{T_j}$ and $\hat{\boldsymbol{\pi}}_j$, respectively, for an example data set with departure type *D3* and $n = 500$.

4.3.4 Departure type *D4*: Selection of an incorrect link function.

When evaluating the GOF tests' ability to recognise lack of fit, the selection of an incorrect link function (*D4*) was included as a type of departure from the true model. The set-up for *D4* was based on the procedure for evaluating a test's power to discern a misspecified link function in Hosmer et al. (1997). This study specified one additional model, however, making the number of models, κ, equal to 6.

In Hosmer et al. (1997), Stukel's generalised model from Section 2.3 was used to generate five different $\boldsymbol{\pi}_{T_j}$ belonging to five models with differing link functions g_j. In this study, we chose another approach when generating $\boldsymbol{\pi}_{T_1}$ and $\boldsymbol{\pi}_{T_2}$. The linear predictor was identical, however, and defined in all six models as

$$\eta_{ji} = 0.8x_i, \tag{4.18}$$

where x_i is $U(-3,3)$. All six $\mathbf{y}_j$ were generated by the `rbinom()` function, with their respective $\boldsymbol{\pi}_{T_j}$ as parameters.

In M_{T_1}, the true link function g_1 is the probit link, hence

$$g_1^{-1}(\eta_{1i}) = \Phi(\eta_{1i}) = \pi_{1i}. \tag{4.19}$$

The `pnorm()` function was used to generate the elements of $\boldsymbol{\pi}_{T_1}$, whereas Hosmer et al. (1997) used Stukel's generalised logistic model to produce π_{1i} in compliance with the probit model (i.e. $\pi_{\boldsymbol{\varphi}}(\eta_{1i})$ where Stukel's shape parameters $\boldsymbol{\varphi} = (0.165, 0.165)$; see Section 2.3). Due to advances made in computational fields of study since 1997, the `pnorm()` function was considered a better alternative than the approach described in Hosmer et al. (1997).

In M_{T_2}, the true link function g_2 is the complementary log-log link, hence

$$g_2^{-1}(\eta_{2i}) = 1 - e^{-e^{\eta_{2i}}} = \pi_{2i}. \tag{4.20}$$

The expression for $g_2^{-1}(\eta_{2i})$ was implemented directly in `R`, instead of computing $\boldsymbol{\pi}_{T_1}$ by using $\pi_{\boldsymbol{\varphi}}(\eta_{2i})$ where $\boldsymbol{\varphi} = (0.62, -0.037)$ as done by Hosmer et al. (1997).

In the remaining four true models, however, it was necessary to use Stukel's generalised logistic model to generate $\boldsymbol{\pi}_{T_3}, \ldots, \boldsymbol{\pi}_{T_6}$ since they were not based on well known tolerance distributions. $M_{T_3}, \ldots, M_{T_5}$ are from Hosmer et al. (1997), whereas M_{T_6} was added specifically for this study.

In M_{T_3}, we wanted the tails of the sigmoid mean function, $g_3^{-1}(\eta_{3i})$, to be longer than the standard logistic function ${}^{e^{\eta_i}}\!/\!{}_{(1+e^{\eta_i})}$. This was achieved by first computing $h_{\boldsymbol{\varphi}}(\eta_{3i})$, where $\boldsymbol{\varphi} = (-1.0, -1.0)$, followed by $\pi_{3i} = \pi_{\boldsymbol{\varphi}}(\eta_{3i}) = {}^{e^{h_{\boldsymbol{\varphi}}(\eta_{3i})}}\!/\!{}_{\left(1+e^{h_{\boldsymbol{\varphi}}(\eta_{3i})}\right)}$. Contrastingly, we wanted the mean function, $g_1^{-1}(\eta_{4i})$, of M_{T_4} to have shorter tails than the standard logistic function. This was achieved by first computing $h_{\boldsymbol{\varphi}}(\eta_{4i})$, where $\boldsymbol{\varphi} = (1.0, 1.0)$, followed by

$\pi_{4i} = \pi_{\boldsymbol{\varphi}}(\eta_{4i}) = {}^{e^{h_{\boldsymbol{\varphi}}(\eta_{4i})}}\!/\!\left({}_{1+e^{h_{\boldsymbol{\varphi}}(\eta_{4i})}}\right)$.

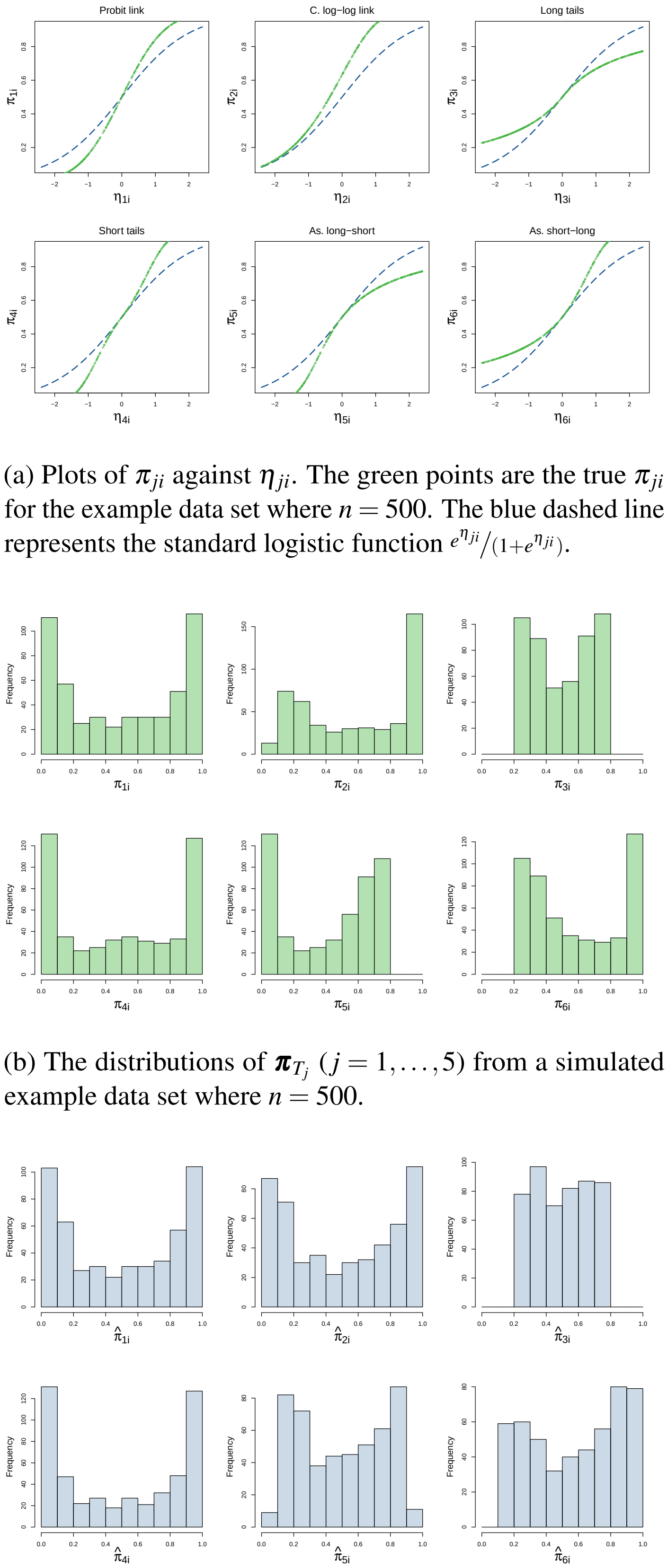
In M_{T_5} and M_{T_6}, the models had asymmetrical mean functions, where one tail was shorter than the other (recall from Section 2.3 that the standard logistic function is symmetric about $\eta_i = 0$). The $\boldsymbol{\pi}_{T_5}$ in M_{T_5} were computed in the same way as for M_{T_3} and M_{T_4}, but with $\boldsymbol{\varphi} = (-1.0, 1.0)$. As mentioned in Section 2.3, φ_1 controls the behaviour of the upper tail, and φ_2 controls the behaviour of the lower tail. When $\boldsymbol{\varphi} = (-1.0, 1.0)$, the mean function $g_5^{-1}(\eta_{5i})$ is asymmetrical with a long upper tail and short lower tail.

The model M_{T_6}, which has the opposite tail heaviness of M_{T_5}, was added to the study to examine whether the statistics performed differently when the upper and lower tails' heaviness was reversed. In this text, the term opposite tail heaviness refers when one model uses $\boldsymbol{\varphi} = (c_1, c_2)$ and another uses $\boldsymbol{\varphi} = (c_2, c_1)$ when $c_1 c_2 < 0$, $c_1, c_2 \in \mathbb{R}$. Hence $\boldsymbol{\pi}_{T_6}$ was computed using $\boldsymbol{\varphi} = (1.0, -1.0)$.

Figure 4.4a compares the true probability curves $\pi_{ji}(\eta_{ji})$ to the standard logistic function ${}^{e^{\eta_{ji}}}\!/_{(1+e^{\eta_{ji}})}$, which is what the probability curves would have looked like if the logit link were the correct link function. All the plots in Figure 4.4 are based on the exact same simulated example data set where $n = 500$. The distributions of the true probabilities in $M_{T_1}, \ldots, M_{T_6}$ are shown in Figure 4.4b. This figure shows a wide variety of distributions, which is not unexpected due to variety of the M_{T_j}. There is also a range of different distributions in Figure 4.4c, which presents the distributions of the estimated probabilities in $M_1, \ldots, M_6$ for the example data set.

Figure 4.4c contains plots which are reminiscent multiple histograms in Figure 3.3b. The example distributions of $\hat{\boldsymbol{\pi}}_1$, $\hat{\boldsymbol{\pi}}_2$, and $\hat{\boldsymbol{\pi}}_4$, are similar to the distribution presented for Situation 2. The plot of the distribution corresponding to Situation 4 is similar to that of $\hat{\boldsymbol{\pi}}_3$. Thus it is expected that X^2_{st} will display very high percentages of rejection for model M_{T_3} due to its problematically large $\hat{\alpha}_4$ in the significance level study.

In contrast, $\hat{\boldsymbol{\pi}}_5$ and $\hat{\boldsymbol{\pi}}_6$ are not as straightforward to compare to the histograms in Figure 3.3b. The example histogram of $\hat{\boldsymbol{\pi}}_5$ is somewhat similar to that of Situation 2, but not to the same degree as $\hat{\boldsymbol{\pi}}_1$, $\hat{\boldsymbol{\pi}}_2$, and $\hat{\boldsymbol{\pi}}_4$. Moreover, the histogram of $\hat{\boldsymbol{\pi}}_6$ does not have a clear candidate in terms of similarity, but its minor left skewness and lack of $\hat{\pi}_{6i}$ close to 0 is worth noting.



(a) Plots of π_{ji} against η_{ji}. The green points are the true π_{ji} for the example data set where $n = 500$. The blue dashed line represents the standard logistic function $e^{\eta_{ji}}/(1+e^{\eta_{ji}})$.

(b) The distributions of $\boldsymbol{\pi}_{T_j}$ $(j = 1, \ldots, 5)$ from a simulated example data set where $n = 500$.

(c) The distributions of $\hat{\boldsymbol{\pi}}_j$ $(j = 1, \ldots, 5)$ from a simulated example data set where $n = 500$.

Fig. 4.4 The distributions of $\boldsymbol{\pi}_{T_j}$ and $\hat{\boldsymbol{\pi}}_j$, respectively, for an example data set with departure type *D4* and $n = 500$.

4.4 The Results of the Power Study

The percentages of replications where the false H_0 was rejected are presented in Table 4.1. Results for each type of departure from the true model is presented in the following sections.

As in the significance level study, there were settings with replications where the computation of Stukel's score test statistic failed. The results from these settings were included if they had less than 25% failed replications. If there was more than 25% computation failure, the resulting empirical power was omitted. The results in Table 4.1 marked with one asterisk belong to settings with failed computations of Stukel's score test statistic, but in no more than 15% of the 1000 replications. Two asterisks indicate over 15% failed replications.

4.4.1 Power assessment results for departure type *D1*

The results from the simulations to assess the statistics' power to detect a lack of fit due to departure type *D1* is listed in Table 4.1a. All seven tests managed rather well to detect that a quadratic term was missing from M_j, $j = 1, \ldots, 5$. The estimated power improves drastically when I is increased from 0.01 to 0.05.

The distributions of $\hat{\boldsymbol{\pi}}_j$ were right skewed. In terms of distributions of probabilities, the most similar situation in the significance level study is Situation 6 (see Figure 3.3b). As I approaches 0.4, the distributions start to look more like that of Situation 11. Therefore, when examining whether the estimated significance levels of the statistics may offer explanations as to the estimated power results, we will refer to the statistics' $\hat{\alpha}_6$ and $\hat{\alpha}_{11}$ from Table 3.2. It should be kept in mind, however, that the distribution of true probabilities produced by Situation 6 and 11, and the distributions of $\hat{\boldsymbol{\pi}}_j$ are not identical.

The statistic which outperformed the others most frequently was $\hat{S}_{st}$. This is promising for the USS test, especially when considering that in Situation 6, its $\hat{\alpha}_6$ were either close to α or slightly less than α. $IMT2$ stood out by having the lowest power in all situations where the all seven statistics did not perform identically. In most of the situations, it was outperformed by a noticeable amount.

Table 4.1 Simulated percent rejection at $\alpha = 0.05$ using sample sizes 100, 500, and 1000, with 1000 replications. The entries are the percentages of replications where the statistic rejected the fit of the misspecified model M_j with one of the four departure types.

(a) *D1*: *Omission of a quadratic term*

Statistic/sample size	Correct model														
	$M_{T_1}, I = 0.01$			$M_{T_2}, I = 0.05$			$M_{T_3}, I = 0.1$			$M_{T_4}, I = 0.2$			$M_{T_5}, I = 0.4$		
	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000
St. Pearson, X^2_{st}	6.2	8.5	10.1	36.2	86.7	98.8	60.8	99.4	100	84.3	100	100	96.9	100	100
USS, $\hat{S}_{st}$	4.7	6.5	10	35.3	89.6	99.3	62.8	99.9	100	87	100	100	98.5	100	100
Stukel's score	5.5	6.9	9.4	30.0*	90	99	56.6*	99.8	100	82.1*	100	100	96.8**	100*	100*
Stukel's LRT1	7	6.1	8.9	26.4	87.4	98.9	51.3	99.6	100	79.6	100	100	96.7	100	100
Stukel's LRT2	7	6.1	8.9	26.4	87.4	98.9	51.4	99.6	100	80.1	100	100	96.7	100	100
IM1	7	7.2	8	32.2	88.5	99	56.5	99.8	100	83	100	100	96.7	100	100
IM2	4.3	5.5	7.1	12.9	76.2	98.7	28.7	98.8	100	56.4	100	100	88.9	100	100

(b) *D2*: *Omission of a* log *term*

Statistic/sample size	Correct model																	
	M_{T_1}, J = 0.5			M_{T_2}, J = 1.0			M_{T_3}, J = 2.0			M_{T_4}, J = 4.0			M_{T_5}, J = 6.0			M_{T_6}, J = 10.0		
	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000
St. Pearson, X^2_{st}	2.8	6.2	6.4	2.7	6.0	6.0	3.1	5.8	6.3	3.2	6.5	8.2	3.8	8.8	10.8	4.6	12.1	19.0
USS, $\hat{S}_{st}$	3.6	5.7	4.7	3.9	5.2	4.7	3.9	4.9	4.3	4.8	4.8	5.5	5.9	5.1	6.4	8.2	13.3	16.9
Stukel's score	5.1	6.4	5.0	4.1	5.6	5.9	4.2	4.6	4.9	6.2	6.0	8.9	6.1	9.6	15.6	7.2	26.2	48.2
Stukel's LRT1	10.3	6.6	5.9	9.2	6.0	6.5	9.7	6.7	6.5	9.9	9.4	10.5	10.5	11.4	16.7	11.4	27.4	49.2
Stukel's LRT2	10.3	6.6	5.9	9.2	6.0	6.5	9.7	6.7	6.5	9.9	9.4	10.5	10.5	11.4	16.7	11.4	27.4	49.2
IM1	4.6	5.7	5.5	4.6	5.8	5.7	4.7	6.4	5.9	5.0	7.9	9.8	6.0	11.9	15.8	7.6	25.6	45.0
IM2	5.8	4.2	3.7	4.7	4.3	5.5	6.7	5.3	5.8	5.6	7.1	7.4	5.7	7.6	10.2	6.1	13.3	26.3

(c) *D3*: *Omission of the main effect of a binary variable and its interaction term.*

Statistic/sample size	Correct model														
	M_{T_1}, K = 0.1			M_{T_2}, K = 0.3			M_{T_3}, K = 0.5			M_{T_4}, K = 0.7			M_{T_5}, K = 0.75		
	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000
St. Pearson, X^2_{st}	56.8	35.5	18.8	46	16.7	8.1	32.8	7.1	4.8	25.7	6	8.7	25.8	8.2	19.5
USS, $\hat{S}_{st}$	5.3	6	4.8	5.2	5	4.7	5.3	5.7	5.5	6.3	10.7	16.8	8	20.5	36.8
Stukel's score	-	-	-	-	-	-	-	-	-	5.5**	9.3*	15.1*	7.9*	18.9*	35
Stukel's LRT1	5.5	5.8	4.8	5.7	4.9	4.6	7	5.3	5.6	7.6	9.8	15.7	10	18.9	35.6
Stukel's LRT2	5.5	5.8	4.8	5.6	4.9	4.6	6.5	5.5	5.5	7.3	10.3	16.7	9.8	19	35.6
IM1	4.3	4.7	4.6	5	4.8	4.4	4.8	5.1	5.7	6.2	7.5	12.7	7.2	15.4	29
IM2	4.5	4.7	4.6	5	4.7	4.5	5.1	5	5.7	6.4	8.1	13.4	7.6	16.5	30

(d) *D4*: *Specification of an incorrect link function.*

Statistic/sample size	Correct model																	
	M_{T_1}, probit link			M_{T_2}, c. log-log link			M_{T_3}, long tails			M_{T_4}, short tails			M_{T_5}, as. long-short tails			M_{T_6}, as. short-long tails		
	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000	100	500	1000
St. Pearson, X^2_{st}	2.2	5.1	11.8	9.8	11.8	15.5	73.8	60.4	60.6	0.5	46.5	95.3	21.8	25	35.6	21.6	27.3	36.4
USS, $\hat{S}_{st}$	5.5	10.1	15.7	6.6	9.3	11.7	8.2	17.7	26.4	14.1	79.9	98.4	8.7	16.1	25.8	8.4	16.9	26.5
Stukel's score	3	6.7	10.7	49.5	100	100	6.7	14.4	21.1	2.4	56.6	94.4	69	100	100	68.6	100	100
Stukel's LRT1	11	10.4	15.1	18	68.8	96.5	6.1	12.8	19.9	21.5	75.8	97	52.9	99.8	100	51.2	99.9	100
Stukel's LRT2	11	10.4	15.1	18	68.8	96.5	6.1	12.8	19.9	21.5	75.8	97	52.9	99.8	100	51.2	99.9	100
IM1	4.6	9.8	15.9	6.3	9.3	11.9	8	17.4	26.4	9.7	79.3	98.5	8.3	16.6	26.1	8.2	17.4	27.4
IM2	8.6	11.9	17.8	6.7	8.3	11.2	7.6	17.4	26.4	24.3	86.1	99	7.6	15.8	25.7	7.8	15.9	27

When $I = 0.01$, the two Stukel's LRT statistics and $IMT1$ had the highest power for the smallest sample sizes. These three statistics have $\hat{\alpha}_6$ which are greater than 6.0%. This may account for their higher rejection percentage. The $IMT2$ statistic, and X^2_{st}, also had empirical significance levels greater than 6.0%. Among the statistics which did not have an $\hat{\alpha}_6 > \alpha$ when $n = 100$, Stukel's score statistic was the one with the highest power.

The $IMT1$ has considerably higher power than $IMT2$ in all situations. The IMT statistics had identical $\hat{\alpha}_6$ for all sample sizes, hence the difference in their estimated power is not due to dissimilar significance levels. It should be kept in mind, however, that $IMT2$ has a smaller $\hat{\alpha}_s$ than $IMT1$ in most of the situations where $n = 100$. Still there is strong evidence of $IMT1$ being a more powerful test statistic as its simulated percentage of rejection is consistently higher and by a considerable amount in many cases.

When $I \geq 0.2$, every statistic has an estimated power of 100% when $n = 500$ and $n = 1000$. However, when $n = 100$, there is a variety of results. When $I = 0.4$, $IMT2$ is the only statistic with a rejection percentage below 90%, whereas the remaining statistics reject H_0 in at least 96.7% of the replications. When $I = 0.2$, it is also $IMT2$ that has the poorest power, close to 50%, whereas the remaining statistics reject in close to 80% of the time.

Stukel's LRT1 and LRT2 performed identically in all but two settings. The LRT2 had a minutely higher rejection rate than LRT1 for M_{T_4} and M_{T_5} when $n = 100$, despite the fact that LRT1 had $\hat{\alpha}_{11} = 6.9$ compared to LRT2 which had $\hat{\alpha}_{11} = 6.8$. The results in *D1* do not give strong evidence for favouring the modified LRT2 over the original LRT1, but it certainly does not discount the potential of the modified version.

The standardised Pearson test performed well compared to the other tests when $n = 500$ and $n = 1000$ in M_{T_1}. This is promising in terms of its power, since X^2_{st} had conservative rejection rates in Situation 6 of the significance level study for those sample sizes. In addition, X^2_{st} had the highest estimated power when $n = 100$ in M_{T_2}. In this setting, however, it only outperforms $\hat{S}_{st}$ by a small amount. In addition, $\hat{S}_{st}$ was conservative in Situation 6 in the significance level study, whereas X^2_{st} had a larger rejection region than desired.

4.4.2 Power assessment results for departure type *D2*

The results from the simulations to assess the statistics' power to detect a lack of fit due to departure type *D2* is listed in Table 4.1b. None of the test statistics had an estimated power exceeding 50%. The percentages of replications where the statistics rejected the false H_0 were disappointingly low.

As mentioned in Section 4.3.2, the distributions of $\hat{\boldsymbol{\pi}}_1,\ldots,\hat{\boldsymbol{\pi}}_4$ are similar to that of the true probabilities in Situation 1 in the significance level study, and $\hat{\boldsymbol{\pi}}_5$ and $\hat{\boldsymbol{\pi}}_6$ bears resemblance to Situation 7. Hence when examining whether the estimated significance levels of the statistics may provide insights about the estimated power results, we will refer to the $\hat{\alpha}_1$ and $\hat{\alpha}_7$ from Table 3.2.

Stukel's LRT1 and LRT2 had the highest power in all the situations, except in M_{T_5} when $n = 500$ where $IMT1$ had a slightly higher rate of rejection. The power of the LRT2 statistic was identical to that of LRT1. In Situation 1, these two statistics performed identically in terms of empirical significance levels. In Situation 7, however, LRT2 was considerably less anti-conservative than LRT1 for sample sizes of 100 and 1000.

The statistics which most frequently had the lowest power in a situation was X^2_{st} and $\hat{S}_{st}$. When $n = 100$, the power of X^2_{st} was below the nominal significance level α for every $M_{T_j}, j = 1,\ldots,6$. This poor performance does not appear to be due to the empirical significance level of the statistic in Situation 1 and 7, which were both greater than α when $n = 100$.

The USS test also resulted in percentages of rejection below 5%. This occurred when evaluating M_1, M_2, M_3, and M_4, and does not appear to be explained by its values of $\hat{\alpha}_1$. The USS test had $\hat{\alpha}_1$ very close to α, and $\hat{\alpha}_1 > \alpha$ when $n = 1000$. Furthermore, the models for which $\hat{S}_{st}$ had power above the nominal α for all three sample sizes are M_{T_5} and M_{T_6}. However, in Situation 7 we saw that $\hat{S}_{st}$ was conservative when $n = 100$ and 500. This implies that the poor power of the USS test for *D2* is not a consequence of its empirical rejection region.

The standardised Pearson test and USS test were not alone in having a limited ability to detect that a log term was missing from M_j. The IMT statistics and Stukel's score test statistic did not perform well either. In almost every situation, $IMT1$ was more powerful than $IMT2$. When $n = 100$ for $M_{T_1},\ldots,M_{T_4}$, however, $IMT2$ was more powerful than $IMT1$, despite $IMT2$ being more conservative in Situation 1 in the significance level study. However, $IMT2$ had only slightly higher percentages of rejections , thus it is not necessarily true that it is better at detecting lack of fit than $IMT1$ in small sample cases where the effect of the log term is less pronounced.

For this particular type of departure, even larger sample sizes in the simulations are necessary for observing higher power. Additional simulations where $n = 10,000$ showed that a power of 100% was only achieved by the three Stukel's statistics and the IMT statistics for M_{T_6} when $J = 10.0$. The remaining two statistics, X^2_{st} and $\hat{S}_{st}$, rejected M_6 in 73% and 63% of the replications, respectively.

The additional simulations also revealed that when $n = 10,000$, the three Stukel's statistics' power was over 50% for M_{T_4}, and over 90% for M_{T_5}. The IMT statistics rejected the false H_0 in over 40% of the replications for M_{T_4} and over 80% of the replications for M_{T_5}. Overall, the standardised Pearson test performed better than the USS test when $n = 10,000$, but both of them had markedly poorer power than the remaining five test statistics. In the additional simulations, the distributions of $\hat{\boldsymbol{\pi}}_4$ and $\hat{\boldsymbol{\pi}}_5$ were more similar to that of Situation 1, as opposed to Situation 7.

4.4.3 Power assessment results for departure type *D3*

The results from the simulations to assess the statistics' power to recognise a misspecified model due to departure type *D3* is listed in Table 4.1c. The results for the additional model M_{T_5}, where $K = 0.75$, showed that the power of $\hat{S}_{st}$, Stukel's score statistic, Stukel's LRT1 and LRT2 statistics were close to 35% when $n = 1000$.

As mentioned in Section 4.3.3, the distributions of $\hat{\boldsymbol{\pi}}_2$ and $\hat{\boldsymbol{\pi}}_3$ are similar to that of the true probabilities in Situation 11 in the significance level study, and $\hat{\boldsymbol{\pi}}_4$ and $\hat{\boldsymbol{\pi}}_5$ bear resemblance to the histogram belonging to Situation 8 in Figure 3.3b. When evaluating whether the estimated significance levels of the statistics may account for certain characteristics displayed in Table 4.1c, we will refer to the $\hat{\alpha}_{11}$ and $\hat{\alpha}_8$ from Table 3.2.

The statistic with the highest estimated power in most situations is X^2_{st}. It appears to perform dramatically better than the six other statistics for all sample sizes when $K = 0.1$ and 0.3, and for the two smallest sample sizes when $K = 0.5$. For M_{T_1}, M_{T_2}, and M_{T_3}, the estimated power of X^2_{st} decreases as n increases. This is most likely a consequence of the rejection region of X^2_{st}, i.e. its empirical significance level which we assume is comparable to $\hat{\alpha}_{11}$ in these cases.

The $\hat{\alpha}_{11} \times 100$ values belonging to the standardised Pearson test statistic in Table 3.2, are respectively 16.4, 5.0, and 6.4 for sample sizes 100, 500, and 1000. In Section 3.3, we saw that X^2_{st} was very unstable for estimated probabilities with distributions where the extremities of one or both tails had either very few or no observed values. This was the case in Situation 4, 9, 10, 11, and 12 for multiple example data sets. Thus the behaviour of X^2_{st} is not surprising.

The statistic which most frequently had the lowest power was $IMT1$, followed closely by $IMT2$ which performed slightly better than $IMT1$ when the omitted covariate effects were the most substantial. In Table 4.1c, $IMT2$ has a higher rejection rate than $IMT1$ in 9 out of 15 situations despite having a smaller rejection region than $IMT2$ in Situation 8 (for all sample sizes) and in Situation 11 for sample sizes 100 and 500.

The computation of Stukel's score test statistic failed in more than 25% of the replications for the data simulated for M_{T_1}, M_{T_2}, and M_{T_3}. This was not surprising as the observed $\hat{\boldsymbol{\pi}}_1$, $\hat{\boldsymbol{\pi}}_2$, and $\hat{\boldsymbol{\pi}}_3$ were highly skewed, and the statistic failed to compute in some of the replications for Situation 11. For the two true models where the estimated power of Stukel's score test statistic was included in the results, its performance was comparable to the USS test.

Observations made in Section 3.3, indicate that Stukel's LRT1 and LRT2 behave slightly differently when the estimated probabilities have highly skewed distributions (like the distributions produced by Situation 6 and 7 in Figure 3.3b), especially when $n = 100$. The LRT2 is more powerful than LRT1 in almost all situations where $n = 100$. This may be explained by the fact that the LRT1 was considerably more anti-conservative than LRT2 in almost all of the settings where their empirical significance levels were not identical.

The USS test was very similar to Stukel's LRT statistics for the models where the estimated probabilities were comparable to that of Situation 11 (M_{T_2} and M_{T_3}). Interestingly, the USS test was mostly conservative in Situation 11, and had $\hat{\alpha}_{11}$ which were smaller than that of Stukel's LRT1 and LRT2 statistics. For the models we compare to Situation 8 (M_{T_4} and M_{T_5}), $\hat{S}_{st}$ had lower power than the LRT statistics when $n = 100$. This coincided with settings where the $\hat{\alpha}_8$ of Stukel's LRT statistics were considerably more anti-conservative. When $n = 500$ and 1000, $\hat{S}_{st}$ performed slightly better. This is not surprising if one considers that the USS test had somewhat larger rejection regions in Situation 8 for these sample sizes.

4.4.4 Power assessment results for departure type *D4*

The results from the simulations to assess the statistics' power to detect a lack of fit due to departure type *D4* is listed in Table 4.1d.

The probit model

The tests had poor power when the correct model was M_{T_1}. This is not surprising considering the similar symmetrical s-shape of the mean function of the probit model, which is relatively similar to the mean function of the logistic regression model. The $IMT2$ statistic performs relatively well, and it is more powerful than $IMT1$ despite being more conservative in Situation 2 in the significance level study.

The complementary log-log model

For M_{T_2}, when the correct link function was the complementary log-log link, Stukel's score statistic was the most powerful. When $n = 100$, it exhibited an estimated power greater than 40%, whereas the remaining statistics all had power less than 20%. Furthermore, Stukel's score statistic has an empirical significance level which is reasonably close to α in Situation 2. Stukel's LRT1 and LRT2 had identical power, and were almost as powerful as Stukel's score test. In contrast to the results for M_{T_1}, in this case the $IMT2$ was the worst performer.

The long tails model

When the correct model was M_{T_3}, where the mean functions had longer tails than the standard logistic model, the highest percentages of rejection belonged to X^2_{st}. This was not unexpected due to how alarmingly anti-conservative the standardised Pearson test was for all sample sizes in Situation 4. When $n = 100$ and 1000, the statistic with the highest power, which also had an acceptable $\hat{\alpha}_4$, was $\hat{S}_{st}$. When $n = 500$, however, it was the IMT statistics.

The short tails model

When the correct model was M_{T_4}, where the mean functions had shorter tails than the standard logistic model, the highest percentages of rejection belonged to the $IMT2$ statistic. The $IMT2$ statistic performed a lot better than $IMT1$ for sample size $n = 100$, and slightly better for larger n. When $n = 100$, the $IMT1$ is relatively conservative, but it has $\hat{\alpha}_2$ close to α for larger n. Similarly to the results reported in Hosmer et al. (1997), in the $n = 100$ version of M_{T_4}, the power of the standardised Pearson test is almost 0% – far below the nominal significance level α.

The asymmetric tails models

There was no obvious corresponding situation from the significance level study in terms of distribution of fitted probabilities. The statistics performed very similarly when the true model was M_{T_5} compared to when the true model was M_{T_6}. This indicates that reversing which tail is short and which is long has no significant consequences in terms of power. The distributions of $\hat{\boldsymbol{\pi}}_5$ and $\hat{\boldsymbol{\pi}}_6$ were fairly similar. Several examples tested out with different simulated data show a distribution of the fitted probabilities more similar to that of the long-short model example, than what we see in Figure 4.4c.

Chapter 5

Summary of Simulation Studies

5.1 The empirical significance levels of the GOF test statistics

Overall, six of the seven statistics had a rejection rate fairly close to 5% for all null hypotheses and all three sample sizes.

In general, the information matrix test (IMT) statistics are the closest to the desirable empirical significance level of 5% in more situations than the other tests when the sample size was equal to 1000. $IMT1$ was preferable over $IMT2$ for the smallest sample size, whereas $IMT2$ appeared to have a more desirable significance level for larger sample sizes. Stukel's score test also performed well for sample size equal to 1000.

For the smaller sample sizes $n = 100$ and $n = 500$, the USS test appears to be the best choice in most situations, closely followed by Stukel's score test. On the other hand, when the sample size is equal to 1000, it is the $IMT2$ statistic that yielded more empirical significance levels closer to α in the study.

Among the three Stukel's tests, the empirical significance level of Stukel's score test was closer to the nominal level than the Stukel's LRT tests in most situations, though the advantage of Stukel's score test statistic was slightly less dominant for larger sample sizes. This was expected due to score tests and likelihood ratio tests having the same asymptotic distributions (see Section 2.2).

In brief, Stukel's LRT1 and LRT2 statistics behaved identically or very similarly most of the time, but not in situations that produced highly skewed distributions of $\hat{\pi}_i$. These types of settings are what initially motivated modifying the original Stukel's LRT1 statistic, so this difference was intended. In these settings, we saw that the LRT2 had a more desirable empirical significance level.

An implication of this study is the possibility that the LRT2 statistic might be a preferable alternative to the original LRT1 statistic with regards to type I errors. Nevertheless, a more comprehensive study of modified Stukel LRT statistics is necessary. The rule of excluding one of Stukel's additional variables from the alternative model when its observed vector consisted of less than 10% non-zero elements was chosen by individual discretion, and was not the outcome of systematic and extensive exploration.

In some situations the standardised Pearson chi-square statistic $\hat{X}_{st}^2$ rejected H_0 at an alarming high rate – even when $n = 1000$. This was surprising considering how widespread Osius-Rojek standardisation procedures are, but in line with weaknesses pointed out in texts like Hosmer (2013).

In Section 3.3, we saw that X_{st}^2 was very unstable for estimated probabilities with distributions closely centred around 0.5, or with highly skewed distributions with fat tails where a considerable proportion of the probabilities were close to 0.5. The USS test was much more stable overall, despite having a similar standardisation method. It was slightly conservative, but the size of its rejection region stabilised as with larger sample sizes.

The rate of large values of X_{st}^2 (which result in rejection of H_0) when $\hat{\pi}_i$ are close to 0.5 is most likely due to the functional form of the classic Pearson chi-square statistic X^2 (see equation 2.1). The squared Pearson residual, which is an observation's contribution to X^2, is defined as

$$r(y_i, \hat{\pi}_i)^2 = \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)}, \tag{5.1}$$

(Hosmer, 2013). Hence when $\hat{\pi}_i$ for a specific observation is close to 0.5, the contribution $r(y_i, \hat{\pi}_i)^2$ will be of a considerable magnitude (close to 1). If the majority of the observations have $\hat{\pi}_i$ close to 0.5 (such as in situation 4), then X^2 is likely to be sizeable.

As noted by Hosmer (2013), observations where either: 1) $y_i = 1$ and $\hat{\pi}_i$ is small, or 2) $y_i = 0$ and $\hat{\pi}_i$ is large, can result in a X^2 large enough to reject H_0. One single observation with that characteristic can be enough to inflate X^2 to a value resulting in rejection. If $y_i = 1$ and $\hat{\pi}_i = 0.18$, for example, the contribution $r(1, 0.18)^2$ is approximately 4.56. In situation 4, where most of the observations produce $r(y_i, \hat{\pi}_i)^2$ close to 1, one single instance where $|y_i - \hat{\pi}_i|$ is large can be enough to skyrocket the value of X^2. The standardised USS statistic $S = \sum_{i=1}^{n} (y_i - \pi_i)^2$ does not have this issue as the squared residual for each observation can be no larger than 1.

Hosmer (2013) also mentioned that observations with large $|y_i - \hat{\pi}_i|$ can also amplify $\hat{\sigma}_P^2$, the estimated variance of X^2. This reduces the size of X_{st}^2, which in turn may cause the standardised Pearson test to fail to reject even though X^2 is large. The observed values of $\hat{\pi}_j(1-\hat{\pi}_j)$ are less influential when estimating the variance of the standardised USS statistic $\hat{S}$. Hence the USS test is favoured in settings like these (Hosmer, 2013). This is fits well with our findings from Section 3.4.

As previously mentioned, parts of our power study design overlaps with the simulation study in Hosmer et al. (1997). We have three GOF tests in common, namely the standardised Pearson, the USS, and Stukel's score test. Table 4.1 shows considerably different power to detect model departures *D3* and *D4* than results reported by Hosmer et al. (1997) for these three tests.

The aforementioned differences could be attributed to the differing methods and software used to generate the true probabilities of M_{T_j}. Different statistical software was used, and the equations used when computing the true probabilities for M_{T_1} and M_{T_2} were dissimilar. This should be taken under advisement when comparing our test performances to the reported results in Hosmer et al. (1997). This is particularly important in relation to X_{st}^2. It can be argued that there is considerable ambiguity in Hosmer et al. (1997) regarding which moment estimates and distribution were actually used when calculating the p-values for their simulation results. In addition, the Osius and Rojek estimator of the mean of the classic Pearson chi-squared statistic used when computing X_{st}^2 is not identical to the estimator in Hosmer et al. (1997) (see Section 2.1).

5.2 The empirical power of the GOF test statistics

5.2.1 Omission of a quadratic term

All seven tests managed very well to detect the omission of a quadratic term when its effect $\hat{\beta}_2$ was greater than 0.2. The large increase in power which occurred when $\hat{\beta}_2$ increased from approximately 0.04 to 0.22 was also present in the equivalent set-up in Hosmer et al. (1997). This is most likely explained by the large increment in $\hat{\beta}_2$. It is plausible that including additional correct models with $\beta_{j2} \in [0.04, 0.21]$ in the study design would have been better for determining how influential the quadratic term must be for the statistics to detect that it is missing from M_j.

Among the statistics which were not anti-conservative when the sample size equalled 100, Stukel's score statistic was the one with the highest power detect the omitted quadratic term. Stukel's LRT2 was marginally more powerful than LRT1 in a few settings, otherwise the two statistics had identical power to detect the missing quadratic term. The results for this departure type do not give strong evidence for favouring the modified LRT2 over the original LRT1, but it certainly does not discount the potential of the modified version.

The $IMT1$ had considerably higher power than $IMT2$ in all situations. The standardised Pearson test performed well when the sample sizes were 500 and 1000 for estimated probabilities where a large proportion of the estimates were close to either 0 or 1.

5.2.2 Omission of a log term

When a log term was omitted, none of the GOF tests had any substantial power for the sample sizes included in the study. Stukel's LRT1 and LRT2 were the most powerful in almost every configuration. The power of the LRT1 and LRT2 statistics were identical, but there are indications that the significance of the LRT2 statistic was more appropriate.

The standardised Pearson test and the USS were the least powerful when assessing whether the specified model was missing a log term. Even when the sample size was specified as $10,000$, none of them exceeded 80% in the cases with the most pronounced lack of linearity, whereas the other five test statistics resulted in 100% rejections. The IMT statistics and Stukel's score test statistic were slightly more powerful than the standardised Pearson test and the USS test. In general, $IMT1$ was more powerful than $IMT2$.

5.2.3 Omission of the main effect of a binary covariate and its interaction with a continuous covariate

For the true models which were the most dissimilar to the specified logistic regression model, none of the statistics had the power to detect the omitted binary variable and its interaction with the continuous variable more than 37% of the time. The tests with the least poor power were the USS test, Stukel's LRTs, and Stukel's score test. However, due to computation failure, Stukel's score test lacks results for most of the settings.

The high rejection rates of X_{st}^2 when the binary covariate and the interaction term are omitted, are undermined by its exceedingly high empirical significance levels. The estimated power of the standardised Pearson test is most likely attributed to its inflated rejection region for $\hat{\pi}_i$ distributed as described in the previous section – the models fitted in this study range from having highly skewed $\hat{\pi}_i$ with a considerable proportion of values close to 0.5, to being U-shaped and spanning the entire unit interval $[0, 1]$. Along with this range of $\hat{\pi}_i$, the standardised Pearson ranges from having high power to low power.

As expected due to the distributions of $\hat{\boldsymbol{\pi}}_j$, the estimated power of Stukel's LRT statistics were dissimilar in several settings. The LRT2 was more powerful than LRT1 in almost all situations with 100 observations. This may be explained by the fact that the LRT1 was considerably more anti-conservative than LRT2 in almost all of the settings where their empirical significance levels were not identical.

The USS test was more powerful than the information matrix tests, but there was no clear indication as to whether it performed better than Stukel's LRTs. The asymptotically equivalent $IMT1$ and $IMT2$ both performed poorly, though there was a tendency for $IMT2$ to be slightly more powerful despite being less anti-conservative in comparable situations from the significance level study.

5.2.4 Incorrectly specified link function

In general, Table 4.1d shows a wide variety of results, and there is no single statistic that stands out as being powerful across all models. As mentioned in Section 4.4, the results show considerably different power to detect a misspecified link function than results previously reported by Hosmer et al. (1997) for the three GOF statistics our studies have in common. These differences are probably partly due to our differing methods and software used to generate the true probabilities of M_{T_j}. Different statistical software was used, and the equations used when computing the true probabilities for M_{T_1} and M_{T_2} were dissimilar.

The three Stukel's test statistics performed relatively well when the true model was M_{T_2}, M_{T_4}, M_{T_5}, and M_{T_6}. The two models where Stukel's score test and Stukel's LRT statistics had relatively low power compared to the other models, was M_{T_1} and M_{T_3}. The probit model M_{T_1} was difficult to detect for all seven statistics. Additionally, since M_{T_3} produces $\hat{\boldsymbol{\pi}}_3$ which are similarly distributed to the fitted probabilities in Situation 4 in the significance level study, the high power displayed by X^2_{st} is seriously undermined.

When considering the aforementioned insights, Stukel's score test statistic and Stukel's LRT1 and LRT2 statistics, appear to be the best all-round alternatives. The LRT statistics performed identically in all the models for all sample sizes. When comparing Stukel's score test statistic to the LRT test statistics, the score test statistic had the highest frequency of greater power. Furthermore, the score test statistic had a much more ideal empirical significance level in Situation 2, and in Situation 4 its $\hat{\alpha}_4$ were very similar to those of Stukel's LRT1 and LRT2 statistics (recall the notes on the distributions of $\hat{\boldsymbol{\pi}}_j$ in Section 4.3.4). Hence Stukel's score test appears to be the better choice if there is suspicion of a different true link function than the specified logit link, though the USS test is also powerful in multiple settings.

Hosmer et al. (1997) noted that the tails contain only a small proportion of total amount of estimated probabilities, and that it is mostly in the tails one finds the differences between the correct links in the true models M_{T_j} and the logit link. I.e. there are relatively few observations in the area where the differences are the most pronounced. In addition, the expression $\hat{\pi}_i(1-\hat{\pi}_i)$ is central in the computation of multiple GOF statistics and, as indicated in Section 3.4, this expression is larger for estimated probabilities close to 0.5.

The tests using standardisation methods, X^2_{st} and $\hat{S}_{st}$, were more susceptible to performing poorly when the logit link function was very similar to the correct link function in the region around $\pi_{ji} = 0.5$. The estimations of the variance of the Pearson statistic and the standardised USS statistic are influenced by the range of $\hat{\pi}_{ji}$ and how clustered the $\hat{\pi}_{ji}$ are around 0.5.

The analogous performances for the two asymmetric tails models makes sense considering that the differences between the true link and the logit link (occurring in the tails) are similar in magnitude. The only distinction between the models when looking at Figure 4.4a is that the tails of the true mean function (or probability curve) are above the logistic probability curve in one model, and below the logistic probability curve in the other. This substantiates the claim that the most central characteristic, regarding whether a GOF test recognises that the link function is misspecified, is how the logistic mean function compares to the correct link function in the area around $\pi_{ji} = 0.5$.

5.2.5 Summary and recommendations

When using the Osius and Rojek standardisation method, the standardised Pearson test is unstable and not recommended as a GOF test. Almost all cases where X^2_{st} was more powerful than the other tests coincided with distributions of $\hat{\pi}_i$ which were very similar to situations in the significance level study were X^2_{st} much too large rejection rates.

In general, one should take into account how the estimated probabilities of one's model are distributed and determine whether: (1) the estimated probabilities are mostly clustered around 0.5; and (2) are they highly skewed, but still have enough observed values close to 0.5 to inflate the classic Pearson chi-square statistic? How to quantify what "enough" observed values of the estimated logistic probabilities is a topic which should be investigated in future studies.

Chapter 6

Data Analysis

6.1 Background

Risk stratifying patients and predicting mortality is important in a clinical setting with patients in need of critical care. There are many advantages to adequately modelling vital status at time of discharge using measures that reflect a patients physiological status and mortality risk when admitted to the ICU. Avoiding premature discharges, for example, may reduce costs for the hospitals and be beneficial to patient survival and recovery (Sluisveld et al., 2017).

Many hospitals have limited ICU capacity and must therefore prioritise their resources. Situations may arise where an ICU does not have the resources to provide optimal care for all their patients and transferrals to other units is not possible (Scales and Rubenfeld, 2014). Resource allocation in these unwanted settings should be fair and the inclusion criteria for receiving certain interventions should be supported by established evidence based models.

The Simplified Acute Physiology Score (SAPS) II, is a illness severity score calculated from information gathered post hospitalisation. This information includes variables such as temperature, systolic blood pressure, heart rate, age, type of admission, AIDS, and metastatic cancer (Moseson et al., 2014). Logistic regression models have been developed to assess effects of the SAPS II score on in-hospital mortality. These are mortality prediction models and have applications in assessing the performance of an ICU, assessing a patient's risk of death during the ICU stay, and also in quality control of clinical trials.

Two such mortality prediction models are the original SAPS II model introduced by Le Gall et al. (1993) and the modified SAPS II model presented by Haaland et al. (2014). The latter model was the result of the first time the original SAPS II model was recalibrated for data from a Nordic country. Le Gall et al. (1993) used vital status at hospital discharge as an endpoint to develop the SAPS II score itself, and fitted a mortality prediction model with the score as the explanatory variable.

A few years have passed since the model in Haaland et al. (2014) was developed using data collected by the Norwegian Intensive Care Registry (NIR). During this time, the population of intensive care patients may have changed (a larger proportion of older patients, for example), individual ICU performance may have improved, and there may have been significant benefits to new medical methods and technologies. It seems reasonable that the SAPS II model should be modified every 2-3 years (Haaland et al., 2014).

There are many possible consequences of poorly calibrated SAPS II score models. Resource allocation guidelines in extraordinary events such as mass casualties may be inadvertently unfair as a result. It is also possible that poor risk estimates can affect hospital budgets, and also give the erroneous impression that an ICU performs better, or worse, than previously or compared to other ICUs. Hence there is a need for evaluating, every so often, if the SAPS II model in use adequately fits the current population of patients.

In this thesis, we will carry out a study where the SAPS II model is fitted to a recent NIR data set with cohorts from 2016 and 2017. An evaluation of how well previously calibrated SAPS II score models predict ICU mortality in a recent data set, compared to our own version of those models will be conducted. We will also examine whether adding an additional explanatory variable to this model improves the fit to the sample data, and thus perhaps an improvement in regards to predicted risk of mortality. This additional variable is a binary covariate indicating whether a patient was admitted due to an acute non-surgical medical event or not.

6.2 The Data Set

This was a registry based study, where the data were provided by NIR. NIR is a Norwegian national quality registry which covers more than 90% of adult patient admissions to Norwegian ICUs. NIR collect data from over 40 intensive care units across the country which are distributed over 38 hospitals. This include both university hospitals, secondary hospitals, and primary hospitals. According to their annual report for 2017, they collected records of 13737 patients comprised of nearly 1.5 million hours of patient treatment spread over 49 different intensive care units.

Table 6.1 *Patient characteristics*

Patient characteristic	**All patients $\geq$ 18 years old**	**Study sample**
n	30,212	30,177
Age (years)		
Missing	0	0
Mean (sd)	64.7 (17.2)	64.7 (17.1)
Q1	56	56
Median	68	68
Q3	77	77
Sex		
Missing	0	0
Female, %	42.3%	42.3%
SAPS II score		
Missing	35	-
Mean (sd)	38.1 (17.4)	38.1 (17.4)
Q1	26	26
Median	36	36
Q3	48	48
Type of admission		
Missing	2	-
Acute medical, %	65.6%	65.7%
Planned surgery, %	22.9%	22.9%
Acute surgery, %	11.5%	11.4%
Vital status at ICU discharge		
Missing	2	-
Died during ICU stay, %	10.4%	10.4%
Survived ICU stay, %	89.6%	89.6%

Characteristics of all the patients aged 18 years or older, and the patients in the study sample. The 35 patients with missing observations of SAPS II score, type of admission, or vital status at ICU discharge, were excluded from the study sample/population. These characteristics were mostly unchanged by the exclusion of these 35 patients.

The source dataset we received from NIR contained records of 30,212 patients, of at least 18 years of age, who were admitted during 2016 or 2017. Re-admissions, admissions of patients transferred from other hospitals, and transfers from a different intensive care unit within the same hospital, were not excluded. The dataset is sizeable despite the registry seeing a decrease in admittance is in 2016.

35 admittances were excluded from the study sample due to missing SAPS II score, type of admission, and/or vital status at discharge from the ICU. The patient characteristics in the source data set and the study sample are presented in Figure 6.1. As can be seen in the table with patient characteristics, this exclusion barely affected the distribution of the study sample variables the study sample contained 30,177 observations of ICU stays.

6.3 Methods

6.3.1 The models

Three logistic regression models were specified in this study: Model A, Model B and Model C. In Model A, the SAPS II score was the only covariate. The specified linear predictor of Model A was

$$\eta_A = \beta_0 + \beta_1(\text{SAPS II}), \tag{6.1}$$

with covariate effect β_1 and intercept β_0. This linear predictor is used to compute the predicted risk of death (PRD) in Model A, defined as

$$PRD_A = \pi(\eta_A) = \frac{e^{\eta_A}}{1+e^{\eta_A}}, \tag{6.2}$$

using the standard logistic function due to the logit link of the logistic regression model. Thus, once the parameters are estimated, we will have a function with which we can predict ICU mortality based on a given SAPS II score.

Model B was based on Model A, but with an added log-transformed covariate, namely $\log(\text{SAPS II}+1)$ to allow for non-linearity of the linear predictor. The specified linear predictor of Model B was

$$\eta_B = \beta_0 + \beta_1(\text{SAPS II}) + \beta_2 \log(\text{SAPS II}+1), \tag{6.3}$$

with intercept β_0 and covariate effects β_1 and β_2. The linear component, η_B, is used to compute the predicted risks of death according to Model B.

The PRD in Model B, $PRD_B = \pi(\eta_B)$, is defined using the standard logistic function in the same manner as in (6.2). Once Model B has been fitted, it will be possible to compare PRD_B to the predicted risk of ICU mortality from the other models.

Model C was based on Model B, but with an added binary covariate. The additional binary covariate, called *acute medical admission* (AMA), was derived from the type of admission variable specifically for this study. In the analysis, AMA is takes on the value $1 =$ "yes" when the stay at the ICU was due to an acute medical admission (non-operative), and the value $0 =$ "no" if the admission was in conjunction with a planned or acute surgery (operative).

The intuition behind this was the assumption that acute medical admission may have a larger tendency toward ending with hospital mortality. When a patient is already admitted to a hospital before the event necessitating an ICU admission, there are factors which are less likely to be of concern. These factors may include incidences of severe traumatic injury, unknown patient identity and health conditions such as allergies, ambulance response time, and other pre-hospital factors.

The linear predictor specified for Model C was

$$\eta_B = \beta_0 + \beta_1(\text{SAPS II}) + \beta_2 \log(\text{SAPS II} + 1) + \beta_3(\text{AMA}), \tag{6.4}$$

with intercept β_0, main covariate effects β_1 and β_3, and the effect of the log term β_2. The predicted risk of ICU mortality in Model C is defined using the same link function as Model A and Model B, where $PRD_C = \pi(\eta_C) = {}^{e^{\eta_C}}/_{1+e^{\eta_C}}$. Thus once the parameters are estimated, we can predict ICU mortality based on a given SAPS II score and whether the ICU admission was non-operative or not.

Model B is identical (i.e. it uses the same link function and form of its linear predictor) to the original SAPS II model specified by Le Gall et al. (1993). Over 20 years ago, Le Gall et al. (1993) estimated the parameters of the original SAPS II model ($\boldsymbol{\beta}_D = [\beta_{D0} \;\; \beta_{D1} \;\; \beta_{D2}]^\mathsf{T}$) based on an international data set with 13,152 patients to provide a method of predicting vital status as hospital discharge.

We will refer to the original SAPS II model and its estimated parameters as Model D. The fitted linear predictor of Model D is

$$\hat{\eta}_D = -7.7631 + 0.0737(\text{SAPS II}) + 0.9971\log(\text{SAPS II} + 1), \tag{6.5}$$

and the estimated PRD based on $\hat{\eta}_D$ is defined as $PRD_D = e^{\hat{\eta}_D}/(1+e^{\hat{\eta}_D})$.

As noted by Haaland et al. (2014), Model D is not optimally fitted for mortality predictions in present-day populations because the estimates of $\boldsymbol{\beta}_D$ are based on more than two decade old data. Using a data set consisting of the 2008-2010 cohorts from NIR, Haaland et al. (2014) fitted the following model, referred to as Model E,

$$\hat{\eta}_E = -9.0917 + 0.0325(\text{SAPS II}) + 1.6698\log(\text{SAPS II} + 1), \tag{6.6}$$

which had more accurate predictions of ICU mortality than Model D. These estimated ICU mortality predictions are defined as $PRD_E = e^{\hat{\eta}_E}/(1+e^{\hat{\eta}_E})$.

It is expected that Model B will result in different parameter estimates compared to that of Model D and Model E. The data set used by Le Gall et al. (1993) is different in several ways, both in terms of time frame and geographical hospital locations. The sample of NIR data used to estimate $\boldsymbol{\beta}_E = [\beta_{E0} \;\; \beta_{E1} \;\; \beta_{E2}]^{\mathsf{T}}$ in Model E does not overlap with the study sample analysed in this thesis, which is restricted to admissions in 2016 and 2017.

In addition, the sample used by Haaland et al. (2014) did not include 2552 patients (around 6% of the source population) due to them being readmitted to the same or different ICU multiple times during one patient care process. The variables needed to filter in this manner were not available during the development of this thesis. Hence a small fraction of the response variables in the study sample are likely to be dependent.

6.4 Statistical analysis

All seven goodness-of-fit (GOF) statistics from Chapter 2 were applied to the models in Section 6.3.1. The power simulation study indicated that the GOF tests had low power when a specified model had omitted a log term for $n \leq 1000$. In the simulation studies, we mentioned that for simulated data sets where $n = 10,000$, the power of the IMT statistics and the three Stukel's statistics to detect an omitted log term was reasonably high when the effect of the omitted term, i.e. the parameter, was greater than 1.

The remaining two statistics, the standardised Pearson and standardised USS test statistic, had considerable poorer power in the same $n = 1000$ simulations. Since we have over 30,000 observations in our study sample, we do not expect Stukel's LRT1 and LRT2, and Stukel's score test, to overrate model adequacy when evaluating the fit of Model A. It was expected that the p-values produced by Stukel's score test statistic, Stukel's LRT1 and LRT2, $IMT1$, and $IMT2$ (when testing H_0 that Model A is correct) would be smaller than those resulting from the standardised Pearson test and the USS test, i.e. that X^2_{st} and $\hat{S}_{st}$ would be less discerning or anti-conservative.

All of the goodness of fit tests performed sub-optimally when the main effect of a binary covariate and its interaction with a continuous covariates was omitted. Since Model C has no interaction term, there is no clear comparison we can make to the departure types studied in the power study. Hence there are no clear indications from the simulation studies of how the GOF tests will perform when applied to the risk estimates of model C. However, we can look at the distributions of the estimated mortality risks, and bear in mind how the test statistics performed when applied to similarly distributed estimated probabilities in Chapters 3 and 4.

6.5 Results

The parameters of Model A, Model B, and Model C, were fitted on the same data set using the `glm()` function in `R`. Figure 6.1 shows the histograms of the estimated probabilities (or predicted mortalities) in each of the three models. These plots are included for comparison to the estimated probabilities in the significance level study and the power study.

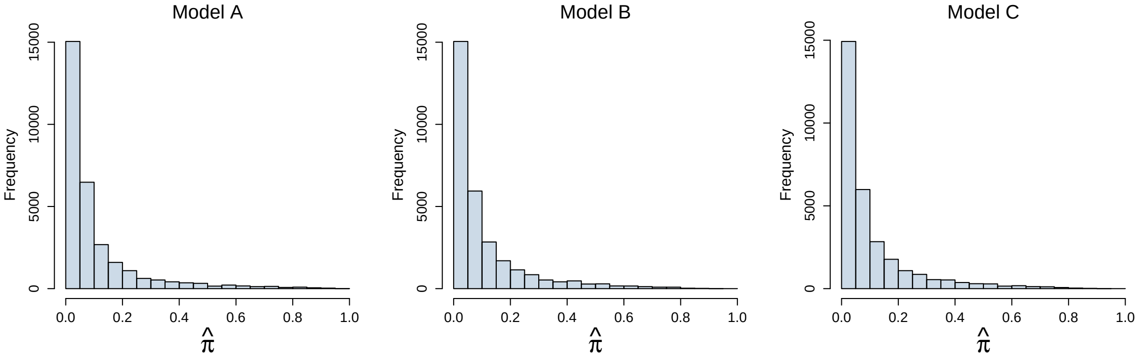


Fig. 6.1 Histograms of the estimated probabilities of Model A, Model B, and Model C.

The estimated linear predictor of Model A was

$$\hat{\eta}_A = -5.6101 + 0.0745(\text{SAPS II}). \tag{6.7}$$

The estimated linear predictor of Model B was

$$\begin{aligned}\hat{\eta}_B = &- 12.0832 + 0.0310(\text{SAPS II}) \\ &+ 2.2311 \log(\text{SAPS II} + 1).\end{aligned} \tag{6.8}$$

The estimated linear predictor of Model C was

$$\begin{aligned}\hat{\eta}_C = &- 12.1799 + 0.0308(\text{SAPS II}) \\ &+ 2.2164 \log(\text{SAPS II} + 1) + 0.2309(\text{AMA}).\end{aligned} \tag{6.9}$$

From visually assessing the histograms belonging to Model A, Model B, and Model C, we see that they are highly right skewed, and arguably comparable in shape to the histogram of $\hat{\pi}_i$ in Situation 11 in Figure 3.3b. None of the example histograms of fitted probabilities in the power study are considered reasonably similar to that of the three models in this study of NIR data.

The resulting estimated logistic regression coefficients of Model A, Model B, and Model C, along with their respective p-values, Akaike's information criteria (AICs), and residual deviances, are given in Table 6.2. In all three models, every covariate is significant according to the chi-square tests performed in R.

In addition, the AIC for Model B is smaller than for Model A, supporting the claim that the linear predictor form suggested by Le Gall et al. (1993) is more appropriate than Model A. Examining Figure 6.2a, shows that the upper tail of Model B has a closer fit to the observed mortalities of patients with a particular SAPS II score. Model B also appears to fit the data better for SAPS II scores less than 75.

When adding the binary variable AMA, the AIC is even smaller than for Model B, hence suggesting that patients admitted due to non-surgical related medical events have a higher risk of not surviving an ICU stay. This is illustrated by Figure 6.2b, where we see that the curve for $\widehat{PRD}_C = \pi(\hat{\eta}_C)$ given AMA= 1 is higher than $\widehat{PRD}_C$ given AMA= 0.

The predicted risk of ICU mortality according to Model D from Le Gall et al. (1993) and the more current Model E from Haaland et al. (2014) were plotted against $\widehat{PRD}_B$ in each their own plot. Figure 6.3a shows that Model D grossly overestimates the ICU mortality risk

Table 6.2 The estimation results of Model A, Model B, and Model C.

	Model A	Model B	Model C
$\hat{\beta}_0$ (std. error)	-5.6101 (0.0686), p-value $< 2.0 \times 10^{-16}$	-12.0832 (0.8978), p-value $< 2.0 \times 10^{-16}$	-12.1799 (0.9010), p-value $< 2.0 \times 10^{-16}$
$\hat{\beta}_1$ (std. error)	0.0745 (0.0012), p-value $< 2.0 \times 10^{-16}$	0.0310 (0.0059), p-value $= 1.65 \times 10^{-7}$	0.0308 (0.0060), p-value $= 2.21 \times 10^{-7}$
$\hat{\beta}_2$ (std. error)	NA	2.2311 (0.3050), p-value $= 2.55 \times 10^{-13}$	2.2164 (0.3060), p-value $= 4.37 \times 10^{-13}$
$\hat{\beta}_3$ (std. error)	NA	NA	0.2309 (0.0483), p-value $= 1.71 \times 10^{-6}$
AIC	15420.15	15360.27	15338.90
Residual deviance	15416.15	15354.27	15330.9

for the NIR study sample. In fact, the only observations $\widehat{PRD}_D$ appears to approximate with some adequacy are the cases with patient with very high SAPS II scores.

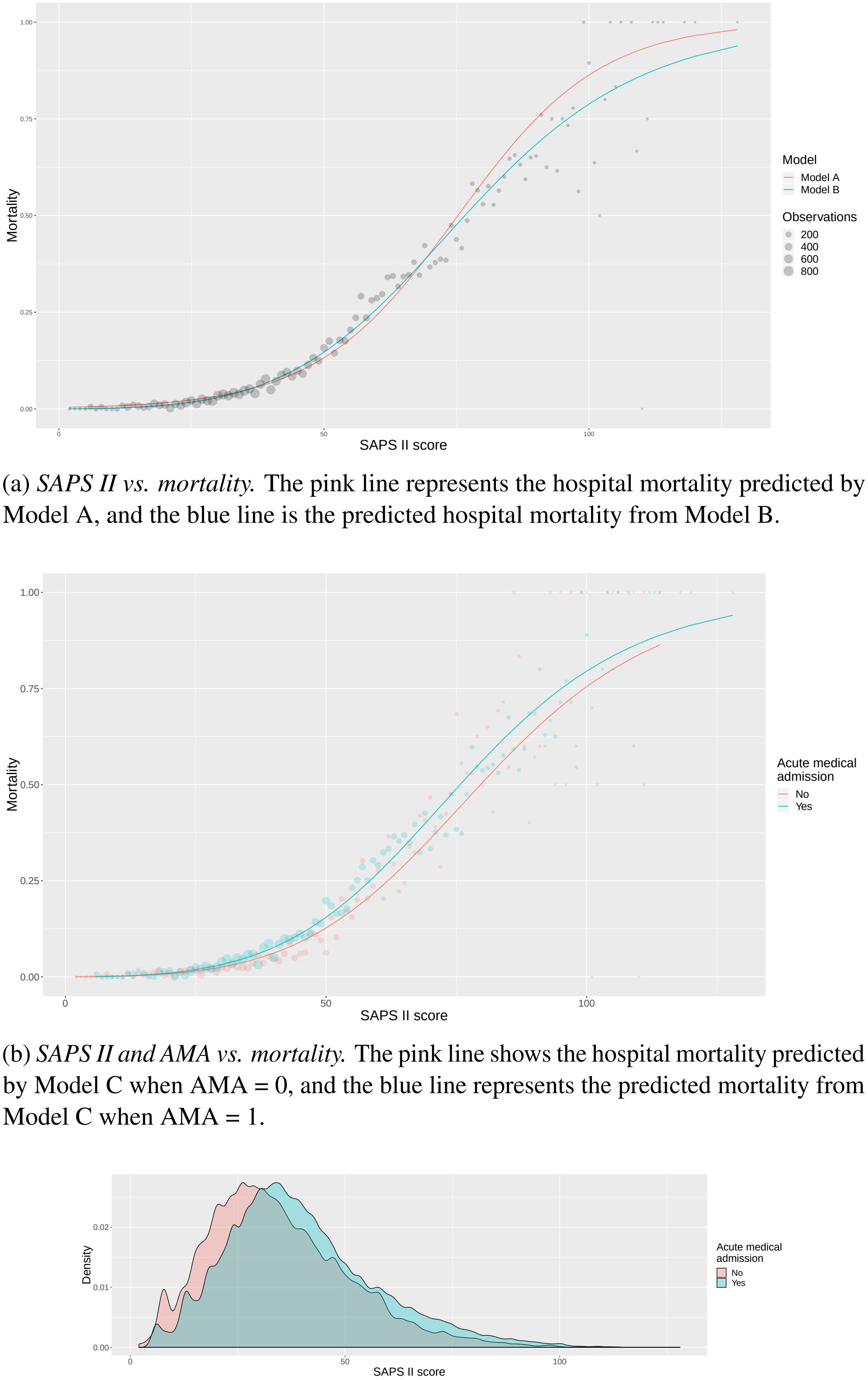
As one would expect from a model fitted to a presumably comparable population, the predicted risk of ICU mortality from Model E in Figure 6.3b shows that Model E is much more similar to Model B than Model D in terms of predicted ICU mortality risk for our study sample. Nevertheless, $\widehat{PRD}_E$ overestimates the mortality risk for the population in this study.

These differences may also in part be attributed to the possibility that the endpoints we received from NIR were vital status at time of discharge from ICU, whereas Le Gall et al. (1993) and Haaland et al. (2014) describe their endpoints as "hospital mortalities" or "vital status at hospital discharge", which may refer to a later point in time during a hospital stay than vital status at ICU discharge.

The p-values resulting from the GOF tests, accompanied by the observed test statistic and degrees of freedom where appropriate, are listed in Table 6.3. These results show that all seven GOF tests prefer Model B over Model A, since all the p-values are substantially larger for Model B. It is encouraging that the USS test, Stukel's score test, Stukel's LRT1, and Stukel's LRT2, would not have rejected Model B as the correct model at the 5% significance level, and that all the tests would have rejected the fit of Model A at even smaller significance levels. It is worth noting that Stukel's LRT2 had a higher p-value than Stukel's LRT1 for Model B, whereas they were approximately equal for Model A.

Table 6.3 does not, however, indicate that the fit of Model C is better than Model B, despite the AIC's preference. All the GOF tests, except the USS test, would reject the fit of Model C at the 5% significance level. The only test which preferred Model C over Model B was the standardised Pearson test, which is not promising for the adequacy of Model C in view of the results highlighted in Chapter 5.

The standardised Pearson test had a relatively large p-value compared to the other tests. This finding confirms part of our expectations from Section 6.4, but rather surprisingly, the USS statistic's p-value was the third smallest out of the seven statistics. This was unanticipated due to how the USS test performed in the power study in Chapter 4.

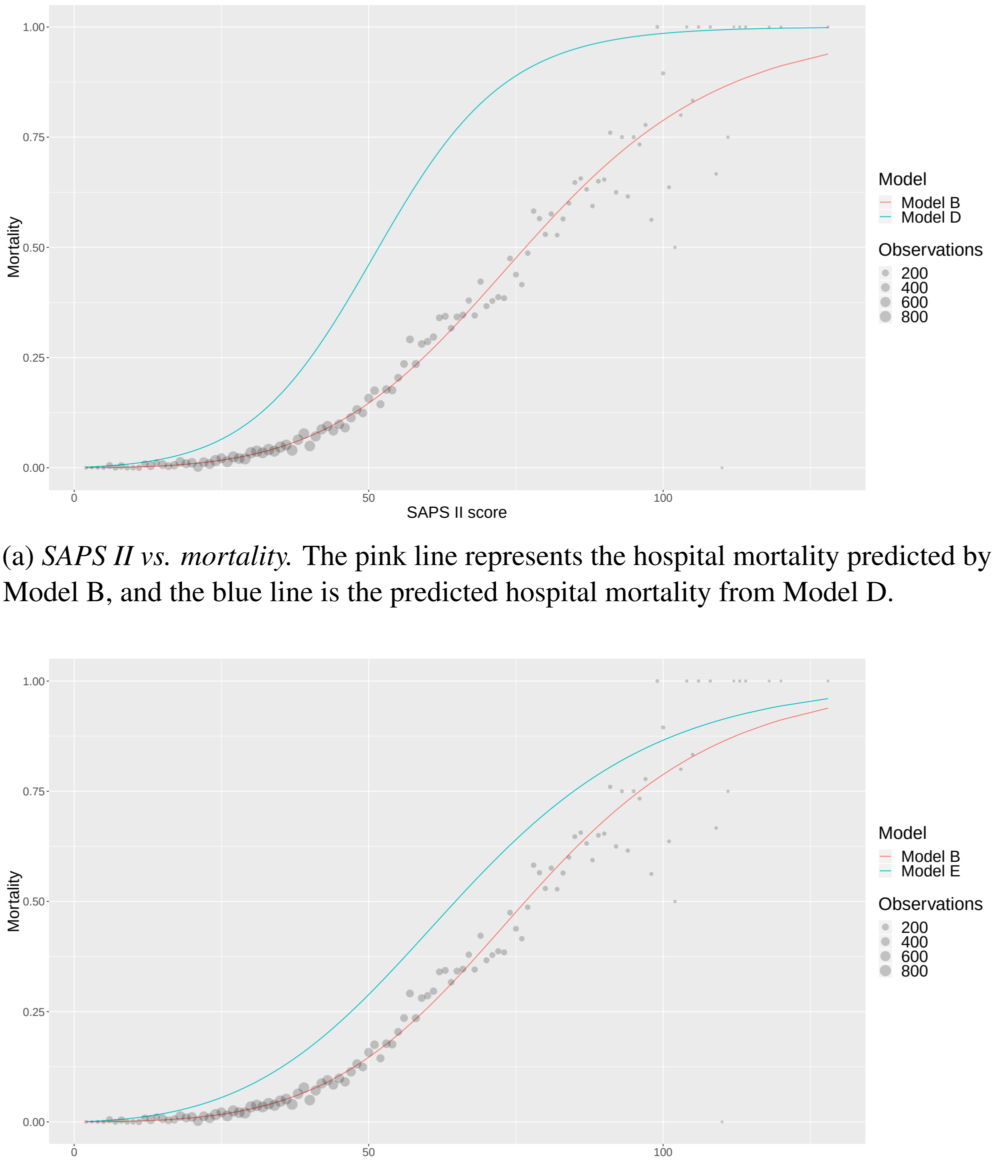


(a) *SAPS II vs. mortality.* The pink line represents the hospital mortality predicted by Model A, and the blue line is the predicted hospital mortality from Model B.

(b) *SAPS II and AMA vs. mortality.* The pink line shows the hospital mortality predicted by Model C when AMA = 0, and the blue line represents the predicted mortality from Model C when AMA = 1.

(c) The distribution of SAPS II score when AMA = 0 ("No"), and the distribution when AMA = 1 ("Yes"), based on the SAPS II score.

Fig. 6.2 The predicted hospital mortalities of the different models. The diameter of the circles reflect the number of observed patients with a particular SAPS II score, and its vertical placement shows the observed mortality for the patient(s) with that particular SAPS II score.



(a) *SAPS II vs. mortality.* The pink line represents the hospital mortality predicted by Model B, and the blue line is the predicted hospital mortality from Model D.

(b) *SAPS II vs. mortality.* The pink line represents the hospital mortality predicted by Model B, and the blue line is the predicted hospital mortality from Model E.

Fig. 6.3 A comparison of the predicted hospital mortalities from Model B vs. Model D, Model B vs. Model E, applied to the study's 2016-2017 NIR data set. The diameter of the circles reflect the number of observed patients with a particular SAPS II score, and its vertical placement shows the observed mortality for the patient(s) with that particular SAPS II score.

Table 6.3 The goodness-of-fit assessments

	Model A			Model B			Model C		
GOF test	p-value	Observed test statistic	df	p-value	Observed test statistic	df	p-value	Observed test statistic	df
The standardised Pearson test	7.69×10^{-10}	-6.151	NA	0.001686211	3.141	NA	0.002697824	3.000	NA
The USS test	3.781546×10^{-17}	8.419	NA	0.3771063	0.883	NA	0.1293031	1.517	NA
Stukel’s score test	3.529×10^{-16}	71.161	2	0.1584	3.685	2	0.001149	13.538	2
Stukel’s LRT1	1.110223×10^{-16}	72.951	2	0.1433656	3.885	2	0.001108824	13.609	2
Stukel’s LRT2	0	72.917	1	0.6601055	0.193	1	0.007734899	7.094	1
IMT1	3.331×10^{-10}	75.202	3	0.01036829	16.720	6	4.469661×10^{-5}	37.591	10
IMT2	0	83.404	3	0.01915316	15.146	6	0.0001766674	34.114	10

Bibliography

Agresti, A. (2013). Categorical data analysis.

Bilder, C. R. and Loughin, T. M. (2014). *Analysis of categorical data with R*. Chapman & Hall/CRC Texts in Statistical Science.

Bliss, C. I. (1934). The method of probits. *Science (New York, N.Y.)*, 79(2037).

Copas, J. (1989). Unweighted sum of squares test for proportions. *Journal Of The Royal Statistical Society Series C-Applied Statistics*, 38(1):71–80.

Davidson, R. and Mackinnon, J. G. (1984). Convenient specification tests for logit and probit models. *Journal of Econometrics*, 25(3):241–262.

Devore, J. L. and Berk, K. N. (2012). *Modern Mathematical Statistics with Applications*. Springer Texts in Statistics. Springer New York, New York, NY, second edition edition.

Dobson, A. J. (2008). An introduction to generalized linear models.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467.

Haaland, . A., Lindemark, F., Flaatten, H., Kvåle, R., and Johansson, K. A. (2014). A calibration study of saps ii with norwegian intensive care registry data. *Acta Anaesthesiologica Scandinavica*, 58(6):701–708.

Hosmer, D. W. (2013). Applied logistic regression.

Hosmer, D. W. and Hjort, N. L. (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine*, 21(18):2723–2738.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning : with applications in r.

Jørgensen, B. (1984). The delta algorithm and glim. *International Statistical Review / Revue Internationale de Statistique*, 52(3):283–300.

Lai, C. D. (2013). Generalized weibull distributions.

Le Gall, J., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24):2957–2963.

McCool, J. (2012). Using the weibull distribution : reliability, modeling, and inference.

McCullagh, P. (1985). On the asymptotic distribution of pearson's statistic in linear exponential-family models. *International Statistical Review / Revue Internationale de Statistique*, 53(1):61–67.

Moseson, E. M., Zhuo, H., Chu, J., Stein, J. C., Matthay, M. A., Kangelaris, K. N., Liu, K. D., and Calfee, C. S. (2014). Intensive care unit scoring systems outperform emergency department scoring systems for mortality prediction in critically ill patients: a prospective cohort study. *Journal of intensive care*, 2.

Orme, C. (1988). The calculation of the information matrix test for binary data models. *The Manchester School*, 56(4):370–376.

Orme, C. (1990). The small-sample performance of the information-matrix test. *Journal of Econometrics*, 46(3):309–331.

Osius, G. and Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal Of The American Statistical Association*, 87(420):1145–1152.

Scales, D. C. and Rubenfeld, G. D. (2014). *The Organization of Critical Care an Evidence-Based Approach to Improving Quality*. Springer New York.

Sluisveld, N. v., Bakhshi-Raiez, F., Keizer, N. d., Holman, R., Westert, G., Wollersheim, H., Hoeven, J. v. d., and Zegers, M. (2017). Variation in rates of icu readmissions and post-icu in-hospital mortality and their association with icu discharge practices. *BMC Health Services Research*, 17:281–6963.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Yan, X. (2009). Linear regression analysis : theory and computing.