UNIVERSITÀ DEGLI STUDI DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXVI

Variance-invariant inference for regression models

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Livio Finos

Co-supervisore: Prof. Jelle J. Goeman

Dottorando/a: Riccardo De Santis

10 January 2024

Abstract

Generalized linear models usually assume a common dispersion parameter, an assumption that is seldom true in practice. Consequently, standard parametric methods may suffer appreciable loss of type I error control. As an alternative, we present a semi-parametric group-invariant method based on sign flipping of score contributions. Our method requires only the correct specification of the mean model, but is robust against any misspecification of the variance. We will further extend the methodology to multivariate responses. Indeed, the weaknesses of the standard approaches can dramatically propagate in multivariate settings. We propose a resampling-based method to handle multivariate generalized linear models which adapts to the unknown correlation structure, resulting in an appreciable gain of power compared to general alternatives when such correlation is present. Finally, we will exploit the benefits of the proposed sign-flipping test to semiparametric models by considering the Cox regression model, where standard approaches rely on asymptotic arguments, while they have a slow convergence to the nominal level of the test.

Sommario

I modelli lineari generalizzati assumono solitamente un unico comune parametro di dispersione, un assunzione che risulta spesso non essere verosimile. Conseguentemente, i metodi parametrici standard possono soffrire una rilevante perdita del controllo dell'errore di I tipo. In alternativa noi presentiamo un metodo semiparametrico basato sul cambio del segno dei contributi allo score. Il nostro metodo richiede solamente la corretta specificazione della funzione della media, mentre risulta robusto rispetto a misspecificazioni della varianza. Successivamente estendiamo il metodo proposto al caso di risposte multivariate. Infatti, le debolezze degli approcci parametrici possono propagarsi in modo importante in problemi multivariati. Noi proponiamo un metodo basato sul ricampionamento per gestire i modelli lineari generalizzati multivariati, che si adatta alla sconosciuta struttura di correlazione delle risposte, risultando in un rilevante guadagno di potenza rispetto ai metodi parametrici specialmente quando questa correlazione è presente. Per concludere estendiamo il test basato sul cambio del segno dei contributi allo score a modelli semiparametrici considerando il modello di regressione di Cox, dove gli approcci standard sono giustificati sulla base di argomenti asintotici, mentre mostrano una lenta convergenza al livello nominale del test.

To my family

Acknowledgements

First of all I must thank professors Livio Finos and Jelle Goeman. They have always supported me with enthusiasm and they were always available for any question or discussion. This manuscript would never have been written without them.

I thank also professors Jesse Hemerik, Hein Putter and Dr. Samuel Davenport for some useful discussions and insights about relevant topics of this manuscript.

I thank all the colleagues I met in Padova, during the three years, and the colleagues I met in Leiden during my six months visiting. We have spent much time together during the breaks, or hanging out in the evenings. I had a lot of fun thanks to them.

Finally I want to thank my family and all my friends in Siena and around the Italy. Many of them were always available for general discussion about the PhD life, sharing thougths and worries, and for distracting from the work having fun together.

Contents

List of Figures **xiii**

List of Tables **xv**

Introduction **3**

Overview . . . 3

Main contributions of the thesis . . . 5

1 Inference in generalized linear models with robustness to misspecified variances **7**

1.1 Generalized linear models . . . 7

1.2 Classic approach to hypothesis testing . . . 10

1.3 Sign-flipping effective score test . . . 12

1.4 Standardized sign-flips score test . . . 14

1.5 Robustness of the test . . . 22

1.6 Testing for multiple regressors . . . 25

1.7 Simulation study . . . 28

2 Permutation-based multiple testing when fitting many generalized linear models **33**

2.1 Multivariate generalized linear models . . . 33

2.2 Local test for multivariate responses and the multiple testing problem . . 34

2.3 Sign-flip tests for multivariate generalized linear models . . . 37

2.4 Simulation Study . . . 42

2.4.1 Univariate . . . 42

2.4.2 Multivariate . . . 44

2.5 Real-data application . . . 46

3 Sign-flip tests for the Cox regression model **49**

3.1 Cox regression model . . . 49

3.2 Connection with Poisson model . . . 52

3.3 Hypothesis testing for the Cox model . . . 55

3.4 Simulation study . . . 57

Conclusions **63**

Bibliography **67**

List of Figures

1.1 Type I error control comparison. **Left:** correct Poisson model. **Right:** correct Logistic model. 29

1.2 Type I error control comparison. **Left:** Normal with nuisance heteroscedasticity. **Right:** Normal with target heteroscedasticity. 29

1.3 Type I error control comparison. **Left:** false Poisson model. **Right:** two groups Negative-binomial. 30

1.4 Power comparison. **Left:** correct Poisson model. **Right:** correct Normal model. 31

2.1 Logit model, univariate. Sample size of 50. On abscissa the Bonferroni-adjusted alpha for the control of FWER at level 5%, i.e. .05/number of tests 43

2.2 Logit model, univariate. Sample size of 100. On abscissa the Bonferroni-adjusted alpha for the control of FWER at level 5%, i.e. .05/number of tests 43

2.3 Logit model, multivariate. Sample size equal to 50, FWER control 44

2.4 Logit model, multivariate. Sample size equal to 50, Power comparison . . 45

2.5 Logit model, multivariate. Sample size equal to 100, FWER control . . . 46

2.6 Logit model, multivariate. Sample size equal to 100, Power comparison . 47

3.1 Type I error control for continuous covariate. Average censoring equal to 25%. 58

3.2 Type I error control for a factor covariate. Average censoring equal to 25%. 58

3.3 Type I error control for continuous covariate. Average censoring equal to 50%. 59

3.4 Power comparison for continuous covariate. Average censoring equal to 25%. 59

3.5 Type I error with omitted covariate. Average censoring equal to 25%. . . 60

3.6 Type I error when the true model is AFT. Average censoring equal to 25%. 60

List of Tables

2.1 FWER-controlled number of rejections 48
2.2 Number of true discoveries 48

3.1 Restructured data for subject i 53
3.2 Toy example: original dataset 54
3.3 Toy example: reconstructed dataset 54

Introduction

Overview

Regression models are popular methods used to quantify the relationship between a set of covariates and one (or more) outcome of interest. The inference procedure is carried on by firstly adopting a statistical model, which is usually assumed to be known except for a finite number of parameters, and the aim is to give statements about the parameters which represents the impact of each covariate on the target outcome. The most common approach for the parameter estimation is the adoption of the maximum likelihood principle which has some properties that makes the procedure attractive (Azzalini, 1996).

Once we have obtained a point estimate we might be interested in conducting a more accurate analysis. In particular, the focus can be on investigating the relevance of one (or more) target parameter. Call β a parameter of interest; we might focus on a hypothesis of the kind $H_0 : \beta = 0$, that is, if the target parameter is statistically significant, or we might be interested in a more general conjecture given by a generic $H_0 : \beta = \beta_0$. Note that H_0 is generally defined as null hypothesis. Hypothesis testing is a crucial step in a statistical analysis as it aims to quantify if the relevance of each covariate is significant in the adopted regression model.

The standard parametric approach for hypothesis testing is based on three alternatives which have similar behavior, especially for growing sample size, the Wald test, the Score test and the Likelihood ratio tests. We will refer to them as parametric tests because they are derived under the chosen model and are therefore exposed to the whole set of assumptions made. The choice of the model is probably the most important step in any statistical analysis as the quality of the inference depends on this choice. However, its appropriateness is often not correctly questioned, or it might be difficult to check all the assumptions made. For example, when testing for equality of means between two groups, statisticians often worry that there may be a difference in variance between the two groups. The task of properly taking this variance into account, known as the Behrens-Fisher problem (Fisher, 1935, 1941), has generated a huge statistical literature (Kim and Cohen, 1998; Chang and Pal, 2008). In the linear regression model, which generalizes the two-group model, the assumption that all error terms have equal variance is less often questioned. While goodness-of-fit tests and other diagnostics for heteroskedasticity exist (Goldfeld and Quandt, 1965; Glejser, 1969; Breusch and Pagan, 1979; Jarque and Bera, 1980; Cook and Weisberg, 1983; Long and Ervin, 2000), there is no simple and general way to follow up on obvious lack of fit (Rochon *et al.*, 2012). Still, the problems arising from misspecified variances in regression can be as severe in regression as they are in the two-group model, especially when the variance depends on the covariates in the model in an unknown way: standard errors are too large or too small, and statistical tests can become severely conservative or anti-conservative.

The situation is worse if we broaden the perspective to generalized linear models (GLMs). Overdispersed GLMs, that allow for additional variation between subjects, have become important in many fields; e.g. negative binomial or quasi-Poisson models in RNA sequencing (Love *et al.*, 2014). GLMs and overdispersed GLMs generally use a single dispersion parameter for all subjects. This assumption is similar to the common variance assumption in the linear model, and violation of the assumption can lead to loss of Type I error control in the same way. The diagnostics to detect a failure of this variance model are more complex in GLMs than in the linear model, since the variance of residuals differs between observations even in a correctly specified model. A general approach that is asymptotically robust to variance misspecification is to use the Sandwich estimator (Eicker, 1967; Huber, 1967; White, 1982). However this is not always an optimal choice, since it often performs poorly for small sample sizes (Boos, 1992; Kauermann and Carroll, 2000; Maas and Hox, 2004; Freedman, 2006).

The issues of the univariate setting can dramatically propagate in multivariate settings. Indeed, the modeling of high-dimensional data represents a stimulating challenge nowadays in many ambit of applications, as neuroimaging and transcriptomics, where it is common to fit several generalized linear regression models in parallel, each with a small sample size (Schaarschmidt *et al.*, 2022; Love *et al.*, 2014; Winkler *et al.*, 2014). The goal of the analysis is to perform hypothesis testing to find relevant associations and the multiplicity issue must be taken into account properly. Several challenges are present: the availability of small sample sizes can make the classic tests unreliable, especially for nonlinear models where exact methods are not available and we always rely on asymptotic arguments. Indeed, the classic normal approximation of the test statistic can be quite unreliable (Schaarschmidt *et al.*, 2022).

Further, the test statistics are often correlated, due to correlations in the underlying biological measurements, and taking correlations into account could increase the power of the test procedure. Indeed, the classic multiple testing corrections (as Bonferroni-Holm) are designed in order to protect against any correlation structure and can be very conservative (Goeman and Solari, 2014). Taking correlations into account could increase the power of the test procedure.

Main contributions of the thesis

In this work we will develop a novel method for inference in generalized linear models (GLMs), Cox regression models and multivariate extension. The proposed methodology is based on sign-flip tests, whose key idea consists in multiplying appropriate quantities by a vector of 1 and -1, in order to obtain a conditional distribution of test statistics (Hemerik and Goeman, 2018). Under some conditions, which must be properly checked, the derived conditional distribution permits to perform valid hypothesis testing.

We will first focus on univariate testing, proposing a method for regression coefficients in GLMs that is robust to misspecification of the variance. Our method is an extension of the method of Hemerik *et al.* (2020), which compares the score to a semiparametric reference distribution calculated by randomly sign-flipping each subject's contribution to the score. This method was known to have robustness to misspecification of the variance by a common multiplicative constant only. We extend this method in two ways. First, we derive an expression for the variance of the flipped score as depending on which of the 2^n possible transformations has been used. We use these expressions to standardize the flipped scores, to obtain a test with faster convergence to the nominal Type I error level. Second, we show that the new standardized test remains asymptotically valid under any misspecification of the variances of the residuals. The resulting test gives valid inference without concern for heteroscedasticity. We find that the convergence to the nominal level is comparable, and sometimes even faster than a parametric test when both methods are asympotically valid, and power is also comparable. Our results actually imply that the same robustness holds for the original test of Hemerik *et al.* (2020), but the new test achieves better error control than the original in small samples.

In the more limited context of linear models, other semiparametric approaches have been proposed before. Winkler *et al.* (2014) provides a systematic comparison of such methods for linear regression in a realistic setting. Our novel method has a connection to the Wild Bootstrap (Davidson and Flachaire, 2008), which addresses the problem of heteroskedasticity only in linear regression models without nuisance parameters.

Further, we start from this univariate test to build a multiple testing procedure based on the max-T method of (Westfall and Young, 1993), which guarantees strong control of the family wise error rate (FWER), a key property which guarantees valid p-values even for postponed choice of the model after seeing the data. Remarkably, the proposed method adapts to the unknown correlation structure. It is especially useful when strong correlation between the individual test statistics is present. In that situation it guarantees a relevant gain in power over alternative methods, as we will show in a simulation.

Finally, we will go beyond generalized linear models by extending the methodology to Cox regression models. We will see that the standard parametric tests suffer of lack of type I error control for small sample size, and we will apply the key idea of sign-flip tests to improve the convergence. We will derive two alternative tests which both show a significant improvement over the parametric tests. Further, from the previous section is then straightforward to obtain a multivariate extension also for this class of models.

Chapter 1

Inference in generalized linear models with robustness to misspecified variances

1.1 Generalized linear models

Assume that we observe n independent observations $Y = \{Y_1, \ldots, Y_n\}^T$ from the exponential dispersion family, i.e. a density of the form (Agresti, 2015)

$$f(y_i; \theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right\},$$

where θ_i and ϕ_i are respectively the canonical and the dispersion parameter. We will assume throughout the chapter that the model chosen fulfills the usual regularity conditions (Azzalini, 1996; Hall and Mathiason, 1990). We derive mean and variance of the observed outcome, respectively, as

$$\mu_i = \mathbb{E}[Y_i] = b'(\theta_i); \qquad \mathbb{V}[Y_i] = b''(\theta_i)a(\phi_i),$$

where primes denote derivatives. We will consider through the chapter, without loss of generality, $a(\phi_i) = \phi_i$. We assume that the mean of Y_i depends on observed covariates (X_i, Z_i) through the following relation, written in vector-matrix form as

$$g(\mu) = \eta = X\beta + Z\gamma$$

where $\mu = (\mu_1, \ldots, \mu_n)^T$ denotes the mean vector, $g(\cdot)$ is the link function, taken to operate elementwise on a vector, (X, Z) is the full rank design matrix with $\dim\{X\} =$

$n \times 1$, $\dim\{Z\} = n \times q$, where q does not depend on n, and (β, γ) are unknown parameters. We consider β the parameter of interest, and γ a nuisance parameter.

Let $\ell(\beta, \gamma)$ denote the log-likelihood function, from which we can derive the score vector with elements

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta} = \ell_\beta = X^T D V^{-1}(Y - \mu); \qquad \frac{\partial \ell(\beta,\gamma)}{\partial \gamma} = \ell_\gamma = Z^T D V^{-1}(Y - \mu),$$

where, in a compact matrix form, we have

$$D = \operatorname{diag}\left\{\frac{\partial \mu_i}{\partial \eta_i}\right\}; \qquad V = \operatorname{diag}\left\{\mathbb{V}[Y_i]\right\}.$$

Taking the derivatives of the score vector we obtain the Fisher information matrix

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\beta,\beta} & \mathcal{I}_{\beta,\gamma} \\ \mathcal{I}_{\gamma,\beta} & \mathcal{I}_{\gamma,\gamma} \end{pmatrix} = \begin{pmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z \end{pmatrix}$$

where $W = D V^{-1} D$. Denote by d_i, v_i, and w_i the i-th diagonal elements of D, V and W, respectively. Note that all the matrices defined here and in the rest of the chapter depend on n. We will suppress this dependence in the notation. The reader may assume that all quantities depend on n, except when explicitly stated otherwise.

The model as defined above has separate dispersion parameters ϕ_i for $i = 1, \dots, n$. These parameters cannot be consistently estimated unless they satisfy some restrictions (Neyman and Scott, 1948). Thus, in practice, the analyst will use some putative model, which imposes such restrictions. This leads to computable estimates $\hat{\phi}_1, \dots, \hat{\phi}_n$, which will generally be inconsistent. Remarkably, the incorrect estimation of the dispersion parameters does not affect the consistency of the estimates of the regression coefficients, as long as the mean (and hence the link function) are well-specified (Agresti, 2015). The corresponding standard errors, however, are no longer reliable.

As an example of the set-up, consider the situation that the data are generated according to a negative binomial model with a log link function and a different and unknown dispersion parameter for every observation. The analyst could estimate these dispersion parameters under the additional constraint that $\phi_i = \phi$ for all i, or even that $\phi_i = 1$ for all i. Alternatively, the dispersion parameters could be modeled as a function of the covariates (McCullagh and Nelder, 1989, Chapter 10). Such strategies will lead to inconsistent estimates of the dispersion parameters unless the true model happens to fulfil the chosen constraints (White, 1982).

About the model and the estimation strategy we make the following assumptions.

First, we demand that the regression parameters β and γ are consistently estimated, and the true variances to be uniformly bounded. Secondly, we assume that the estimates of $\phi_1, \phi_2, \ldots$, though not consistent, will converge, as stated more formally below.

Assumption 1. For $i = 1, 2, \ldots$, we have $\hat{\phi}_i - \tilde{\phi}_i \to 0$ in probability as $n \to \infty$, where $K_1 < \tilde{\phi}_i \leq K_2$, and K_1, K_2 are strictly positive constants not depending on n.

It is known (Huber, 1967) that the use of maximum likelihood estimation in a misspecified model leads under minimal conditions to a well-defined limit of the estimator. In this sense, it is possible to define a "true" density, even in a misspecified model, which is intended to be the closest to the true density which generates the data in terms of Kullback-Leibler distance (White, 1982). This implies that Assumption 1 generally holds in situations where we estimate $\phi_1, \ldots, \phi_n$ using a restricted model.

We place a tilde symbol on all the quantities obtained when plugging the limits $\tilde{\phi}_1, \ldots, \tilde{\phi}_n$ into the model. In particular, let $\tilde{V}, \tilde{W}$ be the variance and weight matrices obtained by fixing $\phi_i = \tilde{\phi}_i$. Note that the putative model only misspecifies the variance, but not the mean: the mean vector and the link function remain correctly specified and can still be consistently estimated (Agresti, 2015).

We further assume the following assumption related to the Fisher information given by the quantities $\tilde{\phi}_1, \ldots, \tilde{\phi}_n$. This is a standard condition in regular models (Van der Vaart, 1998).

Assumption 2. Let $\tilde{\mathcal{I}}$ be the Fisher information matrix for $W = \tilde{W}$. The $\lim_{n\to\infty} n^{-1}\tilde{\mathcal{I}}_{\beta,\beta}$ converges to some positive constant.

Further, the following mild condition is needed to apply the multivariate central limit theorem (Billingsley, 1986, Chapter 5) within the framework of Hemerik *et al.* (2020), that we will apply. It is needed to avoid pathological cases, such as vanishing or dominating observations.

Assumption 3. Let

$$\tilde{\nu}_{i,\beta} = \frac{(Y_i - \hat{\mu}_i)x_i d_i}{\tilde{v}_i}; \qquad \tilde{\nu}_{i,\gamma} = \frac{(Y_i - \hat{\mu}_i)z_i d_i}{\tilde{v}_i};$$

define the element-wise score contribution and

$$\tilde{\nu}^*_{i,\beta} = \tilde{\nu}_{i,\beta} - \tilde{\mathcal{I}}_{\beta,\gamma}\tilde{\mathcal{I}}^{-1}_{\gamma,\gamma}\tilde{\nu}_{i,\gamma}.$$

We require, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(\tilde{\nu}^*_{i,\beta})^2 \mathbf{1}_{\{|\tilde{\nu}^*_{i,\beta}|/\sqrt{n} > \epsilon\}}] \to 0$$

where $\mathbf{1}_{\{\dots\}}$ is the indicator function, and $n^{-1}\sum_{i=1}^{n}\mathbb{V}[\tilde{\nu}_{i,\beta}^{*}]$ to converge to some positive constant.

When the model is correctly specified the last condition implies Assumption 2. In case of misspecified models we do not have this relation, since the so called "information identity" does not hold (Azzalini, 1996).

Throughout the chapter, we will make a distinction between two situations: the situation that the model is correctly specified, i.e. $\tilde{W} = W$, and the putative model is true; or the general case that $\tilde{W} \neq W$.

1.2 Classic approach to hypothesis testing

In the model described in Section 1.1 we are interested in testing the following null hypothesis

$$H_0 :\ \beta = \beta_0 \mid (\gamma, \phi) \in \Gamma \times \Phi \tag{1.1}$$

against a one or two-sided alternative, where we are considering $\dim\{\beta\} = 1$, while $\Gamma \subseteq \mathbb{R}^q$ and $\Phi \subseteq (0, \infty)$. Indeed, the classic approach to GLM assumes ϕ to be known, or unknown but equal among units. When $q = 0$ the observations are identically distributed under the null hypothesis. A large number of tests exist in the literature, as the well-known parametric t-test or some nonparametric alternatives as sign-flip or rank tests (Lehmann and Romano, 2005). When $q > 0$ the observations are no more exchangeable and nonparametric solutions are generally not available. In the literature there are three different approaches to hypothesis testing which are fully parametric, in the sense that they fully rely on the assumptions made in the model's choice (see for instance Azzalini (1996)). Even if they are dated, their use is widespread and they are the default choices in the most popular statistical software. For two-sided alternative they are defined as follows; the Wald test statistic (Wald, 1943) is defined as

$$W_e = (\beta - \beta_0)^T \left(\mathcal{I}^{\beta,\beta}\right)^{-1} (\beta - \beta_0)$$

where $\mathcal{I}^{\beta,\beta}$ is the corresponding element of the inverse of the Fisher information matrix; the parametric Score test statistic (Rao, 1948) which is defined as

$$W_u = \ell_{\beta_0}^T \left(\mathcal{I}^{\beta_0,\beta_0}\right) \ell_{\beta_0}$$

where the symbol β_0 means that the corresponding quantities are evaluated under the null model; and the Likelihood ratio test statistic (Wilks, 1938) is

$$W_l = 2(\ell(\beta) - \ell(\beta_0)).$$

The three test statistic tend to perform similarly and they share the asymptotic distribution which is chi-squared with one degree of freedom. Moreover, when β is a scalar it is possible to define an alternative version of these test statistics that have a standard normal distribution, which are also suitable for testing one-sided alternatives, i.e.

$$\begin{aligned} r_e &= (\beta - \beta_0)(\mathcal{I}^{\beta,\beta})^{-1/2} \\ r_u &= \ell_{\beta_0}(\mathcal{I}^{\beta_0,\beta_0})^{1/2} \\ r_l &= \text{sign}\{\beta - \beta_0\}\sqrt{2(\ell(\beta) - \ell(\beta_0))}. \end{aligned}$$

However, these tests fully rely on the assumptions behind the model choice, and they can perform poorly in presence of variance misspecification or, more in general, for small sample sizes, as we will see in Section 1.7. It is important to note that there are some more recent methods specific to some particular models. For instance, the linear regression model has generate a huge literature (see Winkler *et al.* (2014) and references therein), while the slow convergence of the maximum likelihood estimator - particularly relevant for the logistic regression model (see also the Simulation in Section 1.7) - has been addressed by Firth (1993). However, in this chapter we will focus on methods valid for the whole ambit of generalized linear models with attention to variance misspecification.

A popular answer to protect against variance misspecification is based on the use of the parametric tests adopting the sandwich estimator of the variance (Eicker, 1967; Huber, 1967; White, 1982). This approach is based on the failure of Bartlett second identity when the model is wrongly specified, and it estimates empirically the true variance. Then the significance of the coefficients is tested by means of a Wald-type test with plugged-in this robust estimator. However, we will see in the Simulation study of Section 1.7 that it performs poorly for small sample sizes, even when the model is correctly specified (Boos, 1992; Freedman, 2006; Maas and Hox, 2004; Kauermann and Carroll, 2000). This negative aspect has two main causes; the presence of bias for finite sample size, and the fact that it is often far more variable than the usual parametric estimator.

More complex models tend to jointly model the mean and the variance of the observations. In particular we will analyze the behavior of the generalized additive models for location, scale and shape (Gamlss; Rigby and Stasinopoulos, 2005). They permit to model the mean specifying the family of the response and the link function (as for GLMs), but in addition they model also the variance as function of any potential covariate of interest. A Wald-type test is used to test the significance of the coefficients. However, we will see in Section 1.7 that this class of models tend to show slow convergence, due to the overparametrization, and moreover they are not always able to catch the variance misspecification.

1.3 Sign-flipping effective score test

As introduced in section 1.2 we are interested in testing the following null hypothesis

$$H_0: \ \beta = \beta_0 \mid (\gamma, \phi_1, \dots, \phi_n) \in \Gamma \times \Phi \times \dots \times \Phi \tag{1.2}$$

against a one or two-sided alternative, where $\Gamma \subseteq \mathbb{R}^q$ and $\Phi \subseteq (0, \infty)$. This formulation of the null hypothesis above makes explicit that only the target parameter β is fixed under H_0, but the nuisance parameters are unconstrained.

Since $\phi_1, \dots, \phi_n$, as remarked, are difficult to estimate, we require a test that is robust to misspecification of these parameters. Hemerik *et al.* (2020) proposed a general semiparametric test that has robustness to misspecification of the variance by a constant, and presented promising simulation results suggesting more general robustness properties. We will start from this test.

Hemerik *et al.* (2020) uses the effective score for β as the test statistic. In the model (1.1) the effective score is

$$S = \ell_\beta - \mathcal{I}_{\beta,\gamma}\mathcal{I}_{\gamma,\gamma}^{-1}\ell_\gamma.$$

In the context of generalized linear models, the statistic can be written as

$$S = n^{-1/2} X^T W^{1/2} (I - H) V^{-1/2} (Y - \hat{\mu})$$

where

$$H = W^{1/2} Z (Z^T W Z)^{-1} Z^T W^{1/2}, \tag{1.3}$$

is the hat matrix, and $\hat{\mu}$ is the vector of fitted values of the model under the null hypothesis. Note that, since S is an inner product of the two n-vectors $n^{-1/2}V^{-1/2}(I - H)W^{1/2}X$ and $y - \hat{\mu}$, it can be written as a sum of n terms, which we call the effective score contributions.

To calculate the critical value, Hemerik *et al.* (2020) proposed to use sign-flips of the effective score contributions, randomly multiplying each score contribution by -1 or 1. In matrix notation these sign flips can be represented by a random diagonal matrix $\mathcal{F}$ of dimension n, whose non-zero elements are independent random variables that take values -1 and 1 with equal probability. Consequently, the effective sign-flip score statistic for a given flip matrix $\mathcal{F} = F$ is defined as

$$S(F) = n^{-1/2} X^T W^{1/2} (I - H) V^{-1/2} F (Y - \hat{\mu}). \tag{1.4}$$

where, with a little abuse of notation, we write $S(F) = S(\mathcal{F} | \mathcal{F} = F, Y)$. Therefore we fix F, but not Y hence $S(F)$ is still a random variable. Note that for $\mathcal{F} = \mathbf{I}$ (the identity matrix) we recover the observed effective score.

An asymptotic α level test is then derived as follows. First, Hemerik *et al.* (2020) prove (asymptotic) invariance of the first two moments of the test statistic under the action of $\mathcal{F}$, i.e. that $\mathbb{E}[S(\mathcal{F})] - \mathbb{E}[S(\mathbf{I})] = 0$ and $\mathbb{V}[S(\mathcal{F})] - \mathbb{V}[S(\mathbf{I})] \to 0$ as $n \to \infty$. Note that $S(\mathcal{F})$ depends on two random variables, Y and $\mathcal{F}$, and we do not condition to any of them unless it is explicitly stated. Next, they apply the Lindeberg-Feller multivariate central limit theorem to show that, for independently drawn flip matrices $\mathcal{F}_2, \ldots, \mathcal{F}_g$, where g does not depend on n, the vector $S(\mathbf{I}), S(\mathcal{F}_2), \ldots, S(\mathcal{F}_g)$ converges to a vector of independent and identically distributed random variables. By Lemma 1 of Hemerik *et al.* (2020) is then possible to obtain an asymptotic α level test for the null hypothesis (1.2) against a one or two-sided alternative. Assume we observe values $S_1, \ldots, S_g$, where $S_1 = S(\mathbf{I})$ is the observed test statistic, while the sorted values are $S_{(1)} \leq \cdots \leq S_{(g)}$. Without loss of generality, consider testing (1.2) against $H_1 : \beta > \beta_0$. The test that they consider rejects the null hypothesis if

$$S_1 > S_{(\lceil (1-\alpha) g \rceil)} \tag{1.5}$$

where $\lceil \cdot \rceil$ represents the ceiling function. Analogous procedures can be defined for $H_1 : \beta < \beta_0$ or $\beta \neq \beta_0$ (Hemerik *et al.*, 2020).

The definition of the test involves W, which is unknown. In practice we only have an estimate of W available, which converges to $\tilde{W}$ by Assumption 1. For the theoretical results of the remainder of this chapter, we will treat $\tilde{W}$, though not W, as known. To motivate this, we note, in the first place, that in case of a correctly specified model any error terms relating to estimation of $\tilde{W}$ are of lower asymptotic order (Barndorff-Nielsen and Cox, 1994) with respect to the parameters estimation. Secondly, as we will show below in Section 1.5, the validity of the test is invariant to misspecification of $\tilde{W}$, so it is not too relevant to estimate it correctly.

1.4 Standardized sign-flips score test

The test based on the effective score is proven to be asymptotically exact by Hemerik *et al.* (2020), but simulations in that paper show an anti-conservative behavior for small sample sizes. This problem relates to a lack of null-invariance of the distribution of the test statistic. In this section we revisit the concept of null-invariance in some detail. This discussion will motivate our modification of the test of Hemerik *et al.* (2020).

Tests based on fixed data transformations such as sign-flips or permutations can have exact control of the type I error rate if the observed test statistic $S(\mathbf{I})$ and the flipped $S(F)$ have the same distribution for every F. If this property holds, we say that the transformations are null-invariant. Hemerik and Goeman (2018) showed how null-invariance can be used to construct tests with exact control of the type I error, where they highlight the importance of the group structure. Null-invariance can be achieved in linear models (Solari *et al.*, 2014; Huh and Jhun, 2001; Kherad-Pajouh and Renaud, 2010) or in GLMs with simple experimental designs. With more complex designs, ad hoc solutions sometimes exist, e.g. when all covariates are discrete (Pesarin, 2001). For general GLMs, null-invariant transformations do not exist.

When the equality in distribution holds only for a random flipping matrix $S(\mathcal{F})$ the property of null-invariance has still a major role, but it must be combined with some extra conditions. A sufficient one is outlined in Lemma 1 of Hemerik *et al.* (2020), which uses the equality in distribution for any random sign-flip in addition to a Lindeberg condition required to apply an appropriate version of the central limit theorem. The resulting test is not more an invariance test and the group structure becomes less important.

In the case of the test of Hemerik *et al.* (2020), null-invariance is replaced by an equality in distribution that holds only asymptotically. We restate the result concerning the validity of that test below for the special case of GLMs. Our alternative proof emphasizes the importance of asymptotic null-invariance.

We first introduce some Lemmas needed for the proofs of the Theorems; the projection matrix for GLMs introduced in (1.3) is a proper projection matrix for studentized units as shown in the following Lemma, which is mentioned for instance in (Agresti, 2015, p. 136). We give a full proof for the sake of completeness, since that is omitted in most textbooks.

Lemma 1. The fitted and observed values in a GLM are connected through the relation

$$V^{-1/2}(\hat{\mu} - \mu) = HV^{-1/2}(Y - \mu)\{1 + o_p(1)\}$$

Proof. The first-order approximation of the score function is

$$\begin{aligned}
\ell_{\hat{\beta}} &= \ell_{\beta} + \mathcal{I}_{\beta,\beta}(\hat{\beta} - \beta) + o_p(1) \\
0 &= X^T D V^{-1}(Y - \mu) - X^T W X(\hat{\beta} - \beta) + o_p(1) \\
\hat{\beta} - \beta &= (X^T W X)^{-1} X^T D V^{-1}(Y - \mu) + o_p(1)
\end{aligned}$$

where $(o_p(1))$ is an error term asymptotically negligible. Then, by the Delta method we have

$$\begin{aligned}
\hat{\mu} &= \mu + \frac{d\mu}{d\eta}\frac{d\eta}{d\beta}(\hat{\beta} - \beta)\{1 + o_p(1)\} \\
\hat{\mu} - \mu &= D X^T(\hat{\beta} - \beta)\{1 + o_p(1)\} \\
&= D X^T (X^T W X)^{-1} X^T D V^{-1}(Y - \mu)\{1 + o_p(1)\} \\
&= D W^{-1/2} H W^{1/2} D^{-1}(Y - \mu)\{1 + o_p(1)\} \\
&= V^{1/2} H V^{-1/2}(Y - \mu)\{1 + o_p(1)\} \\
V^{-1/2}(\hat{\mu} - \mu) &= H V^{-1/2}(Y - \mu)\{1 + o_p(1)\}.
\end{aligned}$$

where the asymptotically negligible error term has two sources, one related to the second-order approximation of the likelihood and the other related to possible non-linearity of the link function. □

Lemma 2. Let C be any n-dimensional matrix and G be a nonsingular $n \times n$ matrix, then C and $G^{-1}CG$ have the same set of eigenvalues (with the same multiplicities).

Proof. See (Magnus and Neudecker, 2019, p. 15). □

Lemma 3. Let C be any n-dimensional matrix and $\mathbb{F}$ be the set of all n-dimensional flipping matrices, i.e. the set of all possible n-dimensional diagonal matrices F with elements -1 or 1. Then

$$\sum_{F \in \mathbb{F}} FCF = 2^n \text{diag}\{C\},$$

where diag$\{C\}$ is a diagonal matrix with the same diagonal elements of C.

Proof. First we note that the absolute value of each element of C does not change after the multiplication FCF. The sums for the off-diagonal elements contain an equal number of terms with positive and negative sign, hence their sum over all possible flips is zero, while the sign of each diagonal element is positive for each term. By noting that the total number of flips is 2^n we have the claim. □

The following result was adapted from Huber and Ronchetti (2009), extending their theorem (Proposition 7.1, p. 156) for linear regression models to GLMs.

Lemma 4. Assume that the regression coefficients of a generalized linear model are consistently estimated, in the sense that for every $\varepsilon > 0$, as $n \to \infty$,

$$\max_{1\leq i\leq n} \mathbb{P}(|\hat{\mu}_i - \mu_i| > \varepsilon) \to 0.$$

Then

$$\max_{1\leq i\leq n} h_{ii} \to 0,$$

where h_{ik} is the ik-th element of the matrix H, as defined in (1.3).

Proof. Using Lemma 1 we have, for each i, $\hat{\mu}_i - \mu_i =$

$$h_{ii}(Y_i - \mu_i) + \sum_{k\neq i} v_i^{1/2} v_k^{-1/2} h_{ik}(Y_k - \mu_k) + o_p(1). \tag{1.6}$$

Now we give a general probability result. Let V_1 and V_2 be two independent random variables, for any $\varepsilon > 0$ we have that

$$\begin{aligned}\mathbb{P}(|V_1 + V_2| \geq \varepsilon) &\geq \mathbb{P}(V_1 \geq \varepsilon)\mathbb{P}(V_2 \geq 0) + \mathbb{P}(V_1 \leq -\varepsilon)\mathbb{P}(V_2 \leq 0)\\ &\geq \min\left\{\mathbb{P}(V_1 \geq \varepsilon), \mathbb{P}(V_1 \leq -\varepsilon)\right\}.\end{aligned}$$

Noting that $(y_i - \mu_i)$ is independent from $(y_k - \mu_k)$ for each $k \neq i$ we can apply the result to expression (1.6) to obtain

$$\mathbb{P}\left(|\hat{\mu}_i - \mu_i \geq \varepsilon|\right) \geq \min\left[\mathbb{P}\left\{(Y_i - \mu_i) \geq \frac{\varepsilon}{h_{ii}}\right\}, \mathbb{P}\left\{(Y_i - \mu_i) \leq -\frac{\varepsilon}{h_{ii}}\right\}\right] = m_i^n.$$

We have $\max_{1\leq i\leq n} \mathbb{P}(|\hat{\mu}_i - \mu_i| > \varepsilon) \to 0$, so $\max_{1\leq i\leq n} m_i^n \to 0$. Since ε was arbitrary, this implies that $\max_{1\leq i\leq n} h_{ii} \to 0$. □

We can now turn our attention on the theorem which focuses on the test proposed in Hemerik *et al.* (2020), showing its asymptotic null-invariance.

Theorem 1. Assume that the variances are correctly specified, that is, $\tilde{V} = V$, and that Assumption 1-3 hold. For $n \to \infty$, the test that rejects H_0 if (1.5) holds is an asymptotically α level test.

Proof. By the definition of the random sign-flipping transformations is trivial to observe that the expected value of the test statistic $S(\mathcal{F})$ is zero.

By Lemma 1 we can rewrite the effective score statistic (1.4) as

$$S(\mathcal{F}) = n^{-1/2} X^T W^{1/2} (I - H) \mathcal{F} (I - H) V^{-1/2} (Y - \mu) + o_p(1).$$

Let

$$a = (I - H) W^{1/2} X. \tag{1.7}$$

The variance conditional on $\mathcal{F} = F$ is

$$\begin{aligned} \mathbb{V}\left[S(F)\right] &= n^{-1} a^T F (I - H) V^{-1/2} \mathbb{E}[(Y - \mu)(Y - \mu)^T] V^{-1/2} (I - H) F a + o(1) \\ &= n^{-1} a^T F (I - H) V^{-1/2} V V^{-1/2} (I - H) F a + o(1) \\ &= n^{-1} a^T F (I - H) F a + o(1) \end{aligned}$$

and for $\mathcal{F} = \mathbf{I}$ we have

$$\mathbb{V}\left[S(\mathbf{I})\right] = n^{-1} a^T I a + o(1).$$

Taking the difference

$$\mathbb{V}\left[S(\mathbf{I})\right] - \mathbb{V}\left[S(F)\right] = n^{-1} a^T \left\{ I - F(I - H) F \right\} a + o(1),$$

we note that the first term is a quadratic form and we look at the matrix

$$I - F(I - H)F = I - FF + FHF = FHF.$$

Since H is a projection matrix and $F^{-1} = F$, Lemma 2 implies that FHF is positive semidefinite, so that asymptotically $\mathbb{V}\left[S(\mathbf{I})\right] - \mathbb{V}\left[S(F)\right] \geq 0$.

Define

$$h_{\text{sup}} = \sup_{1 \leq i \leq n} \left\{ h_{ii} \right\}.$$

Let us make the randomness of the flips explicit. Let $\mathbb{F}$ denote the set of all possible flipping matrices, and note that $|\mathbb{F}| = 2^n$. By Lemma 3 we have

$$\begin{aligned} \mathbb{E}\left[\mathbb{V}\left\{S(\mathbf{I})\right\} - \mathbb{V}\left\{S(\mathcal{F}) \mid F\right\}\right] &= (2^n)^{-1} \sum_{F \in \mathbb{F}} [\mathbb{V}\left\{S(\mathbf{I})\right\} - \mathbb{V}\left\{S(F)\right\}] \\ &= n^{-1} a^T \text{diag}\{H\} a + o(1) \\ &\leq n^{-1} \|a\|^2 \cdot h_{\text{sup}} + o(1), \end{aligned}$$

where, by Assumption 2 and Lemma 4, the limiting behavior is

$$\lim_{n \longrightarrow \infty} n^{-1} \|a\|^2 \cdot h_{\text{sup}} = 0.$$

According to the law of total variance we have

$$\mathbb{V}[S(\mathcal{F})] = \mathbb{V}\left[\mathbb{E}\left\{S(\mathcal{F}) \mid F\right\}\right] + \mathbb{E}\left[\mathbb{V}\left\{S(\mathcal{F}) \mid F\right\}\right].$$

We know that $\mathbb{E}\{S(\mathcal{F}) \mid F\}$ does not depend on $\mathcal{F}$, which means that $\mathbb{V}[\mathbb{E}\{S(\mathcal{F}) \mid F\}] = 0$, so $\mathbb{V}\{S(\mathcal{F})\} = \mathbb{E}[\mathbb{V}\{S(\mathcal{F}) \mid F\}]$. It follows that, marginally over F,

$$\lim_{n \to \infty} \mathbb{V}[S(I)] - \mathbb{V}[S(\mathcal{F})] = 0.$$

Since the random sign-flipping transformations are all independent, the corresponding test statistics are all uncorrelated, i.e, for all $1 \leq w < v \leq g$ we have $\mathbb{C}[S(\mathcal{F}_w), S(\mathcal{F}_v)] = 0$ for two random flips $\mathcal{F}_w, \mathcal{F}_v$. By the previous results and Assumption 3, $(S(\mathbf{I}), \ldots, S(\mathcal{F}_g))^T$ converges to a multivariate normal distribution by the multivariate Lindberg-Feller central limit theorem (Van der Vaart, 1998), with mean vector $\mathbf{0}$ and covariance matrix $s^2\mathbf{I}$, where s^2 is the limiting variance of the flipped test statistic. Finally, we use Lemma 1 of Hemerik *et al.* (2020) to conclude that the test that rejects when (1.5) holds is an asymptotic α level test. □

If null-invariance holds only asymptotically, the distributions of $S(\mathbf{I})$ and $S(\mathcal{F})$ are not identical, but both converge to the same distribution. It is natural to suppose that the more aspects of the distributions of $S(\mathbf{I})$ and $S(\mathcal{F})$ are equal to each other in all finite samples, the closer the resulting test is to an exact test. This intuition motivated (Commenges, 2003; Solari *et al.*, 2014) to formulate the concept of second-moment null-invariance. A random transformation $\mathcal{F}$ is second-moment null-invariant if $S(\mathbf{I})$ and $S(\mathcal{F})$ have identical first and second moment. For second-moment null-invariant transformations, asymptotics is only needed for the convergence of the higher-order moments.

The test of Hemerik *et al.* (2020) does not have the second-moment null-invariance property: though the first moments of $S(\mathbf{I})$ and $S(\mathcal{F})$ are equal in finite samples, the variances are not. In fact, we prove in Theorem 2 below that for finite samples $S(\mathbf{I})$ always has a larger variance than $S(\mathcal{F})$, at least in the linear model. As a consequence, the test shows a tendency to anti-conservativeness in finite samples. Since the variance of the observed test statistic is always larger than its flipped counterpart, extreme values in $S(\mathbf{I})$ are more probable than in the reference distribution of $S(\mathcal{F})$. Consequently, the test tends to reject the null hypothesis too often.

Proposition 1. Consider a normal regression model with identity link. Assume that the variances are correctly specified, that is, $\tilde{V} = V$, and that Assumption 1-3 hold. For finite sample size, the effective sign-flip score statistic defined as in (1.4) has $\mathbb{V}[S(\mathbf{I})] > \mathbb{V}[S(\mathcal{F})]$.

Proof. By Lemma 1 we can rewrite the effective score statistic (1.4) as

$$S(\mathcal{F}) = n^{-1/2}X^TW^{1/2}(I-H)\mathcal{F}(I-H)V^{-1/2}(Y-\mu) + o_p(1).$$

The variance conditional on $\mathcal{F} = F$ is

$$\begin{aligned}\mathbb{V}\left[S(F)\right] &= n^{-1}a^TF(I-H)V^{-1/2}E\{(Y-\mu)(Y-\mu)^T\}V^{-1/2}(I-H)Fa\\ &= n^{-1}a^TF(I-H)V^{-1/2}VV^{-1/2}(I-H)Fa\\ &= n^{-1}a^TF(I-H)Fa\end{aligned}$$

and for $\mathcal{F} = \mathbf{I}$ we have

$$\mathbb{V}\left[S(\mathbf{I})\right] = n^{-1}a^TIa.$$

Taking the difference

$$\mathbb{V}\left[S(\mathbf{I})\right] - \mathbb{V}\left[S(F)\right] = n^{-1}a^T\left\{I - F(I-H)F\right\}a,$$

we note that it is a quadratic form and we consider the matrix

$$I - F(I-H)F = I - FF + FHF = FHF.$$

Since H is a projection matrix and $F^{-1} = F$, Lemma 2 implies that FHF is positive semidefinite, and therefore $\mathbb{V}[S(\mathbf{I})] - \mathbb{V}[S(F)] \geq 0$.

We then prove the strict inequality for at least one flipping matrix F. Taking any model with an intercept, by construction

$$h_{\inf} = \inf_{1\leq i\leq n}\{h_{ii}\} > 0.$$

Note that $|\mathbb{F}| = 2^n$. Since $\mathbb{V}[S(\mathbf{I})] - \mathbb{V}[S(F)] \geq 0$ as proven above, it suffices to show that

$$\sum_{F\in\mathbb{F}} n^{-1}a^TFHFa > 0,$$

which means that we have a strictly inequality for some F. Using Lemma 3 we have

$$\sum_{F\in\mathbb{F}} FHF = 2^n\,\mathrm{diag}\{H\}.$$

Therefore

$$\begin{aligned}\sum_{F\in\mathbb{F}} n^{-1}a^T FHFa &= 2^n n^{-1} a^T \text{diag}\{H\}a \\ &\geq 2^n n^{-1}\|a\|^2 \cdot h_{\text{inf}} > 0.\end{aligned}$$

□

Proposition 1 is formulated for the linear model only. For the general case the same positive difference between the variance of $S(\mathbf{I})$ and $S(\mathcal{F})$ is present, but it is not the only term in the asymptotic expansion. However, we see a tendency to small sample anti-conservativeness in all simulations, as was also observed by Hemerik *et al.* (2020).

The concept of second-moment null-invariance suggests that we can improve level accuracy of the test if we can modify the flipped scores to have equal variances. The following result allows for such a procedure. It provides an expression for the variance of the flipped score, depending on the sign flip that has been applied.

Lemma 5. The variance of the sign-flipped score, as depending on F, is

$$\mathbb{V}[S(F)] = n^{-1}X^T W^{1/2}(I-H)F(I-H)F(I-H)W^{1/2}X + o(1).$$

Proof. Let

$$a = (I-H)W^{1/2}X.$$

The variance for a given sign-flix matrix F is

$$\begin{aligned}\mathbb{V}\left[S(F)\right] &= n^{-1}a^T F(I-H)V^{-1/2}\mathbb{E}[(Y-\mu)(Y-\mu)^T]V^{-1/2}(I-H)Fa + o(1) \\ &= n^{-1}a^T F(I-H)V^{-1/2}VV^{-1/2}(I-H)Fa + o(1) \\ &= n^{-1}a^T F(I-H)Fa + o(1)\end{aligned}$$

□

These variances can be estimated by plugging in $\hat{\gamma}$. By dividing the flipped scores by their standard deviations, we obtain what we call the standardized sign-flip score statistics,

$$S^*(F) = S(F)/\mathbb{V}[S(F)]^{1/2}. \tag{1.8}$$

We use the statistics $S^*(F)$ in the same way as the original test uses the statistics $S(F)$. The estimate of $\mathbb{V}[S(F)]$ can be calculated for each F in linear time in n, as we show in the following Lemma.

Lemma 6. The computational cost of the standardization constant in (1.8) is linear in n.

Proof. Define $W^{1/2}Z = U\Delta L^T$, that is, the singular value decomposition of $W^{1/2}Z$, where U is a semiorthogonal $n \times q$ matrix (q equal to the rank of Z), Δ a diagonal q matrix and L a $q \times q$ orthogonal matrix. Therefore (1.3) can be written as

$$H = W^{1/2}Z(Z^TWZ)^{-1}Z^TW^{1/2} = U\Delta L^T(L\Delta U^TU\Delta L^T)^{-1}L\Delta U^T = UU^T.$$

Now, let $a = (I-H)W^{1/2}X$ and $A = \mathrm{diag}\{a\}$. Further, let $\mathbf{1}$ be an n-dimensional vector of ones. The denominator of the standardized test statistic becomes

$$\begin{aligned}
& X^TW^{1/2}(I-H)F(I-H)F(I-H)W^{1/2}X = \\
= \; & a^TF(I-UU^T)Fa = \\
= \; & a^TFIFa - a^TFUU^TFa \\
= \; & a^Ta - \mathbf{1}^TAFUU^TFA\mathbf{1} \\
= \; & a^Ta - \mathbf{1}^TFAUU^TAF\mathbf{1} \\
= \; & a^Ta - f^TCC^Tf,
\end{aligned}$$

where $f = F\mathbf{1}$ and $C = AU$.

Therefore, given that a^Ta is a constant and the computational cost of f^TCC^Tf is linear with n since we can write $f^TCC^Tf = \sum_{j=1}^{q}(\sum_{i=1}^{n} f_iC_{ij})^2$, the result of the lemma follows. □

The proposed standardization has great impact and it substantially improves the type I error control of the original score flipping test. Analogously with the test defined in (1.5), consider without loss of generality testing (1.2) against $H_1 : \beta > \beta_0$. Take the test that rejects the null hypothesis if

$$S_1^* > S^*_{(\lceil(1-\alpha)g\rceil)}. \tag{1.9}$$

The following Proposition concerns the validity of the test.

Proposition 2. Assume that the variances are correctly specified, that is, $\tilde{V} = V$, and that Assumptions 1-3 hold. The standardized sign-flip score statistic is finite sample second-moment null-invariant. The test that rejects H_0 if (1.9) holds is i) an exact α level test in case of normal response with identity link and ii) asymptotically exact in all other cases.

Proof. The proof is immediate. We simply observe that the expected value of the new statistic (1.8) is left unchanged, while the standardization makes the variance of each flipped new statistic equal to 1. The same argument given in the last part of the proof of Theorem 1 applies to deduce the asymptotic exactness of the test.

In the special case of the normal model with the identity link, we have that $\mathbb{V}[S(F)]$ does not depend on any unknown nuisance parameters. Moreover, in this model $S(\mathbf{I})$ and $S(\mathcal{F})$ are normally distributed in finite samples, so the standardized sign-flip score statistic achieves not just second-moment but full null-invariance, resulting in an exact test by Theorem 2 of Hemerik and Goeman (2018). □

1.5 Robustness of the test

Hemerik *et al.* (2020) showed promising simulation results suggesting robustness against variance misspecification, but a formal proof was given only for the special case that all variances were misspecified by the same multiplicative constant. In this section we provide a formal proof of general robustness of the test to variance misspecification, both for the novel test derived in the previous section and for the original effective score test of Hemerik *et al.* (2020). This is the most important result of this chapter.

Wrong specification of the model variance means that

$$\mathbb{E}\left[(Y-\mu)(Y-\mu)^T\right] = V \neq \tilde{V}.$$

We first revisit the simple case considered by Hemerik *et al.* (2020), in which the variance is misspecified by a multiplicative constant. This is the case when the true model has a common dispersion parameter ϕ. It turns out that the properties of the standardized sign-flip score test are not affected by this misspecification. In particular, we have second-moment null-invariance.

Proposition 3. Assume that Assumptions 1-3 hold. If the variances are misspecified by any finite constant $c > 0$, that is, $V = c\tilde{V}$, the standardized sign-flip score statistic is second-moment null-invariant. The test that rejects H_0 if (1.9) holds is i) an exact α level test in case of normal response with identity link and ii) asymptotically exact in all other cases.

Proof. We observe that

$$V\tilde{V}^{-1} = \phi\tilde{\phi}^{-1}I = cI$$

and therefore

$$\mathbb{V}[S^*(F)] = c \quad \forall F \in \mathbb{F}.$$

The result then follows from the proof of Proposition 2. □

The proposition uses the property that the test is invariant to multiplication by a constant. This result is relevant, for instance, in models with a common unknown dispersion parameter, such as normal regression model. In such cases the test can be performed, and retains second-moment null-invariance, without the need to estimate the common dispersion parameter. We can save ourselves the effort of estimating it, taking simply $\hat{\phi} = 1$. This special situation coincides with the standard parametric framework based on the quasi-likelihood approach.

The main result of this chapter concerns the case of general variance misspecification. The following theorem, which is a strong improvement over the properties shown by Hemerik *et al.* (2020), shows that we can still get an asymptotic exact test.

Theorem 2. Assume that the variances are misspecified, that is, $V \neq \tilde{V}$ and that Assumptions 1-3 hold. For $n \to \infty$, the standardized sign-flip score statistic is asymptotically second-moment null-invariant. The test that rejects H_0 if (1.9) holds is an asymptotically α level test.

Proof. It is trivial to see that the expected value of the test statistic is not affected by this misspecification.

Let

$$B = V\tilde{V}^{-1},$$

note that it is a diagonal matrix and all its elements are finite and greater than zero by Assumption 1. We compute again the variance of (1.4), but now we consider the misspecification.

Let $\tilde{H}$ and $\tilde{a}$ be the quantities defined in (1.3) and (1.7) for $W = \tilde{W}$. The variance can be written as

$$\begin{aligned}\mathbb{V}\left[S(F)\right] &= n^{-1}\tilde{a}^T F(I-\tilde{H})\tilde{V}^{-1/2}E\left\{(Y-\mu)(Y-\mu)^T\right\}\tilde{V}^{-1/2}(I-\tilde{H})F\tilde{a} + o(1)\\ &= n^{-1}\tilde{a}^T F(I-\tilde{H})B(I-\tilde{H})F\tilde{a} + o(1).\end{aligned}$$

Take the difference

$$\begin{aligned}\mathbb{V}\left[S(\mathbf{I})\right] - \mathbb{V}\left[S(F)\right] &= n^{-1}\tilde{a}^T\left\{B - F(I-\tilde{H})B(I-\tilde{H})F\right\}\tilde{a} + o(1)\\ &= n^{-1}\tilde{a}^T\left[F\{\tilde{H}B + (I-\tilde{H})B\tilde{H}\}F\right]\tilde{a} + o(1).\end{aligned}$$

We notice that by Lemma 3

$$\sum_{F\in\mathcal{F}} F(\tilde{H}B + B\tilde{H} - \tilde{H}B\tilde{H})F = 2^n \operatorname{diag}\{\tilde{H}B + B\tilde{H} - \tilde{H}B\tilde{H}\}$$

where the i-th element of that diagonal matrix is

$$2\tilde{h}_{ii}b_i - \sum_{k=1}^{n} \tilde{h}_{ik}^2 b_k.$$

Then we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{V}\left\{S(\mathbf{I})\right\} - \mathbb{V}\left\{S(\mathcal{F}) \mid F\right\}\right] &= (2^n)^{-1} \sum_{F\in\mathbb{F}} \left[\mathbb{V}\left\{S(\mathbf{I})\right\} - \mathbb{V}\left\{S(F)\right\}\right] \\ &= n^{-1}\tilde{a}^T \operatorname{diag}\{\tilde{H}B + B\tilde{H} - \tilde{H}B\tilde{H}\}\tilde{a} + o(1) \\ &\leq n^{-1}\tilde{a}^T \operatorname{diag}\{\tilde{H}B + B\tilde{H}\}\tilde{a} + o(1).\end{aligned}$$

By Assumption 1, note that there exists two finite positive constants c_1, c_2 such that for each i-th element

$$2\tilde{h}_{ii}b_i \leq c_1 b_{sup}\tilde{h}_{ii} \leq c_2 \sup_{1\leq i\leq n} \tilde{h}_{ii} = c_2\tilde{h}_{sup}$$

where

$$\tilde{h}_{\mathrm{sup}} = \sup_{1\leq i\leq n} \{\tilde{h}_{ii}\} \quad b_{\mathrm{sup}} = \sup_{1\leq i\leq n} \{b_i\}.$$

Therefore we can derive the upper bound

$$\mathbb{E}\left[\mathbb{V}\left\{S(\mathbf{I})\right\} - \mathbb{V}\left\{S(\mathcal{F}) \mid F\right\}\right] \leq n^{-1}\|\tilde{a}\|^2 c_2 \cdot \tilde{h}_{\mathrm{sup}}.$$

Using Assumption 2 and Lemma 4, the limiting behavior is

$$\lim_{n\longrightarrow\infty} n^{-1}\|\tilde{a}\|^2 c_2 \cdot \tilde{h}_{\mathrm{sup}} = 0.$$

Meanwhile, for two positive constants c_3, c_4 we have for each i-th element

$$-\sum_{k=1}^{n} \tilde{h}_{ik}^2 b_k \geq -\sum_{k=1}^{n} \tilde{h}_{ik}^2 c_3 b_{\mathrm{sup}} = -c_3 b_{\mathrm{sup}}\tilde{h}_{ii} \geq -c_4\tilde{h}_{\mathrm{sup}}.$$

Then using again Lemma 3 we can derive the lower bound

$$\begin{aligned}\mathbb{E}\left[\mathbb{V}\left\{S(\mathbf{I})\right\} - \mathbb{V}\left\{S(\mathcal{F}) \mid F\right\}\right] &\geq n^{-1}\tilde{a}^T \operatorname{diag}(-\tilde{H}B\tilde{H})\tilde{a} + o(1) \\ &\geq -n^{-1}\|\tilde{a}\|^2 c_4 \cdot \tilde{h}_{\mathrm{sup}} + o(1)\end{aligned}$$

where, using Assumption 2 and Lemma 4, the limiting behavior is

$$\lim_{n \longrightarrow \infty} -n^{-1} \|\tilde{a}\|^2 c_4 \cdot \tilde{h}_{\text{sup}} = 0.$$

According to the law of total variance we have

$$\mathbb{V}[S(\mathcal{F})] = \mathbb{V}\left[\mathbb{E}\{S(\mathcal{F}) \mid F\}\right] + \mathbb{E}\left[\mathbb{V}\{S(\mathcal{F}) \mid F\}\right].$$

We know that $\mathbb{E}[S(\mathcal{F}) \mid F]$ does not depend on $\mathcal{F}$, which means that $\mathbb{V}[\mathbb{E}\{S(\mathcal{F}) \mid F\}] = 0$, so $\mathbb{V}[S(\mathcal{F})] = \mathbb{E}[\mathbb{V}\{S(\mathcal{F}) \mid F\}]$. It follows that, marginally over F,

$$\lim_{n \longrightarrow \infty} \operatorname{var}\left[S(I)\right] - \operatorname{var}\left[S(\mathcal{F})\right] = 0.$$

The same argument given in the last part of the proof of theorem 1 applies to deduce the asymptotic exactness of the test considered. □

The following corollary enlarges the robustness property to the effective sign-flip score test, as an immediate consequence from the proof of the preceding theorem.

Corollary 1. Assume that the variances are misspecified, that is, $V \neq \tilde{V}$, and that Assumptions 1-3 hold. For $n \to \infty$, the effective sign-flip score statistic is asymptotically second-moment null-invariant. The test that rejects H_0 if (1.5) holds is an asymptotically α level test.

1.6 Testing for multiple regressors

Until now we have considered hypotheses about a single parameter $\beta \in \mathbb{R}$. We now generalize the test of the previous section to $\beta \in \mathbb{R}^d,\ d < n - q$. We consider a standard asymptotic setting where d is fixed while n increases. The null hypothesis of interest is now given by:

$$H_0 : \beta = \beta_0 \in \mathbb{R}^d \mid (\gamma, \phi_1, \ldots, \phi_n) \in \Gamma \times \Phi \times \cdots \times \Phi,$$

where $\Gamma \subseteq \mathbb{R}^q$ and $\Phi \subseteq (0, \infty)$, which reduces to the null hypothesis (1.2) if $d = 1$.

For this multivariate setting, we have to generalize the assumptions of Section 1.1. Assumption 1 remains unchanged while Assumptions 2 and 3 are replaced by their multivariate counterparts, respectively,

Assumption 4. The $\lim_{n \longrightarrow \infty} n^{-1} \tilde{\mathcal{I}}_{\beta,\beta}$ converges to a positive definite matrix.

Assumption 5. We require, for all $\epsilon > 0$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[||\tilde{\nu}_{i,\beta}^{*}||^{2} \mathbf{1}_{\{||\tilde{\nu}_{i,\beta}^{*}||/\sqrt{n} > \epsilon\}} \right] \to \mathbf{0},$$

and $n^{-1}\sum_{i=1}^{n} \mathbb{V}[\tilde{\nu}_{i,\beta}^{*}]$ to converge to a positive definite matrix, where $|| \cdot ||$ denotes the ℓ^2 norm and $\mathbf{0}$ is a d-dimensional zero vector.

Hemerik *et al.* (2020, Section 4) derived a generalization of the sign-flipped effective score test statistic, as follows. Noting that the effective score is now a d-dimensional vector $S(F) = (S^1(F), \dots, S^d(F))^T$, an asymptotically exact α-level test can be constructed by using the idea of the nonparametric combination methodology (Pesarin, 2001). The test statistic takes the form

$$T(F) = \{S(F)\}^T M \{S(F)\}$$

where M is any non-zero matrix. Usually M is chosen to be a symmetric matrix, and in general this choice influences the distribution of the power between the alternatives (see Hemerik *et al.* (2020) for details). A common choice for M is the inverse of an estimate of the effective Fisher information of β, if available.

As in Section 1.4, we can improve control of type I error by standardizing the score vector. Indeed, the same reasoning of Theorem 2 applies, showing that the sign-flip standardized score vector is finite sample second-moment null-invariant. The definition of the test is analogous by noting that now

$$\mathbb{V}[S(F)] = n^{-1} X^T W^{1/2} (I - H) F (I - H) F (I - H) W^{1/2} X + o(1)$$

is a $d \times d$ matrix and hence the standardized score

$$S^*(F) = S(F) / \mathbb{V}[S(F)]^{1/2}$$

is a d-dimensional vector. We can therefore define the test statistic as

$$T^*(F) = \{S^*(F)\}^T M \{S^*(F)\}.$$

Assume we observe values $T_1^*, \dots, T_g^*$, where $T_1^* = T^*(\mathbf{I})$ is the observed test statistic, while the sorted values are $T_{(1)}^* \leq \dots \leq T_{(g)}^*$. Consider the test that rejects the null hypothesis if

$$T_1^* > T_{(\lceil (1-\alpha) g \rceil)}^*. \tag{1.10}$$

The following Proposition states that the test is second-moment null-invariant and asymptotically exact.

Proposition 4. Assume that the variances are correctly specified, that is, $\tilde{V} = V$, and that Assumptions 1, 4 and 5 hold. The d-dimensional standardized sign-flip score vector is finite sample second-moment null-invariant. The test that rejects H_0 if (1.10) holds is an asymptotically α level test.

Proof. The proof is analogous to Theorem 3 of Hemerik *et al.* (2020). The asymptotic exactness in all the cases is a consequence of the use of the continuous mapping theorem. □

Finally, the following theorem shows that the robustness properties of Theorem 2 are inherited by this multivariate extension.

Theorem 3. Assume that the variances are misspecified, that is, $V \neq \tilde{V}$ and that Assumptions 1,4,5 hold. For $n \to \infty$, the d-dimensional standardized sign-flip score vector is asymptotically second-moment null-invariant. The test that rejects H_0 if (1.10) holds is an asymptotically α level test.

Proof. It is trivial to observe that the expected value is a d-dimensional zero vector. We focus on the variance, which now is a $d \times d$ matrix. We prove the robustness elementwise.

The proof for the diagonal elements follows from Theorem 2. Then, we focus on one covariance term of the covariance matrix of the flipped effective score vector. Let $X = (X_1, \dots, X_d)^T$ and

$$\begin{aligned}\tilde{a}_1 &= (I - \tilde{H})\tilde{W}^{1/2}X_1\\ \tilde{a}_2 &= (I - \tilde{H})\tilde{W}^{1/2}X_2.\end{aligned}$$

When the variances are misspecified

$$\begin{aligned}&\mathbb{C}[S^1(F), S^2(F)] =\\ &\quad = n^{-1}\tilde{a}_1^T F(I - \tilde{H})\tilde{V}^{-1/2}E\left\{(Y-\mu)(Y-\mu)^T\right\}\tilde{V}^{-1/2}(I - \tilde{H})F\tilde{a}_2 + o(1)\\ &\quad = n^{-1}\tilde{a}_1^T F(I - \tilde{H})B(I - \tilde{H})F\tilde{a}_2 + o(1).\end{aligned}$$

Take the difference

$$\begin{aligned}
&\mathbb{C}[S^1(I), S^2(I)] - \mathbb{C}[S^1(F), S^2(F)] = \\
&= n^{-1}\tilde{a}_1^T \left\{ B - F(I - \tilde{H})B(I - \tilde{H})F \right\} \tilde{a}_2 + o(1) \\
&= n^{-1}\tilde{a}_1^T \left\{ F(\tilde{H}B + B\tilde{H} - \tilde{H}B\tilde{H})F \right\} \tilde{a}_2 + o(1).
\end{aligned}$$

From this point the asymptotic second-moment null-invariance can be derived directly from the proof of the Theorem 2, by applying the same reasoning to each covariance term, replacing Assumptions 2-3 with Assumptions 4-5. The asymptotic exactness of the test follows from the proof of Proposition 4. □

1.7 Simulation study

We explore six different settings to compare empirically the type I error control of the usual parametric approach (by considering the Wald test), the effective and standardized Flipscores tests, which are based on 5 000 random sign-flips (we refer to Hemerik and Goeman (2018) for a discussion about the use of a limited number of random flips), the Wald test based on the use of the sandwich estimator of the variance to correct for variance misspecification, and the Gamlss, modeling the variance as a function of the full model. A total of 5 000 simulations have been carried out for each setting. The covariates have been drawn from a multivariate normal distribution, with $X \in \mathbb{R}$ and $Z \in \mathbb{R}^3$. The three nuisance covariates have correlation with the target variable equal to $(0.5, 0.1, 0.1)$, while the true parameter is set to $\beta = 0$. The null hypothesis considered is $H_0 : \beta = 0$ against a two-sided alternative, with a nominal significance level $\alpha = 0.05$. Different sample sizes are considered $(25, 50, 100, 200, 500, 1\,000)$.

The two plots of Figure 1.1 are settings with correctly specified models, respectively a Poisson and a Logistic regression models. We observe that, in case of correctly specified models, the standardized test is close to the nominal level, even with $n = 25$, improving the asymptotic convergence of the effective test. The parametric test shows few rejections with small sample size for the logistic model, due to a poor approximation of the likelihood. The sandwich shows slow convergence in both cases, which is unsatisfactory since we are dealing with a well-specified model. Finally, the Gamlss shows a low convergence, explained by the greater number of parameters involved.

The plots of Figure 1.2 represent normal models with neglected heteroscedasticity, which depends either on a nuisance covariate or on the tested variable, i.e. respectively $\text{var}(y_i) = 4z_i^2$ and $\text{var}(y_i) = 4x_i^2$. In the left plot the parametric test behaves correctly,

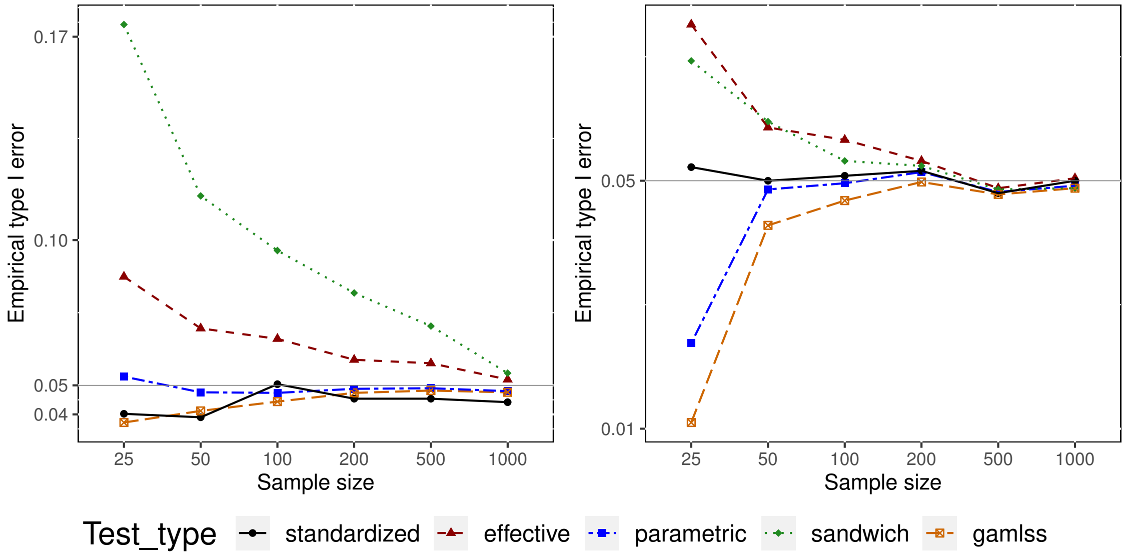

FIGURE 1.1: Type I error control comparison. **Left:** correct Poisson model. **Right:** correct Logistic model.

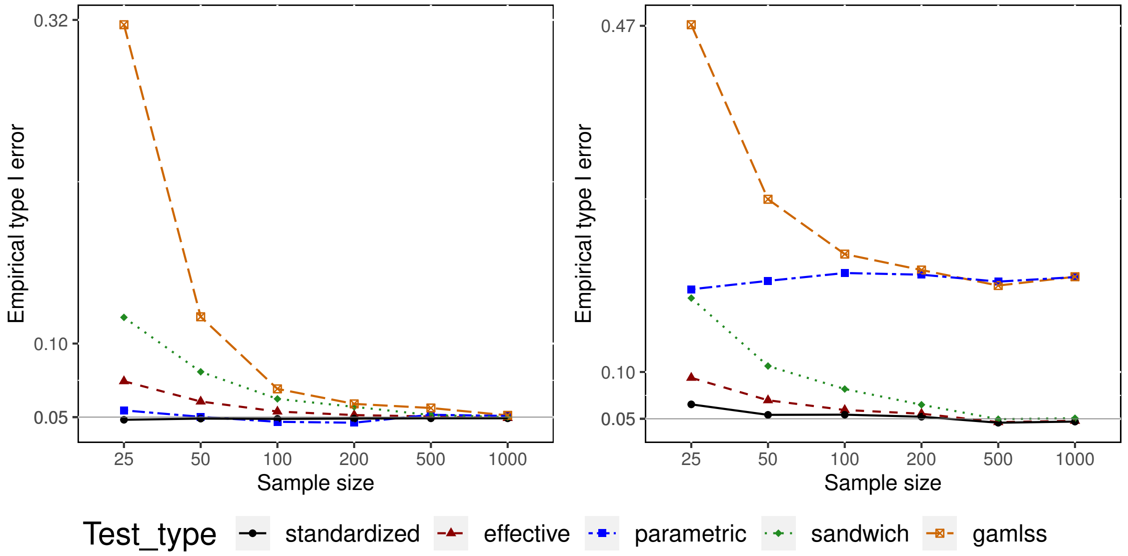

FIGURE 1.2: Type I error control comparison. **Left:** Normal with nuisance heteroscedasticity. **Right:** Normal with target heteroscedasticity.

while we observe its failure in the right plot, even for increasing sample size. The standardized test is always close to the nominal level, confirming its improvement over the effective test. Finally, we observe that the sandwich estimator slowly converges to the nominal level, while the Gamlss has also convergence problems.

In the left plot of Figure 1.3 a Poisson model was fitted when the true distribution was negative binomial with dispersion parameter $\phi = 1$, that is, additional heteroscedasticity relative to the Poisson model that depends on the mean. The right plot displays the results of a two-sample test fitting a negative binomial regression model. A common

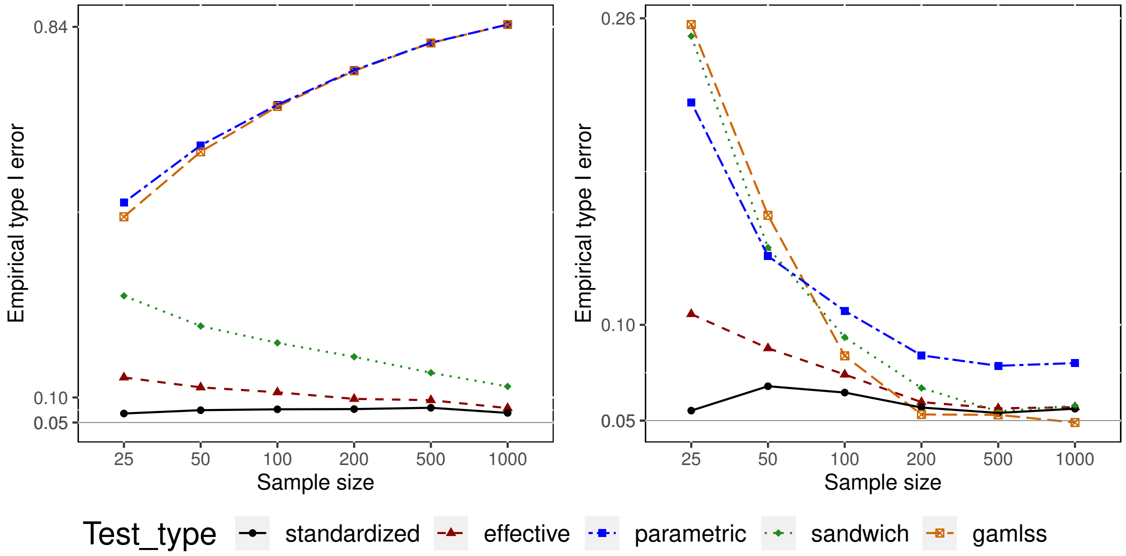

FIGURE 1.3: Type I error control comparison. **Left:** false Poisson model. **Right:** two groups Negative-binomial.

dispersion parameter was assumed for the fitted negative binomial model, but the two groups are unbalanced (proportion equal to 2/3 and 1/3), and generated from two distributions with different dispersion parameter (0.4 and 1). The two plots show the failure of the parametric test in presence of some forms of variance misspecification, where the rejection fraction converges to a level far from the nominal for increasing sample size. The standardized test outperforms its competitors in all cases, being the closest to the nominal level. In particular, the improvement over the effective test and further over the sandwich test is clear. Further, the Gamlss surprisingly shows a failure in the left plot, where the reason is due to the impossibility of modeling the variance adopting a Poisson model. In the right plot it shows a slower convergence compared to the sign-flip tests. Moreover, it is remarkable to see that the standardized test seems to converge to the effective test faster than going to the nominal level, leading to an initial worsening of the true level, with a recovery for larger sample size.

Finally, Figure 1.4 contains an evaluation of the power of the tests with two well-specified models, respectively a Poisson and a Normal model, with true parameter equal respectively to 0.3 and 1. The results for the sandwich estimator are given for completeness, although they are not comparable for small sample size, since that method has no control of the type I error. We see that the improvement of type I error control of the standardized test with respect to the effective naturally costs some power. Analogously, we see some power loss also with respect to the parametric model and the Gamlss as expected, but this difference is remarkably small.

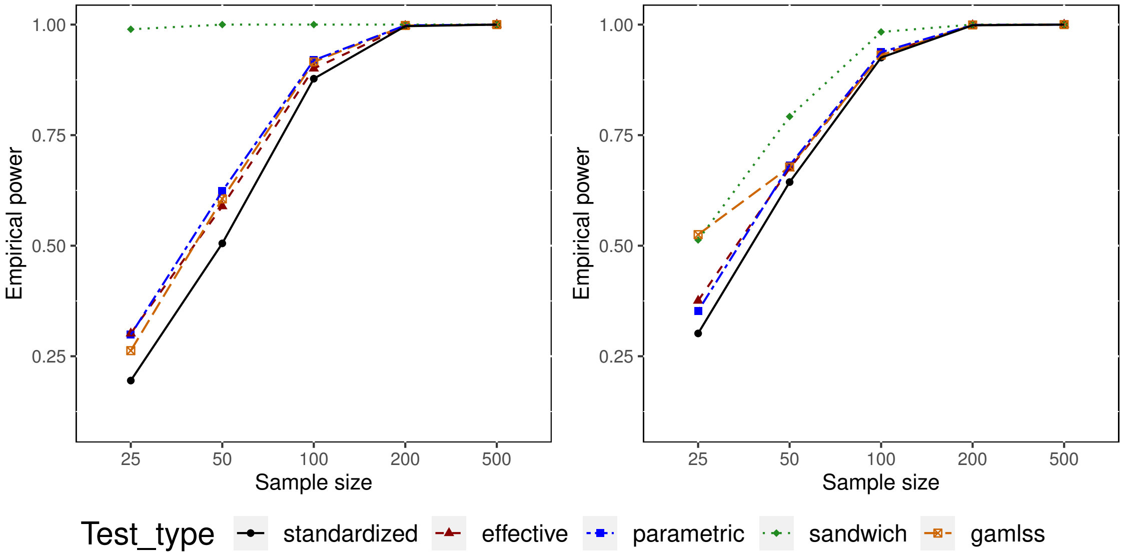

FIGURE 1.4: Power comparison. **Left:** correct Poisson model. **Right:** correct Normal model.

Chapter 2

Permutation-based multiple testing when fitting many generalized linear models

2.1 Multivariate generalized linear models

Suppose to observe n independent observations. For each unit i assume that are observed $m \geq 2$ dependent variables $Y^1, \dots, Y^m$. Therefore, for each response Y^l we have n independent observations $Y^l_1, \dots, Y^l_n$, which we assume to follow some model in the exponential dispersion family. We further observe for each i-th observation some covariates (X_i, Z_i). For each l-th response we have a univariate generalized linear model

$$g(\mu^l) = \eta^l = X\beta^l + Z\gamma^l. \tag{2.1}$$

where $\mu^l = (\mu^l_1, \dots, \mu^l_n)^T$ denotes the mean vector, $g(\cdot)$ is the link function, taken to operate elementwise on a vector, (X, Z) is the full rank design matrix with $\dim\{X\} = n \times 1$, $\dim\{Z\} = n \times q$, where q does not depend on n, and (β^l, γ^l) are unknown parameters. We consider β^l the parameter of interest, and γ^l a nuisance parameter. Throughout the chapter we will consider $\dim\{\beta^l\} = 1$ for simplicity, but the extension to a multivariate parameter is straightforward using the test of Section 1.6. Note that, for each response l, we will use the same notation of Chapter 1. Further, for each model we do the same assumptions of Chapter 1, and we assume the covariances between responses to be uniformly bounded.

Our interest is in testing m null hypotheses $H_1, \dots, H_m$, where H_l is the hypothesis that $\beta^l = \beta^l_0$. This Chapter introduces the multiple testing procedure, which involves computing the effective score for each of the dependent variable $Y^1, \dots, Y^m$. The effective score for the l-th response is then

$$S^l = n^{-1/2} X^T (W^l)^{1/2} (I - H^l)(V^l)^{-1/2}(Y^l - \hat{\mu}^l) = n^{-1/2} \sum_{i=1}^{n} \nu^{*,l}_{\beta,i}.$$

Analogously to Chapter 1, we define the sign-flipped effective score test statistic for a generic flip matrix F as

$$S^l(F) = n^{-1/2} X^T (W^l)^{1/2} (I - H^l)(V^l)^{-1/2} F (Y^l - \hat{\mu}^l).$$

Further, we have

$$\mathbb{V}\left[S^l(F)\right] = n^{-1} X^T (W^l)^{1/2} (I - H^l) F (I - H^l) F (I - H^l)(W^l)^{1/2} X.$$

while the standardized score is

$$S^{*,l}(F) = S^l(F) / \mathbb{V}\left[S^l(F)\right]^{1/2}.$$

2.2 Local test for multivariate responses and the multiple testing problem

When testing multivariate hypotheses, the multiple testing problem is well-known and has generate a huge literature (see Goeman and Solari (2014) for a review of the most known methods). Indeed, it can be easily proven that using individual valid α-level test does not imply the control over the overall tests when performing many hypotheses. Further, defining a valid local test for intersection hypotheses in multivariate generalized linear models is challenging. Assuming normality, in absence of covariates the hypothesis about mean vectors can be performed by the Hotelling one-sample T^2 test (Hotelling, 1931). When covariates are present we can still find a test statistic with known distribution, the Wilks lambda distribution (Wilks, 1962). For general multivariate GLMs we do not have solutions unless we specify a particular correlation structure of the responses, built on a limited number of parameters. This solution might be undesired if the data do not have a particular structure, and it can be unfeasible especially for large value of m.

Several measures have been proposed in the literature (Tukey, 1953; Benjamini and Hochberg, 1995; Goeman and Solari, 2011; Lehmann and Romano, 2012) in order to properly address the multiple testing problem and the related issues, but we will only introduce those of interest for this chapter.

The familywise error rate (FWER) (Tukey, 1953) is defined as the probability of making any type I error. There are formally two types of FWER control; weak FWER means that we control the error for a global null hypothesis, while strong FWER control is obtained when the correction holds for any combination of true and false null hypotheses. In practice, only procedures with strong control of FWER are of interest, while methods with weak control of type I error should be avoided (Goeman and Solari, 2014).

Parametric methods for generalized linear models can be built as follows. We start from univariate valid test as, for instance, the Wald test, assuming that the model is correctly specified. Then some corrections are possible.

The Bonferroni correction (Holm, 1979) is widely popular, thanks to its simplicity, and its control of FWER is exact for any correlation structure between the underlying p-values. Let $q_1, \ldots, q_m$ be the observed p-values. The method rejects all the hypotheses for which $q_i \leq \alpha/m$. However, the method can be quite conservative in the presence of many false null hypotheses and also when the p-values are positively correlated.

A sequential variant of the Bonferroni's procedure is the Holm's method (Holm, 1979). It guarantees strong FWER control under the same assumptions of the Bonferroni's method, and it always rejects at least as much as Bonferroni's method, and often a bit more. The procedure is a multi-step extension as follows. In the first step it rejects all null hypothesis such that $q_i \leq \alpha/m$, just like the Bonferroni's method. Suppose this leaves to h_1 hypotheses unrejected. In the next step we reject all hypotheses such that $q_i \leq \alpha/h_1$. This leaves h_2 hypotheses unrejected. The procedure is thus iterated until there are no more rejections. Again, the procedure can be quite conservative when the p-values are positively correlated.

More procedures are available if we want to make additional assumptions about the underlying dependence of the p-values, with the advantage of a gain in power. We do not want to make stricter assumptions, and we refer to Goeman and Solari (2014) for an overview.

Finally, a different approach consists in using permutations to estimate the underlying dependence instead of making assumptions about the correlation structure. Permutations are not directly usable for GLMs as the outcomes are not exchangeable, as well as the usual test statistics, but they will be the basis of the new proposal of this chapter. We refer to Section 2.3 for further details.

The main critic to FWER-control methods is related to the fact that they might be too stringent, resulting in low power. While being adequate for confirmatory analysis, other measures might be appealing for an exploratory analysis. A measure which answer this question is the true discovery proportion (TDP) Goeman and Solari (2011). This measure quantify a lower bound for the proportion of true discoveries - that is, non-null coefficients - over any set of hypotheses. In order to be helpful, the procedure to estimate a lower bound for the TDP must be simultaneously valid for any subset chosen post-hoc. This is especially useful when the responses are related to different spatial locations and we are interested in inference over a cluster of close responses. Moreover, in some cases reporting the estimated TDP bounds may be more useful than individual adjusted p-values, in the sense that it usually gives an higher number of findings, and it also answers a slightly different research question.

The closed testing procedure (Marcus *et al.*, 1976) is proven to be the optimal way to estimate this lower bound for any subset (Goeman *et al.*, 2021). In detail, the closed testing procedure requires to compute valid α-level tests for any intersection hypothesis. Let $\mathcal{C}$ be the set with all possible combination of all possible indices corresponding to the elementary null hypotheses. Further, let $\mathcal{U}$ be the set containing the raw rejections uncorrected for multiple testing. The procedure rejects every $I \in \mathcal{C}$ for which $J \in \mathcal{U}$ for every $J \supseteq I$. We can then estimate a lower bound for the TDP for any set R of elementary null hypothesis with $t_\alpha(R) = 1 - t_\alpha(R)/|R|$, where $t_\alpha(R)$ is the size of the largest subset of R for which the corresponding intersection hypothesis is not rejected by the closed testing procedure, while $|R|$ is the cardinality of R. The α-level of the lower bound is proven in Goeman and Solari (2011).

The main issue of the closed testing procedure is the computational unfeasibility for large number of hypotheses, approximately around 20-30 (Goeman and Solari, 2011), since it is required to perform $2^m - 1$ local tests. In such case we must rely on shortcuts in order to reduce the computational burden.

A parametric shortcut is based on adjusted p-values compute through the Hommel procedure (Rosenblatt *et al.*, 2018). Note that with the term parametric we mean that the procedure starts from any valid α-level local test. The Hommel procedure is not assumption-free about the joint dependence of the p-values, but it is valid when the Simes inequality holds (Simes, 1986). It says that, with probability $1-\alpha$, simultaneously

$$q_{(i)} > \frac{i\alpha}{m_0}$$

where m_0 is the total of true hypotheses. Note that this condition is shown to be rather flexible (Goeman and Solari, 2011). For any procedure based on the Simes inequality we can apply the shortcut of Goeman *et al.* (2019), to which we refer for the details.

This can have low power in some scenarios, especially if correlated data are analyzed. When we have exchangeable test statistics we can use permutation-based methods to increase power. In this chapter we will apply two methods based on permutations. The shortcut of Vesely *et al.* (2023), which can either use any function of the original test statistics or the corresponding ranks, with the requirement of using the corresponding sum as combining function for intersection hypotheses. It is based on an iterative algorithm which is shown to converge to the optimal estimates, that is, the estimate potentially obtained from the closed testing procedure. The procedure of Andreella *et al.* (2023) is based on finding a critical vector for ordered p-values. The method aims to find the maximum distance between the sorted p-values and the critical vector. We refer to Andreella *et al.* (2023) for how to compute this critical vector.

2.3 Sign-flip tests for multivariate generalized linear models

We can build a multivariate test statistic which takes into account the standardization of the joint variance-covariance matrix as follows. Let

$$H_0 = \bigcap_{l \in L} H_l : \beta^l = \beta^l_0 \tag{2.2}$$

where L is a set containing any combination of the m hypotheses. Note that H_0 can be either a global or a partial null hypothesis for a subset of parameters.

A first idea to perform the test might be given by a full standardization. We can indeed build a global test statistic based on the form of a Mahalanobis distance as follows. Let $S^L(F)$ be the vector with element $S^l(F)$, $l \in L$. Call $L^* = |L|$, i.e. the cardinality of L. Note that by definition $S^L(F)$ is an L^*-dimensional vector. Call $\mathbb{V}[S^L(F)]$ the corresponding variance-covariance matrix, which has dimension $L^* \times L^*$ and is assumed (for the moment) known. Let the joint test statistic be

$$T_L(F) = \left(S^L(F)\right)^T \mathbb{V}\left[S^L(F)\right]^{-1} \left(S^L(F)\right) \tag{2.3}$$

If we assume the normality of the response we have the following theorem;

Theorem 2.1. *Assume to fit the regression model* (2.1) *where the responses are normally distributed with identity link for* $Y_l, l \in L$. *Let* $T_{L,j}$ *be the flipped test statistic for* $F_j, 1 \leq j \leq g$ *as defined in* (2.3). *Let* $T^n_{(1)} \leq ... \leq T^n_{(g)}$ *be the sorted statistics. Consider the test that rejects if* $T^n_1 > T^n_{\lceil(1-\alpha)g\rceil}$. *Under* H_0, *the level of the test is* $\lfloor \alpha g \rfloor / g \leq \alpha$.

Proof. Trivially, note that $S^L(F)$ has a multivariate normal distribution with zero mean and that the test statistic is a quadratic form. The expected value of each test statistic is equal to L^* for every flip, while the variance is equal to $2L^*$ for every flip. This follows from standard properties of random vectors, since the random quadratic form has a generalized chi-squared distribution. It implies that all the test statistics share the first two moments. Thus we can directly apply Theorem 1 of Hemerik and Goeman (2018) to show that the test derived is an α-level test for finite sample size. □

The procedure outlined has some issues. First of all, the test is exact only for normal responses with known variance. For non-linear models we have to follow asymptotic arguments related to the asymptotic normal distribution of the effective score statistic (Marohn, 2002), while when we plug-in an estimate of the variance we rely on asymptotic results of basic probability theorems for the estimation of the variance. Another weakness is the fact that for each flip it is required to invert an $m \times m$ matrix, which might be unfeasible for large values of m. Further, we have to estimate the correlation between responses, which must be assumed to be known except for a limited number of parameters; for instance, we might choose to assume the correlation of the responses to be equal between different units. Moreover, in order to perform an overall analysis with post-hoc validity, we might choose a closed testing approach (Marcus *et al.*, 1976), which requires to perform all the 2^m possible tests, which might be easily large for growing m.

A fast alternative, which is more appealing for large values of m, consists of doing marginal standardization of the test statistic. We will start from the effective scores, showing that we are able to derive a valid procedure for both the sign-flip approaches. Finally, we will see that we still have a remarkable result for the normal linear model.

For each flip j, let $S^{j,l} = S^l(F_j)$. Further, we now explicit the difference between the effective score based on the estimated nuisance parameter $S^{j,l}_{\hat{\gamma}}$ and those based on the true parameter $S^{j,l}_{\gamma_0}$. The corresponding contributions are denoted as $\nu^*_{\hat{\gamma}}$ and $\nu^*_{\gamma_0}$. Let $\boldsymbol{M}^n$ be the g-by-m matrix with (j,l)-th entry equal to $S^{j,l}_{\hat{\gamma}}$. The following lemma will be fundamental in proving that the proposed multiple testing methods are asymptotically exact.

Lemma 2.2. *Let $\boldsymbol{M}^n$ as defined above. Then, for $n \to \infty$, $\boldsymbol{M}^n$ converges in distribution to $\boldsymbol{M}$, where all rows of $\boldsymbol{M}$ have the same multivariate normal distribution.*

Proof. Let $\boldsymbol{M}^n_0$ be the g-by-m matrix with (j,l)-th entry equal to $S^{j,l}_{\gamma_0}$. Note that $\boldsymbol{M}^n_0$ is based on knowledge of the true nuisance parameters $\boldsymbol{\gamma}_0$. The consequence is that each entry of the matrix $\boldsymbol{M}^n_0$ is the sum of n independent (flipped) score contributions. Further, note that each row of $\boldsymbol{M}^n_0$ is uncorrelated with the other rows, due to the independence of the flips. Finally, the correlation structure within each row coincides with the correlation structure of the contributions $\nu^{*1}_{\gamma_0}, \dots, \nu^{*m}_{\gamma_0}$. Consequently, the multivariate central limit theorem (Van der Vaart, 1998) implies that $\boldsymbol{M}^n_0$ converges in distribution to some matrix $\boldsymbol{M}_0$, which has identically distributed multivariate normal rows.

Now it is left to show that $\boldsymbol{M}$, i.e., the matrix based on *estimated* nuisance parameters, also has identically distributed multivariate normal rows. As shown in the proof of Theorem 2 in Hemerik *et al.* (2020) (using the result from Hall and Mathiason (1990)), we have $S^{j,l}_{\hat{\gamma}} = S^{j,l}_{\gamma_0} + o_p(1)$, $1 \leq j \leq g$, $1 \leq l \leq m$. This means that $\boldsymbol{M}$ is asymptotically equivalent to $\boldsymbol{M}_0$. Hence the result holds. □

If instead of effective scores we used standardized effective scores to fill the matrix $\boldsymbol{M}^n$, then Lemma 2.2 will still hold, since the standardized effective scores are asymptotically equivalent to the unstandardized effective scores. This is detailed in Chapter 1.

A special case of interest relates to homoscedastic linear regression models. The matrix of weights becomes an identity matrix, which causes some simplifications in the formulas; in particular

$$S^l(F) = X'(I-H)F(Y^l - \hat{\mu}^l)$$

and

$$\mathbb{V}\left[S^l(F)\right] = X'(I-H)F(I-H)F(I-H)X.$$

while the standardized score is still

$$S^{*,l} = S^l(F)/\mathbb{V}\left[S^l(F)\right]^{1/2}.$$

For the linear regression model, using the standardized scores, we are able to get finite sample results, as stated in the following Lemma. This reflects that we still get the second-moment null-invariance property for linear regression models with normal response.

Lemma 2.3. *Assume to fit the regression model* (2.1) *where the responses are normally distributed with identity link for $Y_l, l \in L$. where the covariance between responses is equal among different units, that is,*

$$\mathbb{E}[(Y^p - \mu^p)(Y^r - \mu^r)^T] = \sigma_{pr}\boldsymbol{I}_n$$

where $\boldsymbol{I}_n$ is the n-dimensional identity matrix. Let $\boldsymbol{M}^S$ be the g-by-m matrix with (j,l)-entry equal to $S^{*,j,l}_{\hat{\gamma}}$. Then $\boldsymbol{M}^S$ has independent rows with the same multivariate normal distribution.

Proof. The rows of the matrix $\boldsymbol{M}^S$ are uncorrelated, due to the independence of the flips. Within the row the covariance structure is

$$\mathbb{C}\left[S^{*,p}(F), S^{*,r}(F)\right] = \mathbb{V}\left[S^p(F)\right]^{-1/2} \mathbb{E}\left[S^p(F)\left(S^r(F)\right)^T\right] \mathbb{V}\left[S^r(F)\right]^{-1/2}$$

Using the result

$$\mathbb{E}[(Y^p - \mu^p)(Y^r - \mu^r)'] = \sigma_{pr}\boldsymbol{I}_n$$

we get, after simple computations,

$$\mathbb{C}\left[S^{*,j,p}_{\hat{\gamma}}, S^{*,j,r}_{\hat{\gamma}}\right] = \sigma_{pr}/(\sigma_p\sigma_r),$$

which is independent of the flip, where σ_p is the variance of the p-th response. It follows that the matrix M^S has independent and identically distributed multivariate normal rows for finite sample size. □

From Lemma 2.2 (or 2.3) we can build a valid local α-level test for any composite hypothesis as follows. Let H_0 as defined in (2.2). Define $\psi : \mathbb{R}^{L^*} \to \mathbb{R}$ as a function non decreasing in its argument. We can derive a global flipped test statistic, for each flip F_j, as

$$T^{\psi}_j = \psi\left(S^{*,1}(F_j), \ldots, S^{*,L}(F_j)\right). \tag{2.4}$$

The following theorem shows that from this test statistic we can get an asymptotic α-level test for the general class of generalized linear models.

Theorem 2.4. *For every $1 \le j \le w$, consider the statistic T^{ψ}_j as defined in* (2.4) *and let $T^{\psi}_{(1)} \le ... \le T^{\psi}_{(g)}$ be the sorted statistics. Consider the test that rejects if $T^{\psi}_1 > T^{\psi}_{\lceil(1-\alpha)g\rceil}$. As $n \to \infty$, under H_0 the rejection probability converges to $\lfloor \alpha g \rfloor / g \le \alpha$.*

Proof. Lemma 2.2 implies that the test statistics $T^{\psi}_1, \ldots, T^{\psi}_g$ are asymptotically independent and identically distributed. Note that, by the definition of ψ, high values of T_1 shows evidence against the null hypothesis H_0. Hence, Lemma 1 of Hemerik *et al.* (2020) directly applies and we derive that we get an asymptotically α-level test. □

Theorem 2.4 implies that we can build an asymptotic valid local test for the hypothesis (2.2), which therefore guarantees weak control of the FWER. Note that for the linear model we can obtain finite sample results, because we can directly apply Lemma 2.3 and it follows that the test has finite sample size properties.

Theorem 2.5. *Assume to fit the regression model* (2.1) *where the responses are normally distributed with identity link for $Y_l, l \in L$. For every $1 \leq j \leq g$, consider the statistic T_j^{ψ} as defined in* (2.4) *and let $T_{(1)}^{\psi} \leq ... \leq T_{(g)}^{\psi}$ be the sorted statistics. Consider the test that rejects if $T_1^{\psi} > T_{\lceil(1-\alpha)g\rceil}^{\psi}$. Under H_0 the rejection probability is $\lfloor \alpha g \rfloor / g \leq \alpha$.*

Proof. Lemma 2.3 implies that the test statistics $T_1^{\psi}, \dots, T_g^{\psi}$ are independent and identically distributed. Note that, by the definition of ψ, high values of T_1 shows evidence against the null hypothesis H_0. Hence, Theorem 1 of Hemerik and Goeman (2018) directly applies and the test derived is an α-level test for finite sample size. □

Multiple choices of the function ψ are available (Pesarin, 2001) and the choice will influence the power properties in different settings. We can subsequently apply the closed testing approach (Marcus *et al.*, 1976) to build a procedure which guarantees strong control of the FWER by computing the 2^n intersection tests, and this procedure is admissible in the sense that every multiple testing procedure is equal or can be improved by applying the closed testing principle (Goeman *et al.*, 2021) .

However, the number of tests might be unfeasible for large values of m. A dramatic shortcut is given by selecting the maximum of the test statistics as combining function in the following way. This is called the max-T approach. For sake of completeness, we will give a direct proof of its validity.

There are roughly two versions of the max-T method by Westfall and Young (Westfall and Young, 1993; Westfall and Troendle, 2008; Meinshausen *et al.*, 2011): the single-step method and the sequential method. The single-step method is simpler and faster, while the sequential method is uniformly more powerful. The single-step max-T method, based on the matrix $\boldsymbol{M}^n$ of test statistics, is defined as follows. Here we formulate a version that employs two-sided tests. For every $1 \leq j \leq g$, let m_j be the maximum of the test statistics $\boldsymbol{M}_{j,l}^n$, $1 \leq l \leq m$. Let $m_{(1)}, \dots, m_{(g)}$ be the sorted values $m_1, \dots, m_g$. Then the multiple testing method rejects all hypotheses with index l for which $\boldsymbol{M}_{1,l}^n > m_{(\lceil(1-\alpha)g\rceil)}$. The sequential max-$T$ method is defined as follows; after the first step defined above is completed, the procedure is continued in a step-down way. We remove from the matrix $\boldsymbol{M}^n$ all rows corresponding to the hypotheses rejected in the first step, then we apply again the same procedure described above. The procedure can be continued until we have no more rejections. Note that the test with standardized test statistic is defined in the same way.

The following theorem states that the single-step and sequential max-T methods provide strong asymptotic FWER control. Write $FWER_n$ to indicate potential dependence of the FWER on n.

Theorem 2.6. *For both the single-step and sequential max-T method,*

$$\limsup_{n\to\infty}(FWER_n) \leq \alpha.$$

Proof. For both the single-step and the sequential max-T method, the argument is as follows. Recall that $\boldsymbol{M}^n$ converges in distribution to $\boldsymbol{M}$. Let $\boldsymbol{M}_{\mathcal{N}}$ be the submatrix of $\boldsymbol{M}$ that only contains the rows corresponding to the true hypotheses. If the matrix $\boldsymbol{M}$ is used as input for the multiple testing procedure, then strong FWER control follows directly from the fact that the rows of $\boldsymbol{M}_{\mathcal{N}}$ are exchangeable, i.e. swapping rows does not change the distribution of the matrix. If instead $\boldsymbol{M}^n$ is used as input, then FWER control follows from the continuous mapping theorem, since $\boldsymbol{M}^n$ converges to $\boldsymbol{M}$ in distribution. This finishes the proof.

Finally, note that for the procedure that uses standardized scores for linear regression model the control of the familywise error rate is obtained for finite sample size, as the matrix $\boldsymbol{M}^S$ has an exact multivariate normal distribution. □

Until now we were focused on some kind of null hypotheses with the aim of controlling the FWER. The sign-flip test statistics can be also used as basis for any permutation-based method which aims to estimate the TDP (Vesely *et al.*, 2023; Andreella *et al.*, 2023). Indeed, the key property is that any random data transformation must preserve the distribution of the test statistics under the null hypothesis, that is, what the sign-flips asymptotically do.

2.4 Simulation Study

2.4.1 Univariate

As a setup for a simulation study we choose a multivariate logistic model. The motivation is given by the ambit of neuroimaging data, in particular by data about the lesions of the voxels which compose the brain. The analysis is generally carried on by fitting several logistic models in parallel, where the presence/absence of the lesion is modelled as a function of some observed covariates. Note that it is generally expected that the responses related to close spatial locations are positively correlated.

We first show a simulation study for univariate tests in a logistic regression model in Figure 2.1 and 2.2, where the respective sample size is equal to 50 and 100. We test $H_0 : \beta = 0$ setting the regression parameter of interest $\beta = 0$ while the nuisance regression parameter γ is equal to 0 (left plots) or 1 (right plots). The correlation

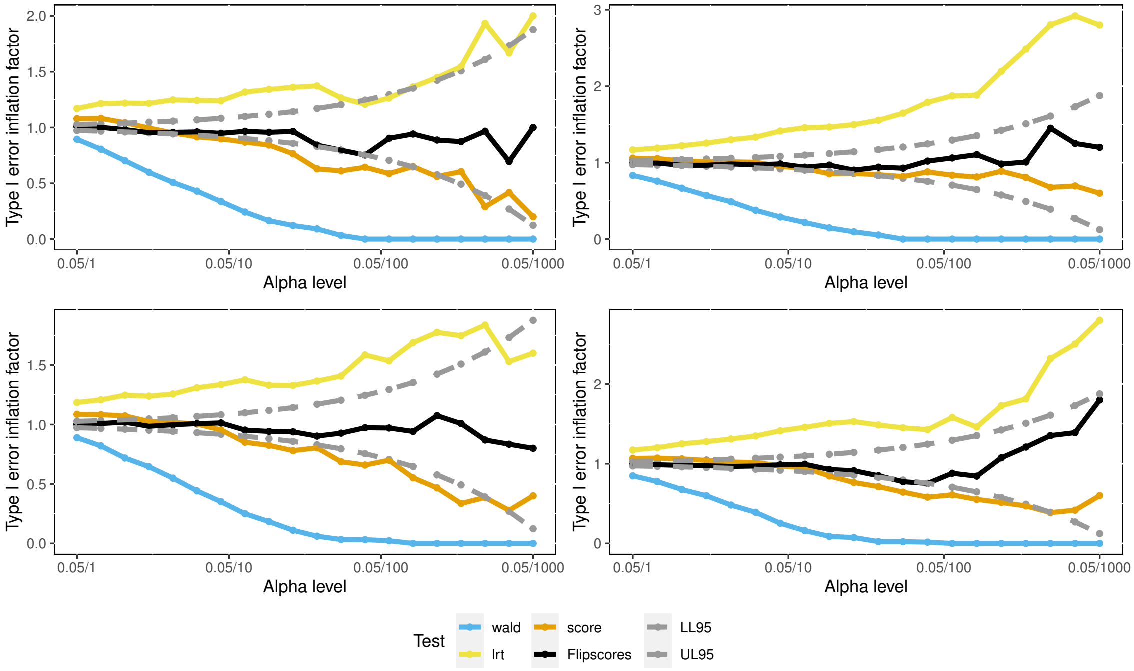

FIGURE 2.1: Logit model, univariate. Sample size of 50. On abscissa the Bonferroni-adjusted alpha for the control of FWER at level 5%, i.e. .05/number of tests

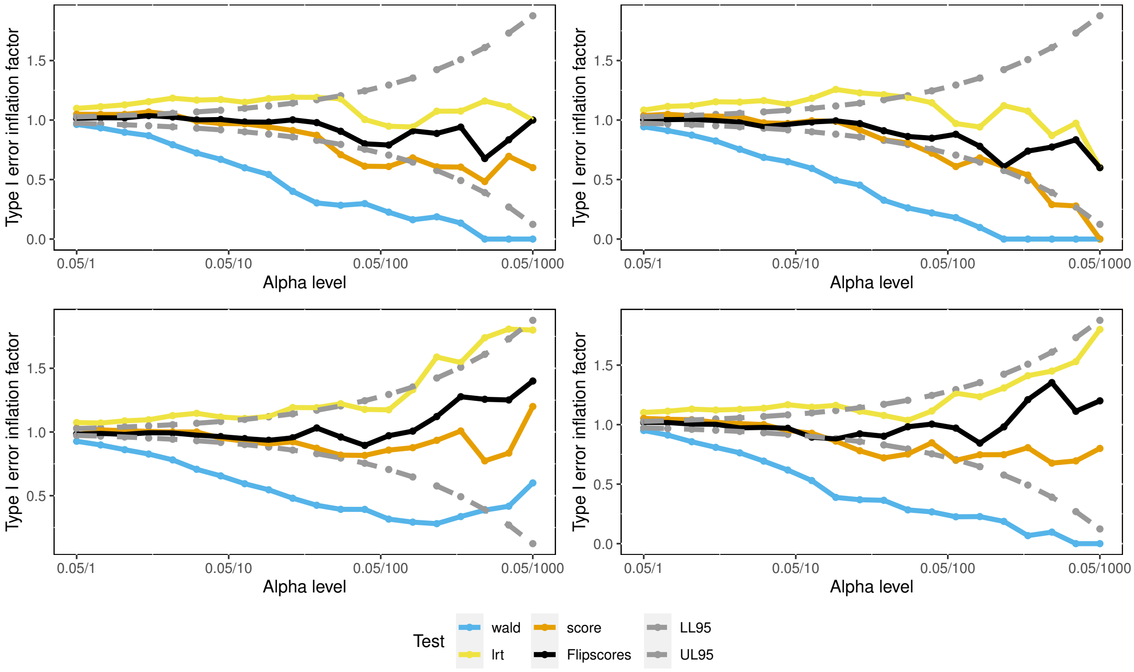

FIGURE 2.2: Logit model, univariate. Sample size of 100. On abscissa the Bonferroni-adjusted alpha for the control of FWER at level 5%, i.e. .05/number of tests

between covariates is equal to 0 (top plots) or 0.5 (bottom plots) while a total of 100 000 simulations have been run. We compare the Flipscores approach (using the standardized version) with the three standard parametric competitors, namely the Wald, parametric Score and Likelihood ratio (LRT) tests. The Flipscores test is based on 2 000 random flips. The x axis represents the multiple testing burden, that is, the Bonferroni factor given by 0.05/the number of tests. The y axis represents the ratio between the Empirical type I error and the nominal level α.

We can observe that the Flipscores is always inside the 95% simulation confidence bands. For a sample size of 50 the Score test is slightly conservative while the two other parametric methods are less satisfactory for the opposite reason, being too conservative. This negative behavior is more evident for small level of alpha, where the confidence bands are larger. The situation improves for the Score test and LRT for a sample size of 100.

2.4.2 Multivariate

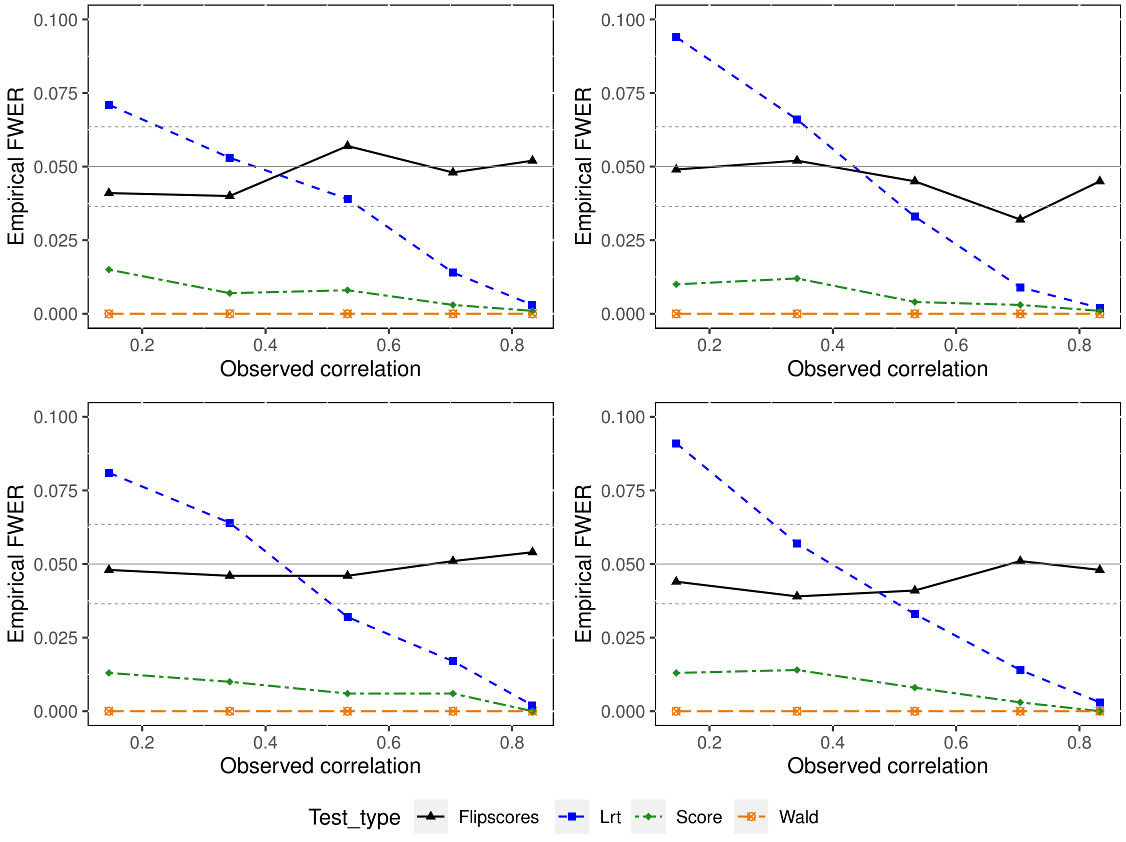

FIGURE 2.3: Logit model, multivariate. Sample size equal to 50, FWER control

Figures 2.3-2.6 represent a multivariate simulation. We set a total of 1 000 dependent variables. The model assumed is a logistic regression model with one target covariate X and one nuisance covariate Z, whose corresponding regression parameters are β and γ. We test the null hypothesis $H_0 : \beta = 0$ for each dependent variable, with a significance level of $\alpha = 0.05$. For 20% they have $\beta = 1$, that is, they are under the alternative

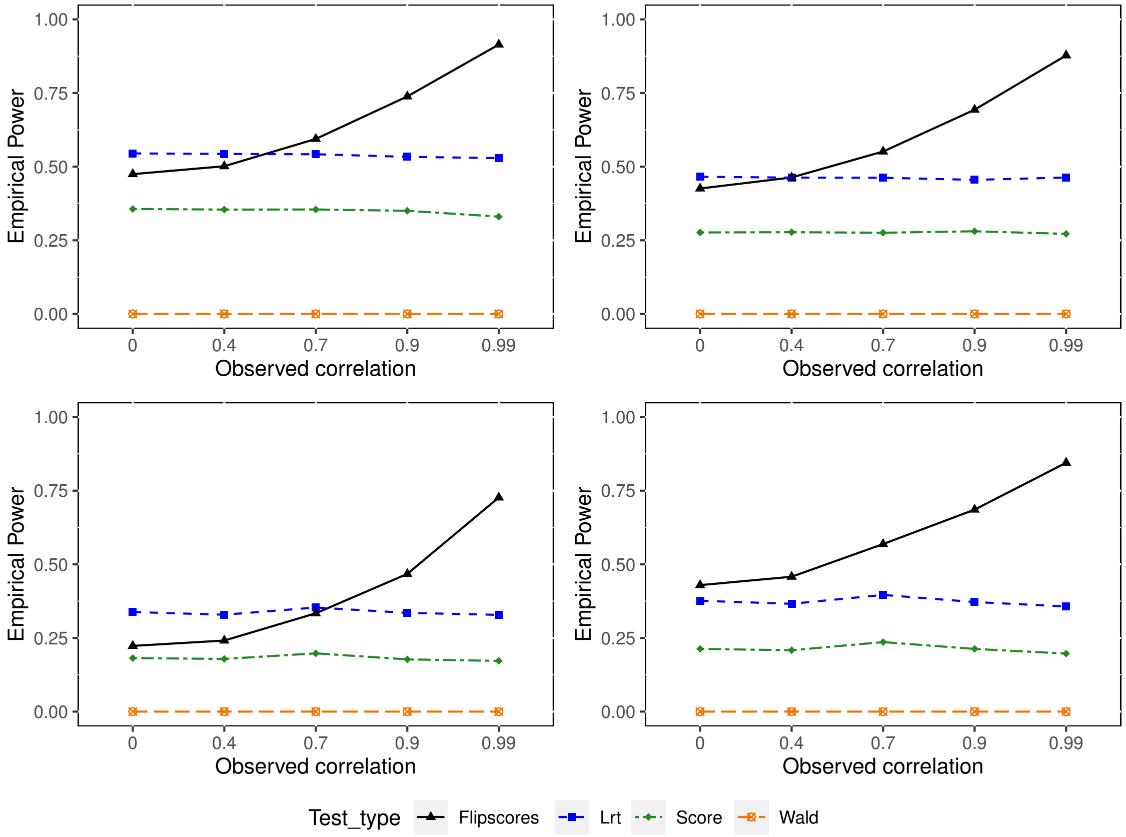

FIGURE 2.4: Logit model, multivariate. Sample size equal to 50, Power comparison

hypothesis, and the remaining 80 % have $\beta = 0$, that is, they are under H_0. We set $\gamma = 0$ (left plots) or -1 (right plots), and the correlation between covariates equal to 0 (top plots) or 0.5 (bottom plots). We use a total of 2 000 random flips, and we run a total of 1 000 simulations. The dotted lines represent the 95% simulation confidence bands. The outcomes are 0-1 variables generated as follows, where the algorithm is built in order to generate correlated 0-1 variables. We first generate the probabilities with given covariates and coefficients through the inverse of the logistic link. Then we generate multivariate normal random variables with pre-specified correlation structure. Finally, the outcome gets the value of 1 if the normal quantile of the generated probability is above the corresponding value generated from the normal distribution, and 0 otherwise.

For controlling the FWER the proposed method is used for the Flipscores approach, while the other parametric competitors are corrected with the Bonferroni-Holm procedure. Figure 2.3 and 2.5 are related to the FWER control, respectively for a sample size of 50 and 100. The x axis of the figures represents the average observed correlation between the dependent variables, while the y represents the observed FWER of the true null hypotheses. In both plots we observe similar behavior. The Flipscores is close to the nominal level, the LRT is anticonservative for low correlation, while the other two

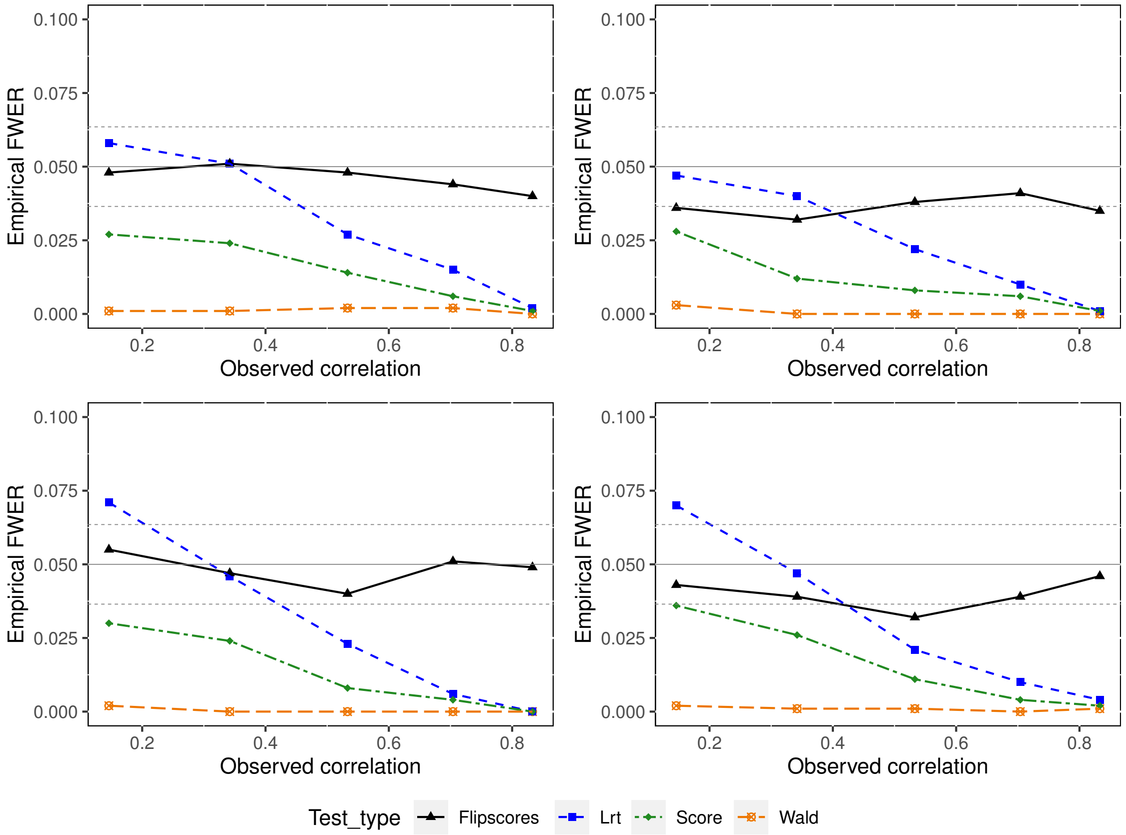

FIGURE 2.5: Logit model, multivariate. Sample size equal to 100, FWER control

methods are highly conservative.

Figure 2.4 and 2.6 represents the empirical power for the true non-null coefficients, respectively for a sample size of 50 and 100. Again, in both plots we observe similar behavior. The Flipscores has greater power, especially for higher correlation, while the other methods do not take advantage of the correlation structure. All the settings give similar results, where we observe that the proposed method seems to be more satisfactory, especially when high correlation between the responses is present.

2.5 Real-data application

In RNA sequencing data (Love *et al.*, 2014) a common aim is to find genes that are differentially expressed across a group of units. The usual analysis adopts a negative binomial regression model, since the observed target variables are counts, and overdispersion relative to the Poisson distribution is standard. However, the variance model generally assumes a fixed mean-variance structure with a common dispersion parameter among the groups of interest, which can be problematic, as we will show.

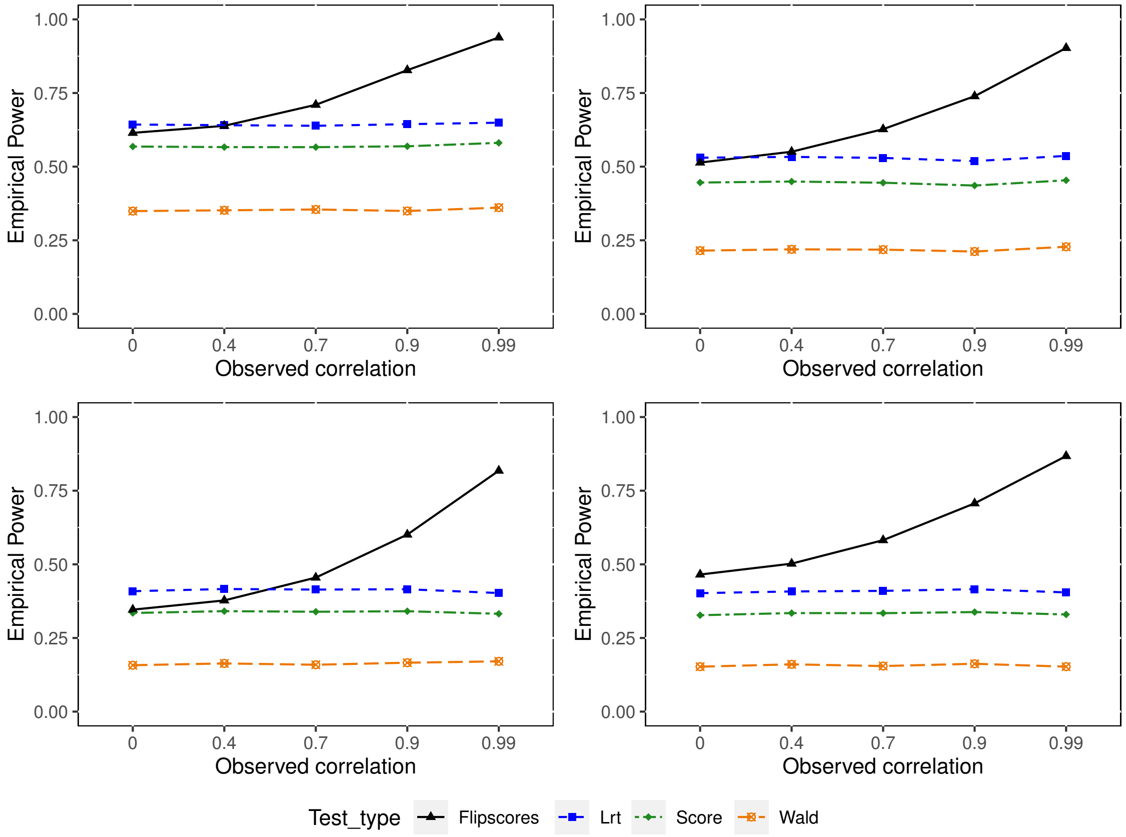

FIGURE 2.6: Logit model, multivariate. Sample size equal to 100, Power comparison

From the Cancer Genome Atlas (TCGA) (Tomczak *et al.*, 2015), we have taken the TCGA-LIHC dataset of Liver Hepatocellular Carcinoma (HCC) (Erickson *et al.*, 2016). The TCGA-LIHC consists of 20 119 genes for 344 patients with a primary tumor. We performed a very limited pre-filtering, deleting only the genes with zero total count. The target covariate is the pathological stage of the tumor. We treat it as a binary variable, splitting it between first pathological stage versus all higher stages. We further included two covariates in the fitted model: gender and age.

The state-of-the-art method for the analysis of these data uses a negative binomial model, which assumes common dispersion parameter between the two tested groups, i.e. the pathological stage. We tested this assumption for each gene using a Gamlss. The null hypothesis that dispersion did not depend on tumor stage was rejected for 5 967 out of 20 119 genes at the unadjusted 5% level. Since we would expect approximately 1 000 rejections if the DESeq2 model fits well, this gives clear indication of lack of fit of that model, at least in some of the genes. Based on the simulations in Section 1.7 we would, therefore, expect the parametric tests to be anti-conservative in this data.

Table 2.1 represents the number of rejections with the FWER at level of 10%. Given the moderate sample size of the dataset, we compare the multivariate extension of the

Method	No. of rejections
Flipscores	312
Sandwich	491
Gamlss	1 342

Table 2.1: FWER-controlled number of rejections

univariate tests which in Figure 1.3 show appreciable behavior against the hypothesis of equality of the dispersion parameter. In particular, we compare the max-T method based on sign-flips with the sandwich and Gamlss methods corrected with the Holm procedure. The Flipscores rejects a smaller number of genes compared to the other methods.

Method	True discoveries
Flipscores Sum test	6 633
Flipscores pARI	4 470
Sandwich ARI	2 342
Gamlss ARI	4 378

Table 2.2: Number of true discoveries

Table 2.2 shows the number of true discoveries for the two permutation-based method of Vesely *et al.* (2023) (Flipscores Sum test) and Andreella *et al.* (2023) (Flipscores pARI), where we choose the default parameters, and the parametric method based on the p-values computed by the sandwich (Sandwich ARI) and the Gamlss (Gamlss ARI), for a significance level of $\alpha = 0.05$. In this case we observe that the two permutation-based methods are able to obtain a greater lower bound for the proportion of true discoveries.

Chapter 3

Sign-flip tests for the Cox regression model

3.1 Cox regression model

Time to event data are particular type of data which arise in many applied fields, such as medicine, biology and demography, and they have generated a huge literature because of their peculiar structure. Let T indicate the time, assumed to be continuous, at which the random variable associated to an event of interest, for example the death, is observed. Researchers are often not interested on the underlying distribution $f(t)$, but instead they focus on the hazard rate $\lambda(t)$ which is defined as

$$\lambda(t) = \lim_{\Delta t \longrightarrow 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t},$$

while a related quantity is the cumulative hazard function, defined as

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

The quantity $\lambda(t)\Delta t$ can be seen as the approximate probability of an individual of age t experiencing the event in the next instant, given that it has not experienced the event yet. This function is useful as it uniquely determine the density function $f(t)$, the survival function, which is defined as $S(t) = 1 - F(t)$, and can describe the way in which the chance of experiencing the event changes with time. Note that $\lambda(t)$ can have any shape, the only requirement being $\lambda(t) \geq 0$.

In many practical situation the individuals are observed for a limited amount of time. The observations are denoted as censored, and in particular two censoring scheme can be considered: right censoring, where we only know that the subject has not experienced the event at a given time, and left censoring, where we only know that the subject has experienced the event before the start of the study. In this chapter we will only consider right-censoring and we will assume that it is non-informative meaning that the event and the censoring times for each i-th unit are independent. In particular, we assume to have a sample of n observation, and for each unit i we observe (Y_i, δ_i), where $Y_i = \min\{T_i, C_i\}$ is the observed time, which might be T_i - the event is observed - or C_i - the individual left the study, while $\delta_i = 1\{T_i \leq C_i\}$ is the event indicator, taking the value of 1 if the event is observed.

When some explanatory variables are observed we might be interested in studying the influence of them on the survival times. A widely used and popular model is the Cox regression model (Cox, 1972), which models the impact of the covariates on the hazard ratio and is defined as

$$\lambda(y_i|x_i, z_i) = \lambda_0(y_i)\exp\{\beta^T x_i + \gamma^T z_i\}.$$

The model is called semiparametric in the sense that the baseline hazard $\lambda_0(y_i)$ is left unspecified, while (X_i, Z_i) represents observed covariates and (β, γ) are regression parameters which have to be estimated. Hereafter, we assume that $\dim\{\beta\} = 1$ and $\dim\{\gamma\} = q$ which is treated as fixed. We will consider β as the parameter of interest, while γ is treated as a nuisance parameter.

The model is generally fitted through a partial likelihood approach, defined as

$$L_p = \prod_{i \in D} \left\{ \frac{\exp\{\beta^T x_i + \gamma^T z_i\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right\}$$

where D represents the set of uncensored observations, while R_i represents the risk set at time i, which is the set of individuals who are still under study at a time just prior to y_i. Note that this likelihood is a product over only the elements which are uncensored, equal to the cardinality of D. The use of the partial likelihood as a valid tool for inference was first justified by (Cox, 1975). However, note that it can be seen also as a profile likelihood from the full-censored likelihood (Johansen, 1983).

The maximum likelihood estimates of the parameters are obtained by maximizing the partial likelihood, which is therefore treated as a usual likelihood. In particular, the marginal scores are

$$\ell_\beta = \sum_{i \in D} \left\{ x_i - \frac{\sum_{j \in R_i} x_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right\}$$

$$\ell_\gamma = \sum_{i \in D} \left\{ z_i - \frac{\sum_{j \in R_i} z_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right\}$$

The Fisher information is minus the second derivative of ℓ. The blocks are

$$\begin{aligned}\ell_{\beta,\beta} = &\sum_{i \in D} \left\{ \frac{\sum_{j \in R_i} x_j x_j^T \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right. \\ &- \left. \left[\frac{\sum_{j \in R_i} x_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right] \left[\frac{\sum_{j \in R_i} x_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right]^T \right\}\end{aligned}$$

$$\begin{aligned}\ell_{\beta,\gamma} = &\sum_{i \in D} \left\{ \frac{\sum_{j \in R_i} x_j z_j^T \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right. \\ &- \left. \left[\frac{\sum_{j \in R_i} x_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right] \left[\frac{\sum_{j \in R_i} z_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \theta_j} \right]^T \right\}\end{aligned}$$

$$\begin{aligned}\ell_{\gamma,\gamma} = &\sum_{i \in D} \left\{ \frac{\sum_{j \in R_i} z_j z_j^T \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right. \\ &- \left. \left[\frac{\sum_{j \in R_i} z_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right] \left[\frac{\sum_{j \in R_i} z_j \exp\{\beta^T x_j + \gamma^T z_j\}}{\sum_{j \in R_i} \exp\{\beta^T x_j + \gamma^T z_j\}} \right]^T \right\}\end{aligned}$$

The maximum partial likelihood estimator has some fundamental properties. In particular, the consistency and asymptotic normality of the estimators are derived elegantly by Andersen and Gill (1982) through an approach based on counting processes theory. Classic hypothesis tests and confidence intervals are built on the same principle of generalized linear models, using the asymptotic normality of the maximum partial likelihood estimator and of the score function. We will turn back to this argument in Section 3.3.

A key assumption of the Cox regression model is the proportionality of the hazard function; it means that, given two individuals i, j, the ratio of their hazards is

$$\frac{\lambda(y_i | x_i, z_i)}{\lambda(y_j | x_j, z_j)} = \exp\{\beta^T (x_i - x_j) + \gamma^T (z_i - z_j)\},$$

which is independent from the time y. This fact means that the ratio of hazards of two subjects is a constant, that is, the hazards are proportional. Model misspecification in the Cox model related to the non-proportionality of the hazard ratio can have several causes and can be sometimes accomodate in more complex models. The main extensions regard the introduction of causal effects for clustered data - which are known as frailty models (Balan and Putter, 2020) - or the introduction of time varying coefficients (Klein *et al.*, 2003), which are beyond the scope of this chapter. Here, and in the simulation study of Section 3.4, we will consider the following issues. Unlike the generalized linear models, the omission of independent covariates will cause the bias of the maximum partial likelihood estimator, which is shrunk towards zero (Bretagnolle and Huber-Carol, 1988; Struthers and Kalbfleisch, 1986). However, if we are interested in testing the nullity of the coefficients the asymptotic distribution of the maximum likelihood estimator is preserved (Struthers and Kalbfleisch, 1986). Another issue comes when we assume a Cox regression model while the true model is an accelerate failure time model (Kalbfleisch and Prentice, 2011), which assumes the following relationship between the observed times and the covariates,

$$\log T_i = \beta^T x_i + \gamma^T z_i + \sigma\varepsilon,$$

where σ is a scale parameter while ε is an error term with a pre-specified distribution. Right-censoring can be easily accomplished in the estimation procedure. In such cases the relative importance of the explanatory variables is preserved, as long as the random censoring is independent from the observed covariates, but the quality of the variance estimator depends on the true distribution of the error term (Solomon, 1984; Struthers and Kalbfleisch, 1986). In Section 3.4 we will consider a standard normal distribution for ε, which implies a lognormal ditribution for T_i.

3.2 Connection with Poisson model

There is a close relationship between the Cox regression model just introduced and a particular Poisson regression model, as was first investigated by Holford (1980) and Laird and Olivier (1981). This connection was originally exploited to build faster algorithms for the parameter estimation. We will further use this connection in order to directly apply the results of Chapter 1 to perform hypothesis testing in the Cox regression model. We first note that the full likelihood of the Cox model is

$$L_f = \prod_{i \in n} \lambda(y_i|x_i, z_i)^{\delta_i} \exp\{-\Lambda(y_i|x_i, z_i)\}. \tag{3.1}$$

The same likelihood is obtained by treating the δ_i as outcome which are Poisson distributed in the following way. Denote $\tau_m, 1 \leq m \leq |D|$ the observed time events. For subject i, consider all the time events $\tau_1 \leq \cdots \leq \tau_j \leq y_i$; rearrange the data for each singleton i in the j lines of Table 3.1.

id	interval	event	covariates
i	1	0	x_i, z_i
i	$\vdots$	$\vdots$	x_i, z_i
i	$j-1$	0	x_i, z_i
i	j	δ_i	x_i, z_i

Table 3.1: Restructured data for subject i

In the Poisson model that will be formulated in each line the event δ_{ij} corresponds to a 0-1 Poisson variable with mean $\mu_{ij} = \lambda_j \exp\{\beta^T x_i + \gamma^T z_i\}$. The contribution to the likelihood of the first $j-1$ lines is equal to

$$e^{-\mu_{ik}} = e^{-\lambda_k \exp\{\beta^T x_i + \gamma^T z_i\}},$$

since $\delta_{ik} = 0$. For the final line, if $\delta_{ij} = 0$ the contribution is equal to

$$e^{-\mu_{ij}} = e^{-\lambda_j \exp\{\beta^T x_i + \gamma^T z_i\}},$$

while if $\delta_{ij} = 1$ - that is, the event is observed - the contribution is

$$e^{-\mu_{ij}} \mu_{ij} = e^{-\lambda_j \exp\{\beta^T x_i + \gamma^T z_i\}} \lambda_j \exp\{\beta^T x_i + \gamma^T z_i\}.$$

If we consider the j lines as independent, the total likelihood contribution for subject i is equal to

$$[\lambda_j \exp\{\beta^T x_i + \gamma^T z_i\}]^{\delta_i} e^{-\sum_{k=1}^{j} \lambda_k \exp\{\beta^T x_i + \gamma^T z_i\}}. \tag{3.2}$$

The likelihood (3.2) is equal to the full likelihood of the Cox model (3.1) and, in fact, they give the same maximum likelihood estimates. In detail, the Poisson model is fitted by first decomposing the original data as described, and then fitting a Poisson regression, which has the original regression parameters (β, γ) plus one nuisance parameter per each time episode, which corresponds to the column interval in Table 3.1, and all of them are modelled as factors.

To be more precise, Holford (1980) and Laird and Olivier (1981) point out that the Poisson model is equivalent to the Cox model with a piecewise constant baseline hazard within time events. However, the maximum likelihood estimates perfectly coincide between the Poisson model and the Cox model without further assumptions. In particular, not only the regression parameters and their corresponding standard errors are equal, but also the parameters λ_j permits to recover exactly the Breslow estimate of the baseline hazard λ_0 (Breslow, 1972).

We will further illustrate this relationship through a toy example. Assume to observe a sample of 5 observations. Table 3.2 represents the original dataset while Table 3.3 represents the reconstructed dataset for the Poisson model. For each observation we observe the time event, the censoring indicator and a covariate X. For the original dataset we fit a usual Cox regression model, where we explicit the time (y_i), the status (δ_i) and the covariate (x_i). For the reconstructed dataset we fit a Poisson model with canonical link, where we explicit the dependent variable, which is the column event, and the covariates, which are the columns x and episode; the latter is modelled as a factor. Using any statistical software it can be verified that the estimates of the regression parameter for x coincides, equal to -0.2321 with a standard error of 0.5953. Further, it can be easily verified that also the score contributions and the Fisher information coincide, and that we can recover the estimate of the baseline hazard by taking the exponential of the coefficients related to the factor episode of the Poisson formulation.

id	time	status	x
1	1.0	1	-0.86
2	2.2	1	-0.16
3	3.0	0	1.69
4	4.6	0	0.01
5	6.0	1	-1.91

Table 3.2: Toy example: original dataset

id	episode	event	x
1	1	1	-0.86
2	1	0	-0.16
2	2	1	-0.16
3	1	0	1.69
3	2	0	1.69
4	1	0	0.01
4	2	0	0.01
5	1	0	-1.91
5	2	0	-1.91
5	3	1	-1.91

Table 3.3: Toy example: reconstructed dataset

3.3 Hypothesis testing for the Cox model

Our interest is now turning on the regression parameters, considering the null hypothesis

$$H_0 : \beta = \beta_0,$$

against a one or two-sided alternative, while the other parameters are nuisance parameters, both the regression parameters γ and the episode parameters λ, which corresponds to the baseline hazard $\lambda_0(y)$ of the Cox model.

In absence of nuisance regression parameters, a large number of proposals have been done about two-sample or general k-sample problems, when the hypothesis is on the equality of the survival curves, considering different censoring schemes. These methods are either based on ranks, sign-flips or permutations, and reflect the nonparametric nature of the estimation of the survival function. A general overview is given by Klein *et al.* (2003) and Arboretti *et al.* (2018). When testing for the effect of a continuous variable in the Cox model, without nuisance covariates, a permutation-based solution is given by Sun and Sherman (1996).

In presence of nuisance regression parameters γ general nonparametric solutions are not available. Concerning the null hypothesis about the target parameter β, as for generalized linear models the parametric approach for this kind of test consists of three alternatives which have similar behavior, especially for growing sample size, respectively the Wald test, the Likelihood ratio test and the parametric Score test. These tests are analogous to the procedure described in Section 1.2. However, we will see in Section 3.4 that these tests show an anti-conservative behavior for small sample sizes; it means that they do not meet the first requirement of any statistical test, that is, the control of the type I error. It should also be remarked that a sandwich-type version of the Wald test exist, which is built following the same idea of the GLM version (Lin and Wei, 1989). This test is implemented to handle correlated observations, which we do not consider in this chapter.

Following the GLM formulation of Section 3.2 we can directly apply the standardized test of Section 1.4 which is shown to be second-moment exact, whose simulation in Section 1.7 shows appreciable behavior for small sample size. The main weakness of this approach is given by the computational burden required. Indeed, a moderate original sample size can easily turn on a massive sample size in the Poisson formulation.

To answer this issue we can build a second test, based on the partial likelihood. Indeed the use of the full likelihood of the Cox model is denied by the presence of the infinite-dimensional nuisance parameter. First of all, remember that the partial likelihood is a profile likelihood with respect to the baseline hazard. That is, it "profiles" out the infinite-dimensional parameter and hence it depends only on nuisance parameters with fixed dimension. Moreover, the effective score can be deduced from the profile likelihood (Murphy and Van der Vaart, 2000). It follows that the effective score contributions of the partial likelihood can be recovered directly from the Poisson contributions by aggregating them per episode event. We can thus think of flipping these contributions, whose cardinality is significantly smaller. Further, we can standardize this flipped test statistic to potentially improve the convergence to the nominal level of the test. We will see in Section 3.4 how the use of the partial likelihood contributions will affect the behavior of the test through a simulation study.

Some words must be spent on the asymptotics. Indeed, one of the main assumption of Chapter 1 is not met. In the Poisson model the dimension of the nuisance parameter λ grows with n. Call n' the sample size in the Poisson formulation. It can be easily shown that the dimension of the nuisance parameters is $O({n'}^{1/2})$. Under minimal regularity assumptions, Fahrmeir and Kaufmann (1985) prove the consistency of the parameters when their dimension is $o(\log n')$ and state that logarithmic or superlogarithmic growth is not admissible. This rate is not satisfied by the Cox model that we have built. Further, Burr (1994) states that the Breslow estimator (Breslow, 1972) of the hazard ratio is inconsistent, confirming the potential issues in applying our sign-flip test.

On the other hand, the use of a profile likelihood as a valid tool for statistical inference in the Cox model - and more generally for semiparametric models - is proven by Murphy and Van der Vaart (2000). The authors prove the consistency and asymptotic normality of the profile likelihood estimator by applying an approximate least favorable submodel which was proposed in their paper. Indeed, it is important to note that all the classic tests and the inference procedure are performed on the basis of the partial likelihood. Murphy and Van der Vaart (2000) states that the profile likelihood behaves like an ordinary likelihood, where the score function and the Fisher information are replaced by the effective score function and effective Fisher information with respect to the nonparametric component. Hemerik *et al.* (2020) derived the sign-flip effective score test for general parametric models. This justify the use of this test for the Cox regression model, where we only ask the nuisance regression parameters γ to be consistently estimated. It is worth noting that, over all the semiparametric models, the Cox model is somehow special, in the sense that it is possible to obtain a close form of the profile likelihood with respect to the nonparametric component.

Further, as we pointed out in Section 3.2 this likelihood and all the estimated quantities perfectly coincides with a particular aggregation of the quantities derived by the Cox model. In this sense we are able to derive the standardized sign-flip score test for both the general Poisson formulation and the partial likelihood formulation. We will further evaluate the empirical performance of the two tests based on the sign-flip approach in the next section. The extension to multivariate parameters $\beta \in \mathbb{R}^d$ and multivariate responses is straightforward following Section 1.6 and Chapter 2.

3.4 Simulation study

We perform a simulation study in order to evaluate the capability of controlling the type I error. We compare the three parametric methods - Wald test, LRT and parametric Score test - with the two Flipscores test, the one based on the full GLM formulation - Flip GLM - and the one based on the partial likelihood - Flip Partial lik. Note that for the GLM formulation we have used the standardized test - the most computational intensive procedure - while for the partial likelihood formulation we have used the effective test, that is, a less computational demanding procedure. A total of 5 000 simulations have been run. The data are generated through the inverse probability method (Bender *et al.*, 2005). The baseline hazard is a Weibull distribution with shape and scale parameters (under the complete null hypothesis) equal to 1. The censoring time are generated through an independent exponential distribution, whose parameter is chosen in order to control the average proportion of censored observations. The model contains two covariates (X, Z). In the Figures 3.1-3.4 the left panel considers a correlation of 0 between X and Z, while such correlation is 0.3 in the right panel.

Figure 3.1 shows the type I error control when testing a continuous variable. The true β is set equal to 0 and we test $H_0 : \beta = 0$ with significance level $\alpha = 0.05$. The parameter γ is set equal to 0.5. The x axis represents the sample size, while the y axis represents the empirical type I error. The dotted lines represent a 95% simulation confidence bands. We see that, even for a sample size of 60, the parametric tests are anti-conservative, beyond the confidence bands. On the other hand, the two Flipscores tests show a similar and more appreciable behavior.

Figure 3.2 shows the type I error control when testing a factor variable, with two levels. The factor has been generated by firstly generated a continuous covariate - with mean 0 - whose correlation with Z is pre-decided, and then we cut off the continuous variable X in 0 to create a 0-1 factor. The true β is set equal to 0 and we test $H_0 : \beta = 0$ with significance level $\alpha = 0.05$. The parameter γ is set equal to 0.5. Among the

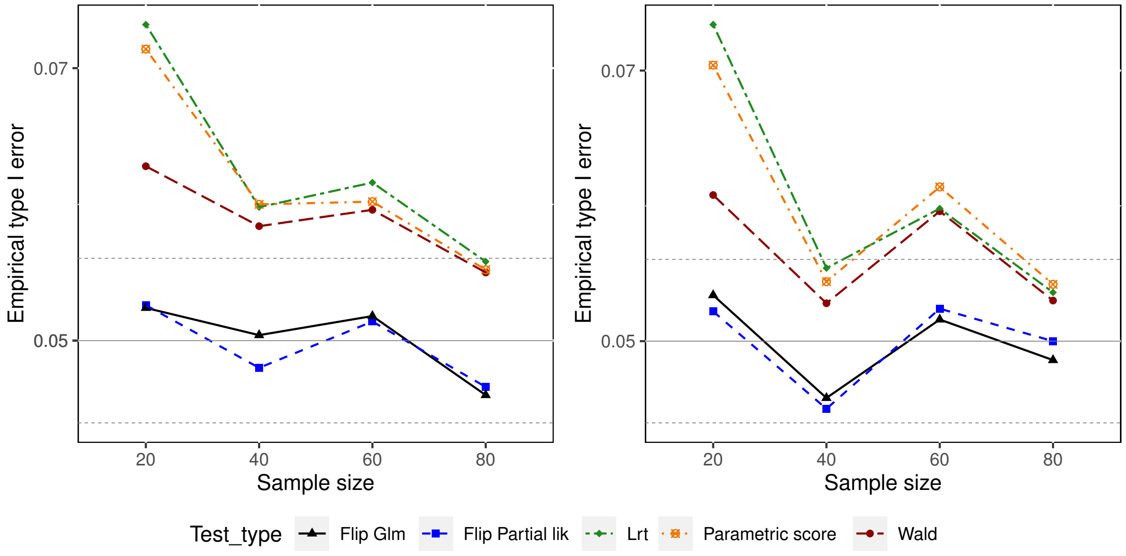

FIGURE 3.1: Type I error control for continuous covariate. Average censoring equal to 25%.

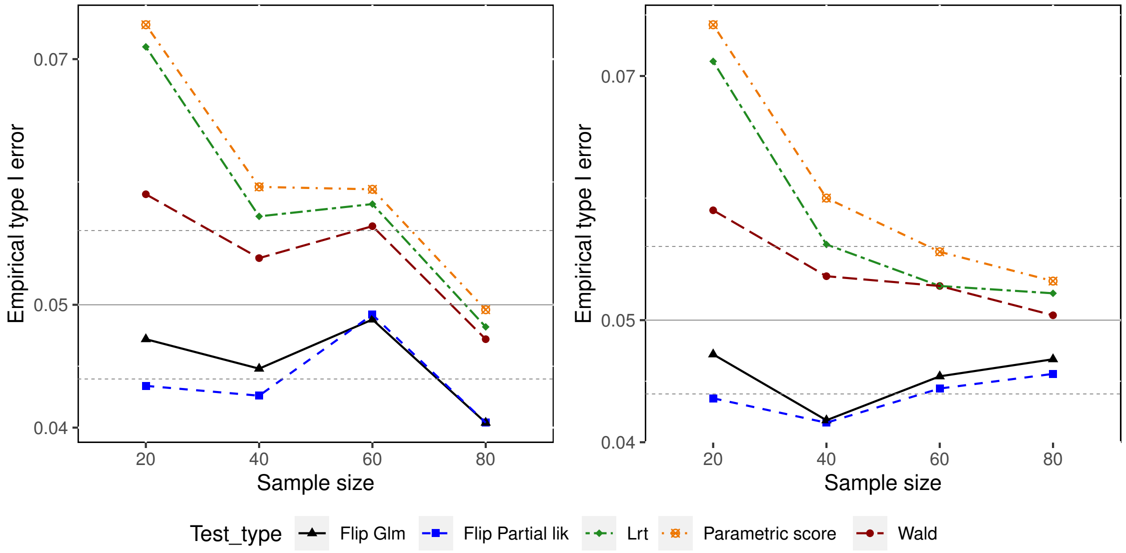

FIGURE 3.2: Type I error control for a factor covariate. Average censoring equal to 25%.

parametric tests, the Wald test is satisfying for a sample size of 40 and more, while the other two methods have slower convergence. The Flipscores tests are more satisfying, being only slightly conservative.

Figure 3.3 shows the type I error control when testing a continuous variable, when the average proportion of censored observations is 50%. The true β is set equal to 0 and we test $H_0 : \beta = 0$ with significance level $\alpha = 0.05$. The parameter γ is set equal to 0.5. Despite the fact that the effective number of observations is lower, we observe a similar behavior as in Figure 3.1. The type I error control is perfect for the

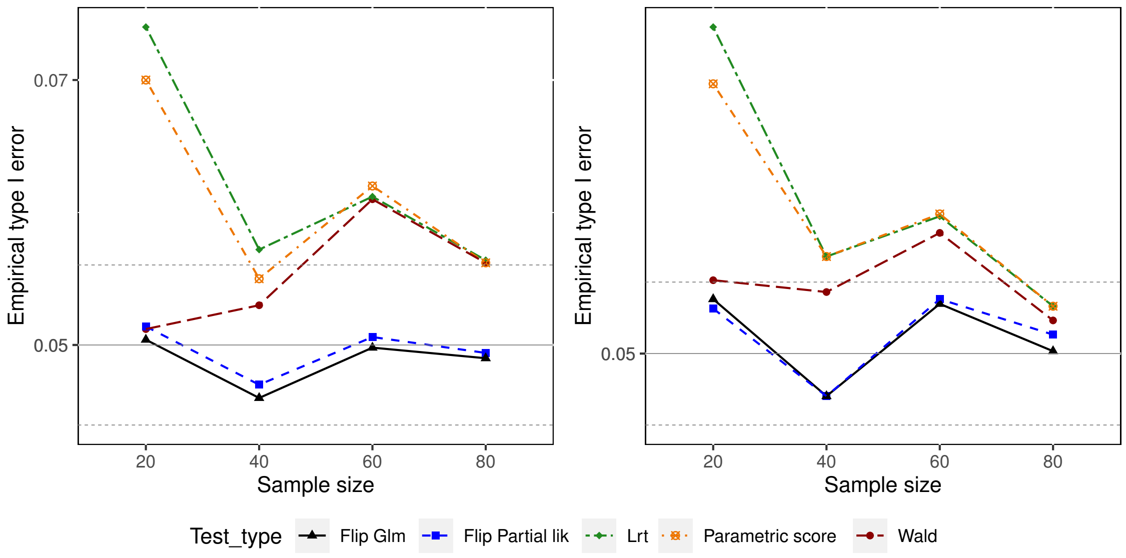

FIGURE 3.3: Type I error control for continuous covariate. Average censoring equal to 50%.

Flipscores approach, even for small sample sizes, while the parametric tests show an anti-conservative behavior, the Wald test being better than the alternatives.

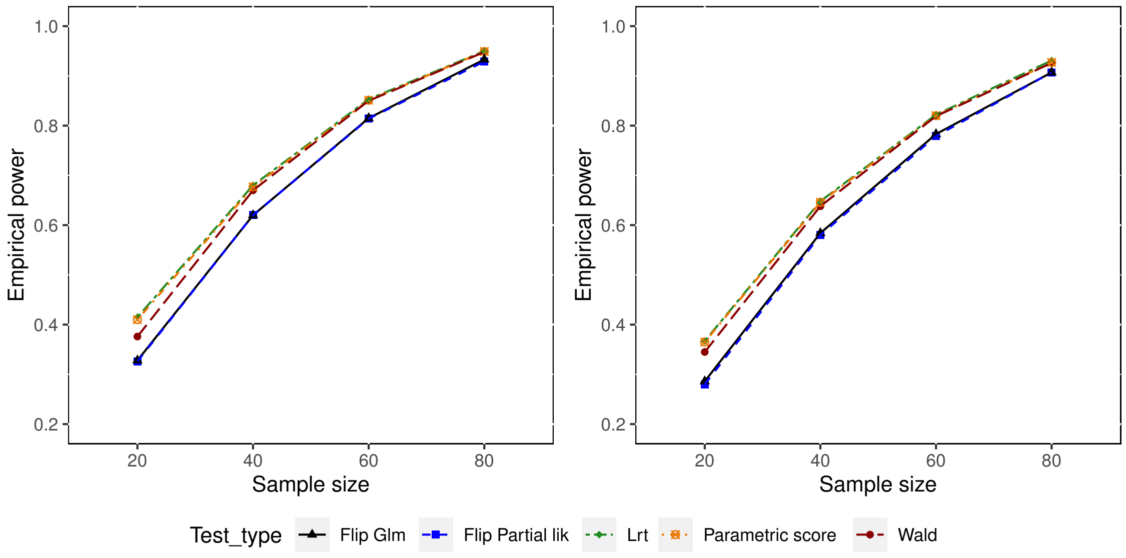

FIGURE 3.4: Power comparison for continuous covariate. Average censoring equal to 25%.

Figure 3.4 shows a power comparison. The true β is set equal to 0.5 while we test $H_0 : \beta = 0$, whose corresponding variable X is assumed continuous. The parameter γ is set equal to 0.5. The parametric tests are slightly more powerful - which is somehow expected from the anti-conservative behavior on the type I error control - but the loss of power for the Flipscores approach is appreciably small, being almost equal for the two methods.

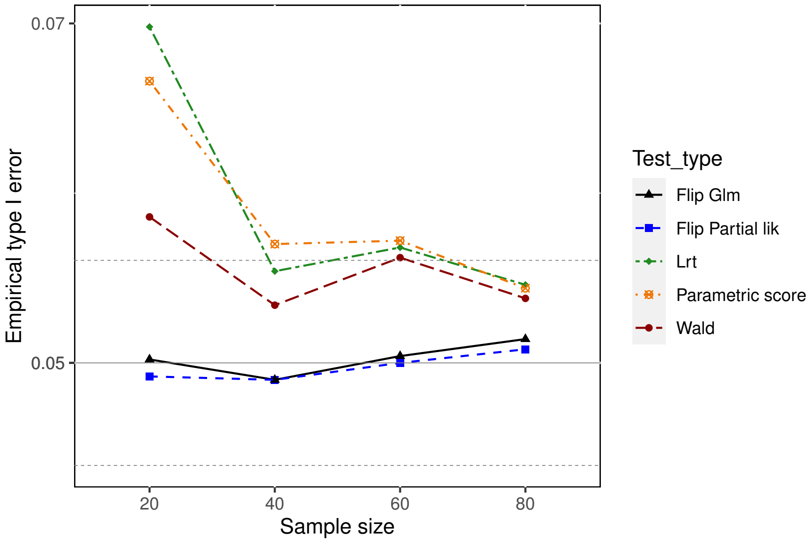

FIGURE 3.5: Type I error with omitted covariate. Average censoring equal to 25%.

Figure 3.5 represents a simulation where an omitted covariate is present, whose coefficients is equal to 0.5, and it is independent from the target covariate. The target parameter is $\beta = 0$ and we test $H_0 : \beta = 0$. Further, the model includes an extra independent nuisance variable which is considered, whose coefficient is equal to 0.5. As expected from the theory in Section 3.1, the results are close to the previous simulations. The parametric tests are anti-conservative, especially for a low sample size, while the sign-flip tests are close to the nominal level and perform similarly.

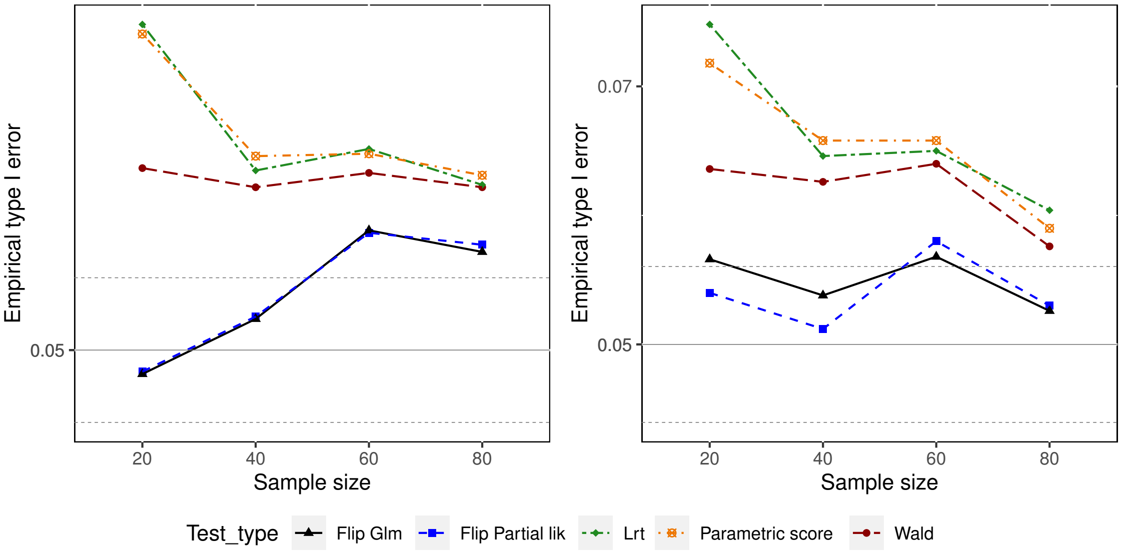

FIGURE 3.6: Type I error when the true model is AFT. Average censoring equal to 25%.

Figure 3.6 shows the result where the true model is an accelerated failure time (AFT) model, while we fit a Cox regression model. The event times are generated from a log-normal distribution, while the censoring time is generated from an independent exponential model. The target parameter is $\beta = 0$ while the nuisance parameter is set equal to 0.5. The two covariates are independent in the left plot, while the correlation is 0.3 in the right plot. We test the null hypothesis $H_0 : \beta = 0$. The parametric tests show a strong anti-conservative behavior. The sign-flip tests perform better in all cases. They are close to the nominal level for smaller sample sizes, while they are slightly anti-conservative for a sample size of 60 and also for 80 in the left plot.

Conclusions

Discussion

In this manuscript we have derived a novel approach to perform hypothesis testing in regression models, which is based on the use of sign-flips in order to obtain null-invariant transformations of the original test statistic. In this sense our approach is semiparametric, since we do a subset of assumptions compared to the standard parametric alternatives. We started from generalized linear models, where we develop a method which shows an appreciable type I error control when the model is correctly specified, while being appreciably robust when the variance is misspecified without any specific pattern. Further, we extended the method to models with multivariate response; the resampling-based solution that we propose is able to capture the unknown correlation structure of the responses, resulting in an appreciable gain of power over the parametric alternatives especially when a strong correlation is present. Finally, we broaden our approach to semiparametric models, focusing on the well-known Cox regression model. Our approach shows a good type I error even for small sample sizes, outperforming the parametric alternatives which are anti-conservative for small sample sizes.

Future directions of research

Given its duality with statistical tests, a natural question that arises is related to confidence intervals. Following the permutation principle, a natural interval is given by taking all the values for which the corresponding test is not rejected. Two issues are present; the computational burden required by this approach, since we would perform the tests over a grid of values, and the fact that, for generalized linear models, this interval is generally non-monotonic. This two open questions require further work.

Another natural extension relates to generalized linear mixed models and frailty models, where the observations are correlated, for instance where they are split in clusters. A naive approach is not possible, as the sign-flips would alter the correlation between the units, resulting in a different distribution of the flipped test statistic.

Further, the Cox regression model requires future theoretical work. Indeed, the behavior of the test based on the effective score of the partial likelihood is surprising, compared to the version for standard generalized linear models. We derived the validity of the test through an implicit argument, i.e. the use of the profile likelihood as a valid tool in semiparametric models, but a future work would go deeper into deriving more explicitly the properties.

Bibliography

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models.* New York: Wiley.

Andersen, P. and Gill, R. (1982) Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**, 1100–1120.

Andreella, A., Hemerik, J., Finos, L., Weeda, W. and Goeman, J. (2023) Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine* .

Arboretti, R., Fontana, R., Pesarin, F. and Salmaso, L. (2018) Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring. *Statistical methods in medical research* **27**(12), 3739–3769.

Azzalini, A. (1996) *Statistical Inference based on the likelihood.* Boca Raton, FL: Chapman and Hall.

Balan, T. A. and Putter, H. (2020) A tutorial on frailty models. *Statistical methods in medical research* **29**(11), 3424–3454.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics.* Boca Raton, FL: Chapman and Hall.

Bender, R., Augustin, T. and Blettner, M. (2005) Generating survival times to simulate cox proportional hazards models. *Statistics in medicine* **24**(11), 1713–1723.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.

Billingsley, P. (1986) *Probability and Measure.* New York: Wiley.

Boos, D. D. (1992) On generalized score tests. *The American Statistician* **46(4)**, 327–333.

Breslow, N. E. (1972) Discussion of the paper by d.r. cox. *Journal of The Royal Statistical Society Series B-statistical Methodology* **34**, 216–217.

Bretagnolle, J. and Huber-Carol, C. (1988) Effects of omitting covariates in cox's model for survival data. *Scandinavian journal of statistics* pp. 125–138.

Breusch, T. S. and Pagan, A. R. (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47(5)**, 1287–1294.

Burr, D. (1994) On inconsistency of breslow's estimator as an estimator of the hazard rate in the cox model. *Biometrics* pp. 1142–1145.

Chang, C.-H. and Pal, N. (2008) A revisit to the Behrens–Fisher problem: Comparison of five test methods. *Communications in Statistics—Simulation and Computation* **37(6)**, 1064–1085.

Commenges, D. (2003) Transformations which preserve exchangeability and application to permutation tests. *Nonparametric statistics* **15(2)**, 171–185.

Cook, R. D. and Weisberg, S. (1983) Diagnostics for heteroscedasticity in regression. *Biometrika* **70(1)**, 1–10.

Cox, D. (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. (1975) Partial likelihood. *Biometrika* **62(2)**, 269–276.

Davidson, R. and Flachaire, E. (2008) The wild bootstrap, tamed at last. *Journal of Econometrics* **146(1)**, 162–169.

Eicker, F. (1967) Reducing tcb complexity for security-sensitive applications: three case studies. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* pp. 59–82.

Erickson, B. J., Kirk, S., Lee, Y., Bathe, O., Kearns, M., Gerdes, C., Rieger-Christ, K. and Lemmerman, J. (2016) The cancer genome atlas liver hepatocellular carcinoma collection (tcga-lihc) (version 5) [data set]. *The Cancer Imaging Archive* .

Fahrmeir, L. and Kaufmann, H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13**(1), 342–368.

Firth, D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1), 27–38.

Fisher, R. A. (1935) The fiducial argument in statistical inference. *Annals of Eugenics* **6**, 391–398.

Fisher, R. A. (1941) The asymptotic approach to Behrens' integral with further tables for the d test of significance. *Annals of Eugenics* **11**, 141–172.

Freedman, D. A. (2006) On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician* **60(4)**, 299–302.

Glejser, H. (1969) A new test for heteroskedasticity. *Journal of the American Statistical Association* **64(325)**, 316–323.

Goeman, J. and Solari, A. (2011) Multiple testing for exploratory research. *Statistical Science* **26(4)**, 584–597.

Goeman, J. J., Hemerik, J. and Solari, A. (2021) Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics, Ann. Statist.* **49**(2), 1218–1238.

Goeman, J. J., Meijer, R. J., Krebs, T. J. and Solari, A. (2019) Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* **106**(4), 841–856.

Goeman, J. J. and Solari, A. (2014) Tutorial in biostatistics: multiple hypothesis testing in genomics. *Statistics in Medicine* **33**(11), 1946–1978.

Goldfeld, S. M. and Quandt, R. E. (1965) Some tests for homoscedasticity. *Journal of the American Statistical Association* **60**(310), 539–547.

Hall, W. J. and Mathiason, D. J. (1990) On large-sample estimation and testing in parametric models. *International Statistical Review* **58**(1), 77–97.

Hemerik, J. and Goeman, J. J. (2018) Exact testing with random permutations. *TEST* **27**, 811–825.

Hemerik, J., Goeman, J. J. and Finos, L. (2020) Robust testing in generalized linear models by sign flipping score contributions. *Journal of The Royal Statistical Society Series B-statistical Methodology* **82(3)**, 841–864.

Holford, T. (1980) The analysis of rates and of survivorship using log-linear models. *Biometrics* **36(2)**, 299–305.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* pp. 65–70.

Hotelling, H. (1931) The generalization of student's ratio. *Annals of Mathematical Statistics* **2**, 54–65.

Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* pp. 221–233.

Huber, P. J. and Ronchetti, E. M. (2009) *Robust Statistics*. New York: Wiley.

Huh, M.-H. and Jhun, M. (2001) Random permutation testing in multiple linear regression. *Communications in Statistics - Theory and Methods* **30**(10), 2023–2032.

Jarque, C. M. and Bera, A. K. (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* **6**(3), 255–259.

Johansen, S. (1983) An extension of cox's regression model. *International Statistical Review* **51**, 258–262.

Kalbfleisch, J. D. and Prentice, R. L. (2011) *The statistical analysis of failure time data*. John Wiley & Sons.

Kauermann, G. and Carroll, R. J. (2000) The sandwich variance estimator: efficiency properties and coverage probability of confidence intervals. *Discussion Paper 189* .

Kherad-Pajouh, S. and Renaud, O. (2010) An exact permutation method for testing any effect in balanced and unbalanced fixed effect anova. *Computational Statistics & Data Analysis* **54**(7), 1881 – 1893.

Kim, S.-H. and Cohen, A. S. (1998) On the Behrens–Fisher problem: a review. *Journal of Educational and Behavioral Statistics* **23(4)**, 356–377.

Klein, J. P., Moeschberger, M. L. *et al.* (2003) *Survival analysis: techniques for censored and truncated data*. Volume 1230. Springer.

Laird, N. and Olivier, D. (1981) Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* **76(374)**, 231–240.

Lehmann, E. L. and Romano, J. P. (2005) *Testing statistical hypotheses.* New York: Springer.

Lehmann, E. L. and Romano, J. P. (2012) *Generalizations of the familywise error rate.* Springer.

Lin, D. and Wei, L. (1989) The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association* **84(408)**, 1074–1078.

Long, J. S. and Ervin, L. H. (2000) Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician* **54(3)**, 217–224.

Love, M. I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15(12)**, 1–21.

Maas, C. J. M. and Hox, J. J. (2004) Robustness issues in multilevel regression analysis. *Statistica Neerlandica* **58(2)**, 127–137.

Magnus, J. and Neudecker, H. (2019) *Matrix Differential Calculus with Applications in Statistics and Econometrics.* New York: Wiley.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**(3), 655–660.

Marohn, F. (2002) A comment on locally most powerful tests in the presence of nuisance parameters. *Communications in Statistics - Theory and Methods* **31(3)**, 337–349.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models.* Boca Raton, FL: Chapman and Hall.

Meinshausen, N., Maathuis, M. H., Bühlmann, P. *et al.* (2011) Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics* **39**(6), 3369–3391.

Murphy, S. A. and Van der Vaart, A. W. (2000) On profile likelihood. *Journal of the American Statistical Association* **95(450)**, 449–465.

Neyman, J. and Scott, E. L. (1948) Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Pesarin, F. (2001) *Multivariate permutation tests: with applications in biostatistics.* Chichester: Wiley.

Rao, C. R. (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* **44**(1), 50–57.

Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C* **54**(3), 507–554.

Rochon, J., Gondan, M. and Kieser, M. (2012) To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology* **12(81)**.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A. and Goeman, J. J. (2018) All-resolutions inference for brain imaging. *Neuroimage* **181**, 786–796.

Schaarschmidt, F., Ritz, C. and Hothorn, L. (2022) The tukey trend test: Multiplicity adjustment using multiple marginal models. *Biometrics* **78**, 789–797.

Simes, R. J. (1986) An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73**(3), 751–754.

Solari, A., Finos, L. and Goeman, J. J. (2014) Rotation-based multiple testing in the multivariate linear model. *Biometrics* **70**(4), 954–961.

Solomon, P. J. (1984) Effect of misspecification of regression models in the analysis of survival data. *Biometrika* **71**(2), 291–298.

Struthers, C. A. and Kalbfleisch, J. D. (1986) Misspecified proportional hazard models. *Biometrika* **73**(2), 363–369.

Sun, Y. and Sherman, M. (1996) Some permutation tests for survival data. *Biometrics* pp. 87–97.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology* **19**, A68 – A77.

Tukey, J. W. (1953) The problem of multiple comparisons. *Multiple comparisons* .

Van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Vesely, A., Finos, L. and Goeman, J. J. (2023) Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85**(3), 664–683.

Wald, A. (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society* **54**(3), 426–482.

Westfall, P. H. and Troendle, J. F. (2008) Multiple testing with minimal assumptions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50**(5), 745–755.

Westfall, P. H. and Young, S. S. (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons.

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

Wilks, S. (1962) *Mathematical statistics.* Chichester: Wiley.

Wilks, S. S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics* **9**(1), 60–62.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. and Nichols, T. E. (2014) Permutation inference for the general linear model. *Neuroimage* **92**, 381–397.

Riccardo De Santis

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 346 6697181
e-mail: riccardo.desantis.1@phd.unipd.it

Current Position

Since October 2020; (expected completion: May 2024)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Variance-invariant inference for regression models
Supervisor: Prof. Livio Finos
Co-supervisor: Prof. Jelle J. Goeman

Research interests

- Hypothesis testing
- Multiple testing

Education

October 2017 – April 2020
Master (*laurea specialistica/magistrale*) degree in Scienze Statistiche per le Indagini Campionarie.
University of Siena, Faculty of Economics and Statistics
Title of dissertation: "Variance estimation for complex measures from complex surveys"
Supervisor: Prof. Gianni Betti
Final mark: 110/110 con lode

October 2014 – October 2017
Bachelor degree (*laurea triennale*) in Scienze Economiche e Bancarie.
University of Siena, Faculty of Economics and Statistics
Title of dissertation: "Calcolo delle scale di equivalenza utilizzando gli housing costs: un confronto tra un panel bilanciato e uno pseudo panel "
Supervisor: Prof. Gianni Betti
Final mark: 110/110 con lode.

Visiting periods

February 2022 – July 2022
LUMC, Department of Biomedical Data Sciences,
Leiden, The Netherlands .
Supervisor: Prof. Jelle J. Goeman

Work experience

January 2020 – October 2020
Employer.
Numeria Srl – Istituto degli Innocenti di Firenze.

Computer skills

- R and Rstudio
- Latex
- Python

Language skills

Italian: native; English: good; French: basic.

Publications

Articles in journals
Andreella, A., De Santis, R., Vesely, A. and Finos, L (2023). . Procrustes-based distances for exploring between-matrices similarity. *Statistical Methods & Applications* **32**, 867–882.

Conference presentations

De Santis, R. (2023). Title Sign-flip test for coefficients in the Cox regression model). *International Society for Clinical Biostatistics (ISCB) 2023*, Milano, Italia, 27-31 August 2023.

De Santis, R. (2022). Inference in generalized linear models with robustness to misspecified variances *International Conference on Multiple Comparison Procedures (MCP) 2022*, Bremen, Germany, 30 August-2 September 2022.

De Santis, R. (2022).Conditional tests for generalized linear models *Società Italiana di Statistica (SIS) 2022*, Caserta, Italia, 22-24 June 2022.

Teaching experience

October 2023 – December 2023
Modelli statistici 2
Degree In Statistical Sciences
Lab, 16 hours
University of Padova
Instructor: Prof. Alessandra Salvan

References

Prof. Livio Finos
University of Padova
Address
e-mail: livio.finos@unipd.it

Prof. Jelle J. Goeman
Leiden University Medical Center
Address
e-mail: j.j.goeman@lumc.nl