2022-08-01

A Computationally Efficient Wald Test in M-Estimation

Denisse Urenda Castañeda
University of Texas at El Paso

A COMPUTATIONALLY EFFICIENT WALD TEST IN M-ESTIMATION

DENISSE URENDA CASTAÑEDA

Master's Program in Statistics

APPROVED:

Xiaogang Su, Ph.D., Chair

Michael Pokojovy, Ph.D.

Tzu-Liang (Bill) Tseng, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School

to my

lovely husband and children

with love

A COMPUTATIONALLY EFFICIENT WALD TEST IN M-ESTIMATION

by

DENISSE URENDA CASTAÑEDA

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Science

THE UNIVERSITY OF TEXAS AT EL PASO

August 2022

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Under the maximum likelihood framework, three asymptotic overall tests have been well developed in generalized linear models (GLM) for testing the single null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, namely, the Wald test, Likelihood Ratio Test (LRT) and Score test also known as the Lagrange Multiplier test (LM). Modified versions of Wald, LR and LM tests can also be found for testing the significance of a portion of the parameter $\boldsymbol{\theta}$, i.e., if $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ it is of interest to test $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$. However, with the constant increase of dimensionality in data, the three tests becomes unfeasible to compute. The computational cost one has to pay seems to be unrealistic and difficult or even untractable.

The approach taken in this document to deal with this issue follows the profile likelihood framework which consists of partitioning the p-dimensional parameter vector $\boldsymbol{\theta}$ into two parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of dimension q and $p - q$, respectively, estimate $\boldsymbol{\theta}_1$ under H_0, say $\tilde{\boldsymbol{\theta}}_1$, and use $\tilde{\boldsymbol{\theta}}_1$ to estimate $\boldsymbol{\theta}_2$. With this approach, one could reduce considerably the execution time when estimating a big number of parameters in the model without losing the asymptotic properties and the power of the traditional tests. Also, one could test the null hypothesis even if the dimension of $\boldsymbol{\theta}$ is moderately bigger than the sample size n as long as both q and $p - q$ are smaller than n.

This document is organized as follows. Chapter 2 gives an extensive background where topics as linear regression, generalized linear models, profile likelihood, and stochastic convergence are covered. Chapter 3 describes the two proposed methods and shows the derivation of the asymptotic distribution. Several applications are also discussed at the end of this chapter. In chapter 4, simulations to study the empirical distribution, power, and size of the proposed tests will be performed as well as the execution time. Comparison of the

proposed methods and ordinary counterparts will be done. In chapter 5, it will be explored the practical use of the proposed method with the use of a real data, one for each of the three models considered in the simulation. Comparison of the performance among the ordinary and proposed tests is made. Finally, in Chapter 6, summary of the procedure followed to derive the proposed tests is made. Advantages and disadvantages of the proposed tests are stated. Conclusions and future work will be discussed.

Chapter 2

Background

In this chapter basic concepts and methods will be introduced starting by linear regression models and the estimation of coefficients by the two most common frameworks, namely, ordinary least squares and maximum likelihood. Generalized linear models will be described with the aim of concepts in exponential family of distributions. Derivation of general tests for testing single hypothesis such as likelihood ratio, Wald and score tests will be conducted. Concepts on stochastic convergence will be given and used to introduce M and Z-estimators. Chapter closes with the definition of consistency and the derivation of the asymptotic normality of Z-estimators and Maximum Likelihood Estimators.

2.1 Linear Regression Models

The framework for linear regression models is one of the most known and developed ones. In a linear model, it is assumed that the continuous response variable Y is linearly related to a set of non-random predictor variables X_j plus a random term ϵ. When only one predictor variable is used, the model can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, \ldots, n$, where y_i is the ith realization of the response random variable Y, x_i is the ith observation of predictor variable X, β_0 and β_1 are unknown parameters to be estimated and ϵ_i is the ith random error. It is also assumed that the ϵ_is are independent and identically distributed such that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ where σ is regarded as an unknown constant. Two general approaches regarding the predictors can be considered. One could

assume either the predictors to be non-random or treat them as random as well. The former approach is taken in the discussion unless otherwise stated. According to the assumptions, it can be deduced that $\mu_i = \mathrm{E}(Y_i) = \beta_0 + \beta_1 X_i$, the Y_is are independent $\mathcal{N}(\mu_i, \sigma^2)$.

One of the approaches to estimate β_0 and β_1 is by least squares, which attempts to minimize the criterion function

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2,$$

known as sum of squares. Taking partial derivatives of Q with respect to β_0 and β_1, equating to zero and solving for β_0 and β_1 simultaneously yield the least squares estimators

$$b_0 = \bar{y} - b_1 \bar{x}, \quad b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$. An estimate for σ^2 can be found by using the residual sum of squares (SSE)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the ith fitted value of y_i defined as $\hat{y}_i = b_0 + b_1 x_i$ and e_i is the e_i is the ith residual of y_i defined as $e_i = y_i - \hat{y}_i$. One can prove that $E(SSE) = \sigma^2(n-2)$ and therefore $SSE/(n-2)$ becomes a natural unbiased estimator of σ^2, i.e., the estimator of σ^2 is

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This quantity is also referred to as the *mean square residuals* and denotes as MSE.

Another approach is to use the maximum likelihood framework which consists of maximizing the likelihood or log-likelihood function of $Y_1, \ldots, Y_n$, i.e., maximizing

$$\ell = \ell(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i; \beta_0, \beta_1, \sigma^2).$$

Under the assumption that $Y_i \sim N(\mu_i, \sigma^2)$

$$\ell = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

which equating to zero yields to the sum of squares Q. Therefore, the maximum likelihood estimators (MLEs) of β_0 and β_1, $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively, are the same as the ones obtained in least squares. Additionally, one can easily estimate the variance σ^2 as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

an biased estimator of σ^2. One can note immediately that $\hat{\sigma}^2 = (n-2)s^2/n$.

To make inferences about the parameters β_0 and β_1, one can derive the sampling distribution of b_0 and b_1 by noticing that both b_0 and b_1 are linear combinations of normal random variables

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} a_i y_i$$

where $a_i = (x_i - \bar{x})/\sum_{i=1}^{n}(x_i - \bar{x})^2$. Since x_i is not random and $\bar{y}$ is normal, $b_0 = \bar{y} - b_1\bar{x}$ is also a linear combination of normal random variables. It is easy to prove that both b_0 and b_1 are unbiased estimators of β_0 and β_1, respectively and that

$$\mathrm{Var}(b_0) = \frac{\sigma^2\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

so that

$$b_0 \sim \mathcal{N}\left(b_0, \frac{\sigma^2\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right), \quad b_1 \sim \mathcal{N}\left(b_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Regarding the multivariate linear regression model, it is assumed that

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$

where Y_i is the ith realization of the response variable Y and X_{ij} is the ith realization for the jth predictor variable X_j. In matrix notation, the model can be written as compact as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix},$$

where $\mathbf{X}$ is called the design matrix, $\boldsymbol{\beta}$ is the unknown vector of parameter to be estimated, $\mathbf{Y}$ is a random vector of responses and $\boldsymbol{\epsilon}$ is the random vector of errors following multivariate normal distribution with mean zero and variance-covariance matrix $\sigma^2\mathbf{I}$, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$.

To estimate $\boldsymbol{\beta}$ one can follow any of the two approaches mention above: least squares or maximum likelihood framework. For the least square estimates, it is needed to minimize

$$Q(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

with respect to $\boldsymbol{\beta}$, which yield on the least square estimator $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ with fitted vector $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$. The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the *hat matrix* and plays an important role in model diagnosis. The residual vector is defined as $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. This can be used, as before, to estimate σ^2 through the residual sum of squares

$$SSE = \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Last equality holds because $\mathbf{H}$, and therefore $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent. It can be proved that

$$\mathrm{E}(SSE) = \sigma^2(n - p - 1), \quad \mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}, \quad \mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

which can be used to make inferences about individual parameter coefficients in a similar fashion as in simple linear regression.

Linear regression models can be more flexible than they look. X_j can be transform in many ways as long as the parameter coefficients remain linear. They can represent categorical variable by introducing dummy variables. Interaction between predictor variables can also be included in the model through the product of variables. Another advantage of linear models is its interpretability. Since the response variable is directly related to a linear combination of predictors, it can be easy to interpret what each coefficient means and the effect that a change in the predictor may produce in the response variable.

One of the principal assumptions in linear models is normality of the error term. Without this assumption, inferences regarding the estimation coefficient could be invalid. Sometimes it is also inappropriate to use a linear model to relate the mean of the response variable with one or several predictor variables since the relation may not be linear. Another problem is multicollinearity which refers to high correlation among predictor variables. Multicollinearity may produce larger standard errors and so less reliable results. The violation of the assumption that the variance of the error term is constant is also a big problem. This is also referring to heteroscedasticity. Test result may not be valid since all theory has been derived on base of a homoscedasticity assumption.

Several other models have been developed to fix these issues such as weighted linear regression to deal with heteroscedasticity, polynomial regression to deal with non-linearity, partial least squares regression to deal with multicollinearity, or *generalized linear models* (GLM) to deal with non-linear relationships, heteroscedasticity, and/or discrete response variables. All the assumption in GLMs will be discuss in the next sections but one of the most important is that the response variable follows a distribution belonging to the exponential family.

2.2 Exponential Family of Distributions

Distribution functions from the exponential family plays a main role in generalized linear models. A distribution function f belongs to the single-parameter exponential family if it can be rewritten in the form

$$f(x, \theta) = h(x)g(\theta)\exp[a(x)b(\theta)] \tag{2.1}$$

where $a(x)$ and $h(x) > 0$ are continuous functions independent of the parameter value θ and $b(\theta)$ and $g(\theta) > 0$ are continuous function of θ free of x. Alternatively

$$f(x, \theta) = \exp[a(x)b(\theta) + c(\theta) + d(x)] \tag{2.2}$$

where $c(\theta) = \log g(\theta)$ and $d(x) = \log h(x)$. The function $b(\theta)$ is called the *natural parameter* and sometimes denoted as $\eta(\theta)$, $a(x)$ is known as the *sufficient statistic* and sometimes wrote as $T(x)$. When $a(x) = x$, f is said to be in *canonical form*. When the distribution function has more than one parameter which are of no interest, they are referred to as *nuisance parameters*, regarded as known and can be absorbed in any of the functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, or $d(\cdot)$.

One can prove that, if f is twice differentiable and the derivatives are interchangeable with the integral sign,

$$\mathrm{E}[a(X)] = -\frac{c'(\theta)}{b'(\theta)}, \quad \mathrm{Var}[a(X)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \tag{2.3}$$

Proof can be found in [5].

The log-likelihood function of a random variable X which distribution belongs to the exponential family is then

$$\ell(\theta, X) = a(X)b(\theta) + c(\theta) + d(X)$$

and for n random variables $X_1, \ldots, X_n$ the log-likelihood function becomes

$$\ell(\boldsymbol{\theta}, \boldsymbol{X}) = \sum_{i=1}^{n} a_i(X_i)b_i(\theta_i) + \sum_{i=1}^{n} c_i\theta_i + \sum_{i=1}^{n} d_i(X_i)$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$, $\boldsymbol{X} = [X_1, \ldots, X_n]$, $X_i \sim f_i(x_i, \theta_i)$ and

$$f_i(x_i, \theta_i) = \exp[a_i(x_i)b_i(\theta_i) + c_i(\theta_i) + d_i(x_i)]$$

Not only for exponential families but for any distribution function, the partial derivative of ℓ with respect to θ satisfies $\mathrm{E}[\dot{\ell}(\theta, X)] = 0$ (See proof in [5]). Moreover, the variance of $\dot{\ell}(\theta, X)$ is what is known as the *expected information matrix* or *Fisher's information* of X and it is denoted as $\mathcal{I}(\theta)$. In [5] it has been shown that the information matrix is the negative of the expectation of the matrix of second partial derivatives of ℓ, $-\mathrm{E}[\ddot{\ell}(\theta, X_i)]$. The information matrix plays an important role in making inferences as will be seen in next sections. Also, the information matrix of a random sample drawn from a distribution

belonging to the exponential family can be shown to be negative definite for all θ which have a direct implication on the concavity of ℓ. Indeed, all distribution in the exponential family are called log-concave since their log-likelihood is concave. This result will be useful when investigating the asymptotic normality of MLEs in exponential families.

Some distributions belonging to the exponential family are the Normal, Binomial, and Poisson distributions. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$f(x, \mu) = \exp\left[x\left(-\frac{\mu}{\sigma^2}\right) - \frac{\mu}{2\sigma^2} - \frac{x^2}{2\sigma^2} - 2\log(2\pi\sigma^2)\right]$$

Here σ^2 is considered as known and therefore regarded as a nuisance parameter. If $X \sim \text{Binom}(n, \pi)$ then

$$f(x, \pi) = \exp\left[x \log\frac{\pi}{1-\pi} + n\log(1-\pi) + \log\binom{x}{n}\right].$$

For the binomial distribution, π is the parameter of interest and n is regarded as constant. Finally, if $X \sim \text{Poiss}(\lambda)$ then

$$f(x, \lambda) = \exp[x\log\lambda - \lambda - \log x!].$$

All these distributions are in the canonical form, and it can be easily verified that their expectations and variances satisfy (2.3). Other distributions belonging to the exponential family are the exponential, gamma, beta, and geometric distributions but some of them are not in the canonical form. Table 2.1 shows the principal components for some distributions belonging to the exponential family.

For a parameter vector $\boldsymbol{\theta}$, f belongs to a vector exponential family if it can be rewritten as

$$f(x, \boldsymbol{\theta}) = h(x)g(\boldsymbol{\theta})\exp[\mathbf{b}^T(\boldsymbol{\theta})\mathbf{a}(x)]$$

where $\mathbf{a}$ and $\mathbf{b}$ are vector-valued functions. It is also common to write f as

$$f(x, \boldsymbol{\theta}) = h(x)\exp[\boldsymbol{\eta}^T(\boldsymbol{\theta})\mathbf{T}(x) - \psi(\boldsymbol{\theta})]$$

Table 2.1: Components of generalized linear models in common distributions

Distribution	**Parameter** θ	**Nuisance parameter**	**Sufficient statistic** T	**Natural parameter** η	**Inverse**
Normal	μ	σ^2	x	$-\frac{\mu}{\sigma^2}$	$-\sigma^2\eta$
Binomial	π	n	x	$\log\frac{\pi}{1-\pi}$	$\frac{e^\eta}{1+e^\eta}$
Poisson	λ		x	$\log\lambda$	e^η
Bernoulli	π		x	$\log\frac{\pi}{1-\pi}$	$\frac{e^\eta}{1+e^\eta}$
Exponential	λ		x	$-\frac{1}{\lambda}$	$-\frac{1}{\eta}$

In this case, $\mathbf{T}(x)$ is called the natural sufficient statistic. If the natural parameter $\boldsymbol{\eta}(\boldsymbol{\theta})$ is a one-to-one function of $\boldsymbol{\theta}$, f can be reparametrized so that

$$f(x, \boldsymbol{\eta}) = h(x)\exp[\boldsymbol{\eta}^T\mathbf{T}(x) - \psi^*(\boldsymbol{\eta})]$$

It can be prove that $\mathrm{E}[T(X)] = \partial\psi^*(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ and $\mathrm{Var}[T(X)] = \partial^2\psi^*(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^T\boldsymbol{\eta}$.

2.3 Generalized Linear Models

Linear regression models relate a response variable Y with the predictor variables $X_1, \ldots, X_p$ in such a way that the expectation of the response is equal to a linear combination of the predictors, i.e., $\boldsymbol{\mu} = \mathrm{E}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$. The main idea of generalized linear models is to allow this relationship to be nonlinear by relating Y with the predictors $X_1, \ldots, X_p$ through a function of its mean. The most popular models are the linear regression (normal), logistic (binomial) and the log-linear (Poisson) models.

In generalized linear models there are a couple of assumptions that need to be satisfied. First, it is assumed that $Y_1, \ldots, Y_n$ are independent and follow the same distribution function belonging to the exponential family which is fully specified up to the parameters in $\boldsymbol{\theta}$. This is, $Y_i \sim f(y_i, \theta_i)$, where f has the form in (2.2) for all $i = 1, \ldots, n$. It is also assumed

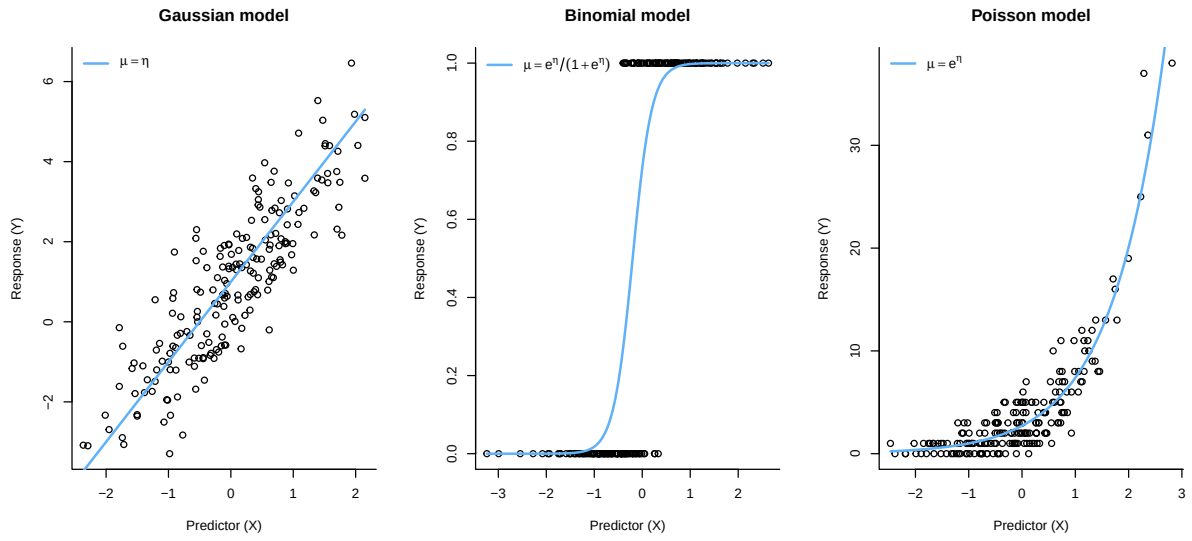

Figure 2.1: Three generalized linear models

that f is in the canonical form so that

$$f(y_i, \theta_i) = \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)], \quad i = 1, \ldots, n$$

and that the expectation of Y_i, μ_i, is related to a linear combinations of predictor variables through a monotone and continuous function g called the *link function*, i.e., $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$, where $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{ip}]^T$ is the ith row of the design matrix $\mathbf{X}$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$ is the vector of unknown parameters to be estimated.

For linear regression, the link function is the identity, i.e., $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$; for the logistic model, the link function is what is called the logit function, this is, $\log \frac{\mu_i}{1-\mu_i} = \mathbf{x}_i^T \boldsymbol{\beta}$; and finally the log-linear model uses the logarithm as the link function, $\log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$. See Table 2.1 and Figure 2.1.

Under the maximum likelihood framework, one can estimate $\boldsymbol{\beta}$ by maximizing the log-likelihood function of $Y_1, \ldots, Y_n$, this is, maximizing

$$\ell(\boldsymbol{\theta}, \mathbf{Y}) = \sum_{i=1}^{n} \ell_i(\theta_i, Y_i)$$

The derivative of the log-likelihood function with respect to β_j is what is called the *score*

function and can be verified that

$$S_j = \frac{\partial \ell(\boldsymbol{\theta}, \mathbf{Y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{Y_i - \mathrm{E}(Y_i)}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) x_{ij}$$

It can also be proved that

$$\mathrm{E}(S_j) = 0, \quad \mathrm{cov}(S_j, S_k) = \sum_{i=1}^{n} \frac{x_{ij} x_{jk}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

for all $j, k = 1, \ldots, p$. In matrix form, the variance-covariance matrix can be rewritten as $\mathrm{cov}(\mathbf{S}) = \mathbf{X}^T \mathbf{W} \mathbf{X} = (\mathcal{I}_{jk})$ where $\mathbf{W}$ is the $n \times n$ diagonal matrix with diagonal components

$$w_{ii} = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2,$$

and $\mathcal{I}_{jk}$ are the components of the information matrix $\mathcal{I}$. It is needed to find $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]$ for which $\mathbf{S} = [S_1, .., S_p]^T$ is the zero vector, i.e., $\mathbf{S}(\boldsymbol{\beta}) = \mathbf{0}$. Using the Newton's method, $\boldsymbol{\beta}$ can be estimated by an iterative process using the formula

$$\mathbf{b}^{(r)} = \mathbf{b}^{(r-1)} - [\mathbf{J}^{(r-1)}]^{-1} \mathbf{S}^{(r-1)}$$

and an initial guess $\mathbf{b}^{(0)}$ of $\boldsymbol{\beta}$. Here $\mathbf{b}^{(r)}$ is the estimation of $\boldsymbol{\beta}$ in the rth iteration, $\mathbf{S}^{(r-1)}$ is the evaluation of $\mathbf{S}$ in $\mathbf{b}^{(r-1)}$, and $\mathbf{J}$ is the Jacobian matrix of $\mathbf{S}$ evaluated at $\mathbf{b}^{(r-1)}$, that is, the matrix of partial derivatives of $\mathbf{S}$ with respect to $\boldsymbol{\beta}$ at $\mathbf{b}^{(r-1)}$

$$J_{jk} = \frac{\partial S_j}{\partial \beta_k} = \frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}$$

Since $\mathrm{E}(\partial^2 \ell / \partial \beta_k \partial \beta_j) = -\mathcal{I}_{jk}$ one can estimate $\mathbf{J}$ with its expectation $-\mathcal{I}$

$$\mathbf{b}^{(r)} = \mathbf{b}^{(r-1)} + [\mathcal{I}^{(r-1)}]^{-1} \mathbf{S}^{(r-1)} \tag{2.4}$$

which is called the *scoring method*. Multiplying last equation by $\mathcal{I}^{(r-1)}$ both sides, right hand side can be rewritten as

$$\sum_{i=1}^{n} \frac{x_{ij}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \left[\sum_{k=1}^{p} x_{ik} b_k^{(r-1)} + (y_i - \mathrm{E}(Y_i)) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)\right]$$

Then, (2.4) becomes

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(r)} = \mathbf{X}^T\mathbf{W}\mathbf{z} \tag{2.5}$$

where $\mathbf{z}$ is the $n \times 1$ vector which components are defined by

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(r-1)} + (y_i - \mathrm{E}(Y_i)) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)$$

Note that (2.5) has the same form of the estimate $\boldsymbol{\beta}$ in weighted least squares, except that (2.5) is solve by an iterative process because both $\mathbf{W}$ and $\mathbf{z}$ depend on $\mathbf{b}$. Once $\mathbf{b}$ is obtained, the inference part can be derived using Taylor expansion of $\mathbf{S}$ around $\mathbf{b}$

$$\mathbf{S}(\boldsymbol{\beta}) \approx \mathbf{S}(\boldsymbol{b}) + S(\boldsymbol{b})(\boldsymbol{\beta} - \boldsymbol{b}) = -S(\mathbf{b})(\mathbf{b} - \boldsymbol{\beta}) = \mathcal{I}(\mathbf{b})(\mathbf{b} - \boldsymbol{\beta})$$

since $\mathbf{S}(\mathbf{b}) = \mathbf{0}$ and $-S(\mathbf{b}) = \mathcal{I}(\mathbf{b})$. If $\mathcal{I}$ is invertible at $\mathbf{b}$ and regarded as constant $\mathrm{E}(\mathbf{b} - \boldsymbol{\beta}) \approx \mathbf{0}$ and

$$\mathrm{Var}(\mathbf{b}) = \mathrm{E}[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T] \approx \mathrm{E}[\mathcal{I}^{-1}(\mathbf{b})\mathbf{S}(\boldsymbol{\beta})\mathbf{S}^T(\boldsymbol{\beta})\mathcal{I}^{-1}(\mathbf{b})] = \mathrm{E}[\mathcal{I}(\mathbf{b})] = \mathcal{I}(\mathbf{b}),$$

since $\mathbf{S}(\boldsymbol{\beta})\mathbf{S}^T(\boldsymbol{\beta}) = \mathrm{Var}[\mathbf{S}(\boldsymbol{\beta})] = \mathcal{I}(\boldsymbol{\beta})$. Then, $(\mathbf{b} - \boldsymbol{\beta})^T\mathcal{I}(\mathbf{b})(\mathbf{b} - \boldsymbol{\beta}) \sim \chi^2_p$ approximately. The statistic $(\mathbf{b} - \boldsymbol{\beta})^T\mathcal{I}(\mathbf{b})(\mathbf{b} - \boldsymbol{\beta})$ is what is called the *Wald test*. A more rigorous proof can be given by following the concepts of consistency and asymptotic given in next sections and can be found in Appendix A.

2.4 Likelihood Ratio, Wald, and Score Tests

When a statement about the true parameter $\boldsymbol{\theta}$ is made, one wants to test whether the statement is either true or false. This is done through a hypothesis test. The standard form of a hypothesis test is $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs. $H_1 : \boldsymbol{\theta} \in \Theta_0^c$; H_0 is called the null hypothesis and H_1 the alternative hypothesis. When $\boldsymbol{\theta}_0$ consists of a single value or vector H_0 is said to be simple, in other case H_0 is said to be composite. A test statistic $W(X_1, \ldots, X_n)$ which depends on the sample $X_1, \ldots, X_n$ is used to decide about the feasibility of H_0. The set of values for which H_0 is rejected, i.e., considered as false, is called the rejection region R.

Three general methods are used for testing hypotheses: the Likelihood Ratio Test (LRT), the Wald test, and the Score test.

Consider testing the simple hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ so that $\Theta_0 = \{\boldsymbol{\theta}_0\}$ and $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. Let $X_1, \ldots, X_n$ be a random sample drawn from a population with distribution function $f(x, \boldsymbol{\theta})$. Let $\mathbf{X} = [X_1, \ldots, X_n]^T$, $L(\boldsymbol{\theta}, \mathbf{X})$ be the likelihood function of $\mathbf{X}$, $\ell(\boldsymbol{\theta}, \mathbf{X})$ be the log-likelihood function of $\mathbf{X}$, $S(\boldsymbol{\theta}) = \dot{\ell}(\boldsymbol{\theta}, \mathbf{X})$ be the score function, $\mathcal{I}(\boldsymbol{\theta})$ be the information matrix, $\hat{\boldsymbol{\theta}}$ be the unrestricted MLE of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}_0$ the restricted MLE of $\boldsymbol{\theta}$ under H_0, i.e., $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$.

The Likelihood Ratio Test consists of comparing the two likelihoods in both the null parameter space Θ_0 and the entire parameter space $\Theta = \Theta_0 \cup \Theta_0^c$

$$LRT = \frac{\sup_{\Theta_0} L(\boldsymbol{\theta}, \mathbf{X})}{\sup_{\Theta} L(\boldsymbol{\theta}, \mathbf{X})} = \frac{L(\boldsymbol{\theta}_0, \mathbf{X})}{L(\hat{\boldsymbol{\theta}}, \mathbf{X})}. \tag{2.6}$$

An alternative form for the LRT is

$$LRT = 2[\ell(\hat{\boldsymbol{\theta}}, \mathbf{X}) - \ell(\boldsymbol{\theta}_0, \mathbf{X})] \tag{2.7}$$

which can be obtained by applying logs to (2.6) and multiplying by -2. This modification makes sense once the asymptotic distribution of LRT wants to be derived. Then H_0 is rejected when the LRT value is smaller than a critical value c which depends on the significance level α. Under appropriate conditions, the log-Likelihood Ratio Test has a chi-square asymptotic null distribution with p degrees of freedom.

One of the main disadvantages of the LRT is that it does not consider the curvature of the log-likelihood function. If the log-likelihood is flat enough $\boldsymbol{\theta}_0$ could be far from $\hat{\boldsymbol{\theta}}$ but $\ell(\boldsymbol{\theta}_0)$ will be closed to $\ell(\hat{\boldsymbol{\theta}})$ then, LRT will fail to reject H_0. See left graph in Figure 2.2 for the one-dimensional case. Note that green curve, referred to as ℓ_A in the graph, has more curvature than red curve ℓ_B, moreover $\ell(\hat{\theta}) = \ell_A(\hat{\theta}) = \ell_B(\hat{\theta})$ but the vertical distance $\ell(\hat{\theta}) - \ell(\theta_0)$ is greater for curve A than for curve B.

The Wald test, on the other hand, uses both the square distance between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ and the curvature of $\ell(\boldsymbol{\theta})$ which can be estimated by the information matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$. This is

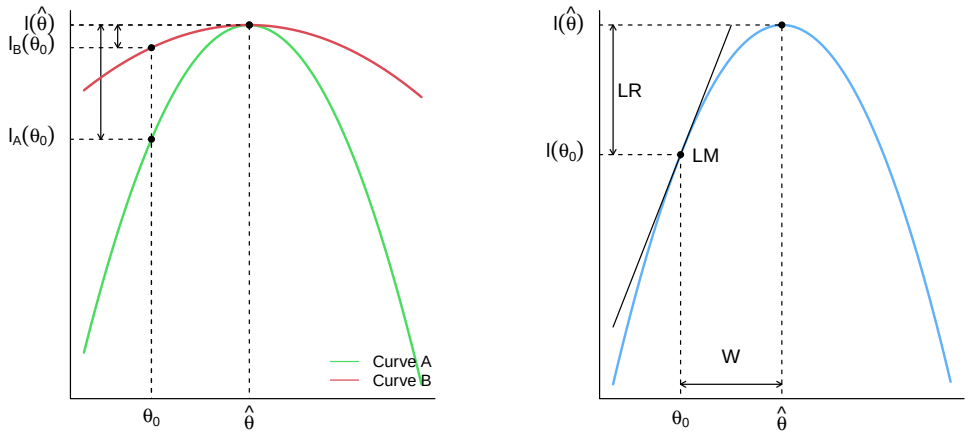

Figure 2.2: Graphical representation of the LR, Wald and score tests

defined by

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{2.8}$$

Note that the square distance is weighted by the curvature of ℓ. This is because a larger curvature will produce a greater departure of $\ell(\boldsymbol{\theta}_0)$ from $\ell(\hat{\boldsymbol{\theta}})$ as shown in left graph in Figure 2.2.

Just as before, $W \sim \chi^2_p$ asymptotically. When θ is one-dimensional, the Wald test takes a simpler form which is more familiar and common to see in literature:

$$W = \frac{\hat{\theta} - \theta_0}{\sigma(\hat{\theta})} \sim N(0, 1).$$

The Wald test will reject H_0 if $\boldsymbol{\theta}_0$ is farther enough to $\hat{\boldsymbol{\theta}}$, or equivalently, if W is larger than a critical value $c(\alpha)$. It will fail to reject if W is small.

The score test, also named as the Lagrange multiplier test (LM), considers both the curvature and the slope of the log-likelihood function at the null value $\boldsymbol{\theta}_0$. The curvature of ℓ is again estimated by the information matrix $\mathcal{I}$ and the slope is computed with the score statistic S which is, by definition, the partial derivative of ℓ with respect to $\boldsymbol{\theta}$, then

the Lagrange multiplier test becomes

$$LM = S^T(\boldsymbol{\theta}_0)\mathcal{I}^{-1}(\boldsymbol{\theta}_0)S(\boldsymbol{\theta}_0). \tag{2.9}$$

Recall that $S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ since $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ then if $\boldsymbol{\theta}_0$ is far from $\hat{\boldsymbol{\theta}}$, $S(\boldsymbol{\theta}_0)$ will be far from zero and then LM will reject H_0. Conversely, LM will fail to reject it. Similar to the Wald test, the square slope, $S^T(\boldsymbol{\theta}_0)S(\boldsymbol{\theta}_0)$, has been weighted but this time with the inverse of the curvature since the more the curvature, the more the departure of $S(\boldsymbol{\theta}_0)$ from $\mathbf{0}$, and therefore the bigger the square slope will be. In left plot of Figure 2.2, note that, for fixed θ_0, the square slop of $\ell_A(\theta)$ at θ_0 is greater than the square slop of $\ell_B(\theta)$. If curvature in LM was not taking into account, LM would reject H_0 for random samples with associated log-likelihood function ℓ_A but would fail to reject it if the associated log-likelihood function was ℓ_B. Again, LM $\sim \chi^2_p$ asymptotically, under suitable conditions.

Right graph in Figure 2.2 shows these three tests graphically when testing the simple null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Although Wald, LR, and LM tests are asymptotically equivalent, they possess different advantages and disadvantages. One of the most remarkable disadvantages of LRT is that two models must be fitted to compute it: the full model and the reduced one. In the Wald and LM tests only one model must be fitted, instead. Another advantage of Wald an LM tests is that one can easily construct confidence intervals which could be analytically difficult (or even impossible) for the LRT. An advantage over the Wald test is that LR and LM tests are invariant under monotone functions, which allows to compute (or estimate) confidence intervals not only for $\boldsymbol{\theta}$ but for $g(\boldsymbol{\theta})$, as well. One of the major advantages of LM test is that it does not require the computation of $\hat{\boldsymbol{\theta}}$, however it requires the computation of the inverse of the information matrix, which may not even exist. An interesting property when the three tests are applied to a linear model is that $W \geq LR \geq LM$ as shown in [4]. This relationship implies that, in testing the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, Wald test rejects the null hypothesis most often than the other two.

2.5 Profile Likelihood

In the maximum likelihood framework, one intents to estimate $\boldsymbol{\theta}$ by maximizing the log-likelihood function $\ell(\boldsymbol{\theta}, \mathbf{X})$ with respect to $\boldsymbol{\theta}$, that is,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}, \mathbf{X})$$

However, one could be sometimes interested in making inferences about a portion of a paramater $\boldsymbol{\theta}$. In this situation, one could consider $\boldsymbol{\theta}$ being composed of two parts: the parameter of interest $\boldsymbol{\psi}$ and the nuisance parameter $\boldsymbol{\eta}$. More specifically, suppose that the p-dimensional parameter $\boldsymbol{\theta}$ can be partitioned into two subparameters $(\boldsymbol{\psi}^T, \boldsymbol{\eta}^T)^T$ where $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ are of dimensions q and $p - q$, respectively. In this case, it could be unnecessary to estimate $\boldsymbol{\theta}$ all at ones as it must be done under the Maximum Likelihood framework. Instead, it could be better to estimate $\boldsymbol{\theta}$ in two stages. First, one could suppose that $\boldsymbol{\psi}$ is known and estimate $\boldsymbol{\eta}$ by maximizing $\ell(\boldsymbol{\psi}, \boldsymbol{\eta})$ over $\boldsymbol{\eta}$

$$\hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, \ell_{\boldsymbol{\psi}}(\boldsymbol{\eta}) = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\psi}, \boldsymbol{\eta}). \tag{2.10}$$

The subscript in $\hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}}$ is used to emphasize that the estimation is in terms of $\boldsymbol{\psi}$. Second, use $\hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}}$ to find the MLE of $\boldsymbol{\psi}$,

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}}\, \ell(\boldsymbol{\psi}, \hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}})$$

Note that $\ell(\boldsymbol{\psi}, \hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}})$ is only in terms of $\boldsymbol{\psi}$. Finally, $\hat{\boldsymbol{\eta}}$ is found by replacing $\hat{\boldsymbol{\psi}}$ in $\hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}}$, i.e. $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_{\hat{\boldsymbol{\psi}}}$.

To illustrate the idea discussed above consider the following example. Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables following a $\mathcal{N}(\mu, \sigma^2)$. One can be interested on estimating $\theta = (\mu, \sigma^2)$ but making inferences only about μ then σ^2 is regarded as the nuisance parameter. The log-likelihood function is then

$$\ell(\mu, \sigma^2; \boldsymbol{X}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

where $\mathbf{X} = [X_1, \ldots, X_n]^T$. Assume, for now, that μ is known and estimate σ^2 by maximizing $\ell(\mu, \sigma^2; \boldsymbol{X}) = \ell_\mu(\sigma^2; \boldsymbol{X})$ with respect to σ^2. Then

$$\hat{\sigma}^2_\mu = \underset{\sigma^2}{\operatorname{argmax}}\, \ell_\mu(\sigma^2; \boldsymbol{X}) = \underset{\sigma^2}{\operatorname{argmax}} \left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \right)$$

It is straightforward to find that $\hat{\sigma}^2_\mu = n^{-1}\sum_{i=1}^{n}(X_i - \mu)^2$. Now, $\hat{\sigma}^2_\mu$ can be used to estimate μ by replacing $\hat{\sigma}^2_\mu$ in $\ell(\mu, \sigma^2; \boldsymbol{X})$ and maximizing ℓ with respect to μ,

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}}\, \ell(\mu, \hat{\sigma}^2_\mu; \boldsymbol{X}) = \underset{\mu}{\operatorname{argmax}} \left[-\frac{n}{2}\log\left(\frac{2\pi}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \right) - \frac{n}{2} \right]$$

Then $\hat{\mu} = n^{-1}\sum_{i=1}^{n} X_i = \bar{X}$ and hence $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Note that $\hat{\mu}$ and $\hat{\sigma}^2_\mu$ are the usual MLE $(\hat{\mu}, \hat{\sigma}^2) = \operatorname{argmax}_{\mu,\sigma^2}\, \ell(\mu, \sigma^2; \boldsymbol{X})$. This is not an special case, in fact, it can be proved that the profile likelihood framework generates the usual MLEs.

In Figure 2.3, the two profile log-likelihood function for μ (left graph) and σ (right graph) have been plotted for a random sample $\mathbf{X} = [X_1, \ldots, X_n]^T$ where $X_i \sim \mathcal{N}(0, 1)$. Note that the maximum of $\ell(\mu, \hat{\sigma}_\mu)$ is reached at some value closed to zero and the maximum of $\ell(\hat{\mu}_\sigma, \sigma)$ is reached at some value closed to one which is expected since true μ and σ equal zero and one, respectively.

As it will be proved in next sections, the MLE of $\hat{\boldsymbol{\theta}}$ has an asymptotic $\mathcal{N}_p(\boldsymbol{\theta}, \mathcal{I}(\boldsymbol{\theta})^{-1})$ distribution under suitable conditions, where $\boldsymbol{\theta} = (\boldsymbol{\psi}^T, \boldsymbol{\eta}^T)^T$ is the true parameter vector. Then, if $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\psi}}^T, \hat{\boldsymbol{\eta}}^T)^T$

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix} \sim \mathcal{N}_p \left(\mathbf{0}, \begin{pmatrix} \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\psi}} & \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\eta}} \\ \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\psi}} & \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} \end{pmatrix}^{-1} \right)$$

asymptotically. Here $\mathcal{I}_{\boldsymbol{u}\boldsymbol{v}} = \mathcal{I}_{\boldsymbol{u}\boldsymbol{v}}(\boldsymbol{\psi}, \boldsymbol{\eta}) = -\mathrm{E}\left[\partial^2 \boldsymbol{\ell}(\boldsymbol{\psi}, \boldsymbol{\eta})/\partial \boldsymbol{u} \partial \boldsymbol{v}^T\right]$ for $\boldsymbol{u}, \boldsymbol{v} = \boldsymbol{\psi}, \boldsymbol{\eta}$. Notice that for an invertible block matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

its inverse is

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}CB(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

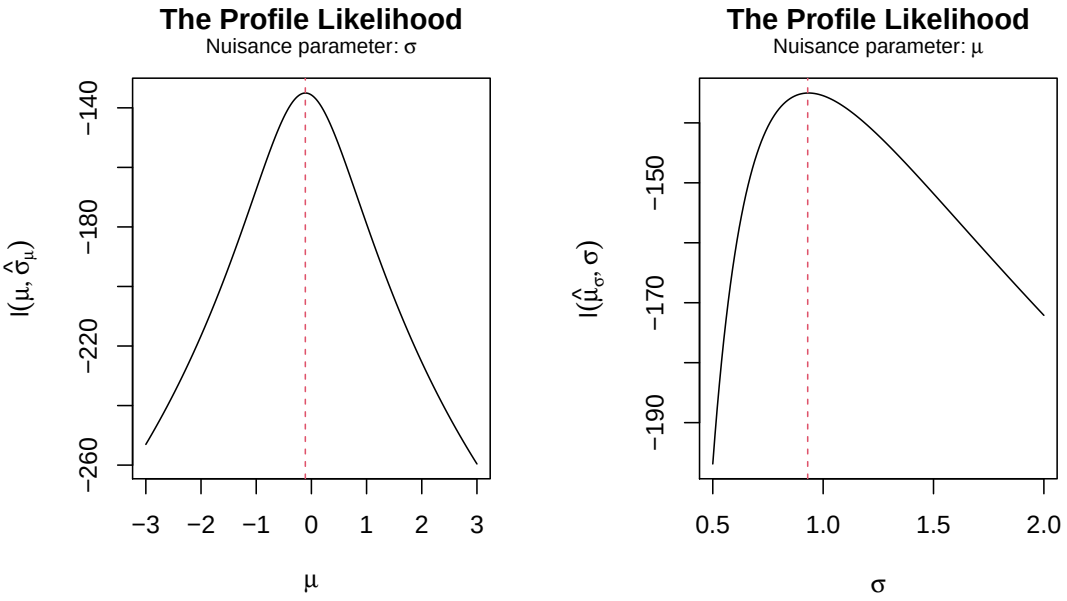

Figure 2.3: Profile log-likelihood functions for μ (left graph) and σ (right graph)

and, therefore

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \sim \mathcal{N}_q(\mathbf{0}, (\mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\psi}} - \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\psi}})^{-1})$$

asymptotically. Henceforth, the Wald, LR, and LM tests for testing $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ become

$$W = n(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T(\mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\psi}} - \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\psi}})(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0), \tag{2.11}$$

$$LRT = 2[\ell(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}) - \ell(\boldsymbol{\psi}_0, \hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}_0})], \tag{2.12}$$

$$LM = S^T(\boldsymbol{\psi}_0, \hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}_0})(\mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\psi}} - \mathcal{I}_{\boldsymbol{\psi}\boldsymbol{\eta}}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\psi}})^{-1}S(\boldsymbol{\psi}_0, \hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}_0}), \tag{2.13}$$

where the $\mathcal{I}_{\boldsymbol{u}\boldsymbol{v}}$'s are evaluated at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ and $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}_{\boldsymbol{\psi}_0}$. As expected, Wald, LRT, and LM have an asymptotic χ_q^2 distribution. Proof of the asymptotic distribution of LRT in profile likelihood can be found in [9]. In fact, this result is known as the *Wilk's Theorem*.

2.6 Stochastic Convergence

The results presented throughout all this section are valid for both a single parameter θ and a vector parameter $\boldsymbol{\theta}$ as well as for a random variable X or vector $\mathbf{X}$, therefore distinction will not be made unless otherwise stated. Also, the following notations will be used.

2.6.1 About the Notations

Let P be a measure in a measurable space $(\mathcal{X}, \mathcal{B})$ and $f : \mathcal{X} \to \mathbb{R}^k$ be a measurable function. For a random vector X the expectation of $f(X)$, $\mathrm{E}[f(X)]$, is denoted as Pf. The empirical measure of a random sample $X_1, \ldots, X_n$ is denoted and defined as

$$\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(X_i)$$

For a sequence fo random vectors X_n the symbol $\xrightarrow{d}$ is set to indicate convergence of the sequence in distribution; the symbol $\xrightarrow{p}$ to indicate convergence in probability; and $\xrightarrow{as}$ to indicate almost surely convergence. These concepts will be defied shortly.

When n is used as an index, it is assumed that n tends to infinity and when referring to the *asymptotic* behavior of a (sequence of) random vector, it is meant taking limits as n goes to infinity.

Since convergence is a key concept in, rate of convergence will be discuss as well. The symbol $o_p(1)$ means convergence in probability to zero and $O_p(1)$ means bounded in probability. The general notations $o_p(R_n)$ and $O_p(R_n)$ will be discussed in short.

2.6.2 Type of Convergence

Before defining type of convergence in probability, two main types of convergence in calculus will be defined: *pointwise* and *uniform convergence*. Let f_n be a sequence of functions defined all on the same domain as a function f. f_n is said to converge pointwise to f if $\lim_{n\to\infty} f_n(x) = f(x)$ for every x in the domain of f. This is denoted by $f_n \to f$. On the other

hand, f_n is said to converge uniformly to f if for every $\epsilon > 0$ and for every x in the domain of f there exists $n_0 \in \mathbb{N}$, such that for all $n > n_0$, $\sup_x |f_n(x) - f(x)| < \epsilon$.

For a sequence of random variables X_n, we say that X_n converges in distribution to a random vector X if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x at which F is continuous, where F_n and F are the cumulative distribution functions of X_n and X, respectively. This type of convergence is also called *weak convergence* or *convergence in law*.

A stronger type of convergence is what is called *convergence in probability*. A sequence of random vectors X_n converges in probability to X if for every $\epsilon > 0$, $\lim_{n\to\infty} \Pr(d(X_n, X) > \epsilon) = 0$, where d is any distance metric defined in the metric space (X, d). When X_n and X are random variables, $d(X_n, X)$ can be replace by $|X_n - X|$, then X_n converges in probability to X if $\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0$.

Finally, the sequence X_n is said to *converge almost surely* to X if $\Pr(\lim_{n\to\infty} d(X_n, X) = 0) = 1$. Again, $d(X_n, X)$ can be replaced by $|X_n - X|$ when X_n and X are random variables.

An obvious question regarding convergence of a sequence of random vectors X_n is about the convergence of a function of random vector. One could want to know whether convergence of a sequence implies or not convergence of function of that sequence. This is true if the function is continuous, along with other conditions, and is stated in the Continuous Mapping Theorem. Proof can be found in [8].

Theorem 1 (Continuous Mapping)**.** *Let X_n be a sequence of random vectors and X a random vector. Let $g : \mathbb{R}^P \to \mathbb{R}^q$ be a continuous function in a nonempty set $A \subseteq \mathbb{R}^p$ with the property that $\Pr(X \in A) = 1$. Then*

(i) if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$

(ii) if $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$

(iii) if $X_n \xrightarrow{as} X$, then $g(X_n) \xrightarrow{as} g(X)$

The following theorem states the relation between the three types of convergence discussed above. The order in which they were defined the stronger the convergence is. In fact,

almost sure convergence implies convergence in probability, convergence in probability implies convergence in distribution, being the last mentioned the weakest type of convergence among the three. Other relations are also stated. Proof can also be found in [8].

Theorem 2. *Let X_n and Y_n be two sequences of random vectors, X and Y two random vectors, and c a constant vector. Then,*

(i) if $X_n \xrightarrow{as} X$, then $X_n \xrightarrow{p} X$

(ii) if $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$

(iii) $X_n \xrightarrow{p} c$ if and only if $X_n \xrightarrow{d} c$

(iv) if $X_n \xrightarrow{d} X$ and $d(X_n, Y_n) \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$

(v) if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $(X_n, Y_n) \xrightarrow{d} (X, c)$

(vi) if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$

One important lemma is one called Slutsky's Lemma. This states the distribution of the sum and the product of two random sequences combining the Continuous Mapping Theorem and statements (iii) and (v) from Theorem 2. Proof of the this theorem is straightforward.

Theorem 3 (Slutsky's Lemma)**.** *Let X_n and Y_n two random sequences, X be a random vector and c a constant such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, then*

(i) $X_n + Y_n \xrightarrow{d} X + c$

(ii) $X_n Y_n \xrightarrow{d} cX$

(iii) $Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$, $c \neq 0$

Two last definitions are about boundness. A random vector X is said to be *tight* if for every $\epsilon > 0$ there exists a constant M such that $\Pr(||X|| > M) < \epsilon$. A sequence of random vectors X_α is said to be *uniformly tight* or *bounded in probability* if for every $\epsilon > 0$ there exists a constant M such that $\sup_\alpha \Pr(||X_\alpha|| > M) < \epsilon$.

2.6.3 Rate of Convergence

As mentioned before, a random vector X is $o_p(1)$ if X converges to zero in probability and a sequence of random vectors X_n is $O_p(1)$ if X_n is bounded in probability. If X_n and Y_n are sequences of random variables, $X_n = o_p(Y_n)$ if and only if $X_n/Y_n = o_p(1)$, i.e., if X_n/Y_n converges in probability to zero, and $X_n = O_p(Y_n)$ if and only if $X_n/Y_n = O_p(1)$, that is, if X_n/Y_n is bounded in probability. If X_n and Y_n are sequences of random vectors,

$$X_n = o_p(Y_n) \text{ if and only if } X_n = Z_n Y_n \text{ and } Z_n \xrightarrow{p} 0,$$
$$X_n = O_p(Y_n) \text{ if and only if } X_n = Z_n Y_n \text{ and } Z_n = O_p(1).$$

As a special property that can derived directly from the discussion above is that

$$o_p(Y_n) = Y_n o_p(1) \text{ and } O_p(Y_n) = Y_n O_p(1).$$

Moreover,

$$o_p(1) + o_p(1) = o_p(1),$$
$$o_p(1) + O_p(1) = O_p(1),$$
$$O_p(1) o_p(1) = o_p(1).$$

The first arithmetic property is a direct consequence of the Continuous Mapping Theorem and statement (vi) in Theorem 2. The other properties can be proved by writing $o_p(1)$ and $O_p(1)$ in terms of random vectors. More general properties can be found involving $o_p(a_n)$ and $o_p(b_n)$ for positive sequences a_n and b_n converging to zero.

2.7 M and Z-Estimators

When one wants to estimate a parameter θ of a population f it is usual to use a criterion function involving a random sample drawn from that population. Let $X_1, \ldots, X_n$ a random sample from a population with distribution function f. Consider maximizing the criterion

function

$$M_n(\theta) = n^{-1} \sum_{i=1}^{n} m(X_i, \theta) = \mathbb{P}_n m(\theta) \tag{2.14}$$

where $m : \mathcal{X} \to \mathbb{R}$ is a known function which depends on the true parameter θ. Usually, when maximizing some function, one appeals to solving equations of the type

$$\Psi_n(\theta) = n^{-1} \sum_{i=1}^{n} \psi(X_i, \theta) = \mathbb{P}_n \psi(\theta) = 0 \tag{2.15}$$

where ψ is a known vector-valued function. When θ is p-dimensional, last equation can be transformed into the system of equations

$$\sum_{i=1}^{n} \psi_j(X_i, \theta) = 0, \quad j = 1, \ldots, p. \tag{2.16}$$

In this case, ψ can be thought as a vector of vector-valued functions, i.e. $\psi = (\psi_1, ..., \psi_p)$. Equations such as the ones defined in (2.16) are called *estimating equations*. Finding vector $\hat{\theta}_n$ which maximizes $M_n(\theta)$ in (2.14) yields what it is called an *M-estimator*. Similarly, finding $\hat{\theta}_n$ which solves (2.15) yields what it is known as a *Z-estimator*.

Some examples of *M-Z* estimators are the sample mean and sample median. Note that $\bar{X} = \text{argmax}_\theta \, n^{-1} \sum_{i=1}^{n} [-(X_i - \theta)^2]$. Here $m(X, \theta) = -(X - \theta)^2$. Also $\bar{X}$ can be thought as the solution of $n^{-1} \sum_{i=1}^{n} (X_i - \theta) = 0$. Other examples are the least square estimators and maximum likelihood estimators. For a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

the least square estimates are $(b_0, b_1) = \text{argmax}_{\beta_0, \beta_1} \sum_{i=1}^{n} -(Y_i - \beta_0 - \beta_1 X_i)^2$. Least squares estimators can also be found by solving an equation of the form shown in (2.15). Recall that, for a random sample $Y = (Y_1, ..., Y_n)$, the MLE of β is defined by

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \, \mathcal{L}(\beta, Y) = \underset{\beta}{\text{argmax}} \, \ell(\beta, Y) = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \ell(\beta, Y_i)$$

Then, MLE is also an M-estimator. Moreover, $\hat{\beta}$ can be thought as the solution of the equation

$$\sum_{i=1}^{n} \frac{\partial \ell}{\partial \beta}(\beta, Y_i) = \sum_{i=1}^{n} S(\beta, Y_i) = 0$$

where S is the score function. If β is a p-dimensional parameter vector, $S = (S_1, ..., S_p)$ where $S_j = \partial\ell/\partial\beta_j$ and the estimating equations become

$$\sum_{i=1}^{n} S_j(\beta, Y_i) = 0 \quad \text{for} \quad j = 1, \ldots, p.$$

Henceforth, MLE can also be thought as a Z-estimator.

2.7.1 Consistency of M and Z-Estimators

One necessary property to establish asymptotic normality in M and Z-estimators is *consistency*. An estimator $\hat{\theta}_n$ is said to be *consistent* to θ if $\hat{\theta}_n$ converges in probability to θ. If $\hat{\theta}_n$ is the maximizer of $M_n(\theta) = \mathbb{P}_n m(\theta)$ and θ_0 the maximizer of $M(\theta) = Pm(\theta)$ by the Law of Large Numbers (LLN) $M_n(\theta) \xrightarrow{p} M(\theta)$ for all θ. An obvious question is whether this convergence of the criterion function ensures the converges of the maximizer $\hat{\theta}_n$ to θ_0. Unfortunately, the answer is no. Convergence in probability is too weak and therefore a stronger type of convergence is needed: the *uniform convergence*

$$\lim_{n\to\infty} \sup_{\theta} \Pr(|M_n(\theta) - M(\theta)| \geq \epsilon) = 0, \quad \text{for all} \quad \epsilon > 0$$

However, uniform convergence is too strong and difficult to verify in some situations. A way to relax this condition is to allow $\hat{\theta}$ being a nearly maximizer of $M_n(\theta)$. Moreover, it is also needed to ensure that θ_0 uniquely maximizes $M(\theta)$. The following theorem states the sufficient conditions to ensure consistency.

Theorem 4 (Consistency)**.** *Let $M_n(\theta) = \mathbb{P}_n m$ and $M(\theta) = Pm$. Suppose that*

(i) $\sup_\theta |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$,

(ii) $\sup_\theta\{M(\theta) : d(\theta, \theta_0) \geq \epsilon\} \leq M(\theta_0)$ *for every* $\epsilon \geq 0$,

(iii) $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$

Then, $\hat{\theta}$ is a consistent estimator of θ_0.

Proof. First condition implies that $M_n(\theta)$ converges uniformly to $M(\theta)$ and therefore $M_n(\theta) \xrightarrow{p} M(\theta)$. Consequently $M_n(\theta_0) \xrightarrow{p} M(\theta_0)$. By third condition

$$\begin{aligned} M_n(\hat{\theta}_n) &\geq M(\theta_0) - o_p(1) \\ M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_p(1) \\ &\leq \sup_\theta |M_n(\theta) - M(\theta)| + o_p(1) \end{aligned}$$

Notice that the first condition also can be seen as $\sup_\theta |M_n(\theta) - M(\theta)| = o_p(1)$, then $M(\theta_0) - M(\hat{\theta}_n) \leq o_p(1) + o_p(1) = o_p(1)$. In other words, $M(\theta_0) - M(\hat{\theta}_n)$ converges in probability to zero.

By the second condition an the definition of supremum there exists $\eta > 0$ such that for every $\epsilon > 0$, $M(\theta) < M(\theta_0) - \eta$ for every θ such that $d(\theta, \theta_0) \geq \epsilon$. Therefore, it is clear that

$$\begin{aligned} \{d(\hat{\theta}_n, \theta_0) \geq \epsilon\} &\subseteq \{M(\hat{\theta}_n) < M(\theta_0) - \eta)\} \\ \Pr\{d(\hat{\theta}_n, \theta_0) \geq \epsilon\} &\leq \Pr\{M(\hat{\theta}_n) < M(\theta_0) - \eta)\} \to 0 \end{aligned}$$

Therefore, $\hat{\theta}_n$ converges in probability to θ_0, i.e., $\hat{\theta}_n$ is a consistent estimator of θ_0. □

The same result can be obtained for Z-estimators by letting $M_n(\theta) = -||\Psi_n(\theta)||$, $M(\theta) = -||\Psi(\theta)||$, $\hat{\theta}_n$ a be nearly zero of $\Psi_n(\theta)$, and θ_0 be a zero of $\Psi(\theta)$ in the preceding theorem.

2.7.2 Asymptotic Normality of Z-Estimators

Although M–estimators have been the main focus of this section, it can be proved easier that Z–estimators are asymptotically normal under suitable conditions than M–estimators. Also, recall that some M–estimators are also Z–estimators, specially the maximum likelihood estimations, which are the basis of the proposed test stated in the next section.

Consider the case where θ is a real number and $\hat{\theta}_n$ is a Z – estimator of θ. The following theorem shows that if θ_0 is a zero of $\Psi(\theta)$ and that $\hat{\theta}_n$ is consistent to θ_0 then $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal. Other conditions are also stated in the theorem.

Theorem 5 (Asymptotic Normality)**.** *Let* $\Psi_n(\theta) = \mathbb{P}_n\psi(\theta)$ *and* $\Psi(\theta) = P\psi(\theta)$*. Suppose that* $\hat{\theta}_n$ *is a zero of* Ψ_n*,* θ_0 *is a zero of* Ψ*, and* $\hat{\theta}_n \xrightarrow{p} \theta_0$*. If* $P\psi^2 < \infty$*,* $P\dot{\psi} < \infty$*, and* $\ddot{\Psi}_n(\tilde{\theta}_n)$ *is bounded in probability for some* $\tilde{\theta}_n$ *between* $\hat{\theta}_n$ *and* θ_0*, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{P\psi^2(\theta_0)}{(P\dot{\psi}(\theta_0))^2}\right)$$

Proof. Assuming that $\hat{\theta}_n$ is closed enough to θ_0 and that $\psi(\theta)$ is twice differentiable at θ_0, there must exists a $\tilde{\theta}_n$ between $\hat{\theta}_n$ and θ_0 such that

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^2 \tag{2.17}$$

which corresponds to the Taylor series of $\Psi_n(\hat{\theta}_n)$ around θ_0. Equation (2.17) can be rearranged so that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)} \tag{2.18}$$

Note that, since $\Psi_n(\theta_0)$ is an average and assuming that $P\Psi(\theta_0)^2$ is finite, by the Central Limit Theorem, $\Psi_n(\theta_0)$ it is asymptotically normal. Moreover, the expectation of $-\sqrt{n}\Psi_n(\theta_0)$ is zero since

$$P\Psi_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} P\psi(X_i, \theta_0) = P\psi(\theta_0) = \Psi(\theta_0) = 0$$

Therefore, the numerator of (2.18) converges in distribution to $\mathcal{N}(0, P\psi^2(\theta_0))$. Concerning the first term of the denominator in (2.18), note that $\dot{\Psi}_n(\theta_0)$ converges in probability to the constant $P\dot{\psi}(\theta_o)$ by the law of large numbers. By assumption $\ddot{\Psi}_n(\tilde{\theta}_n) = O_p(1)$ and $\hat{\theta}_n - \theta_0 = o_p(1)$, therefore $\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) = O_p(1)o_p(1) = o_p(1)$ and

$$\dot{\Psi}_n(\theta_0) + \frac{1}{2}\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) \xrightarrow{d} P\dot{\psi}(\theta_0)$$

Then,

$$\frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)} \xrightarrow{d} \mathcal{N}\left(0, \frac{P\psi^2(\theta_0)}{(P\dot{\psi}(\theta_0))^2}\right)$$

by Slutsky's lemma. □

Theorem 5 and its proof can be extended for a p – dimensional parameter θ_0, so that if $\hat{\theta}_n$ is a consistent estimator of θ_0, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}_p\big(0, (P\dot{\psi}(\theta_0))^{-1} P\psi(\theta_0)\psi^T(\theta_0)(P\dot{\psi}(\theta_0))^{-1}\big)$$

In this case, $\Psi_n : \mathbb{R}^p \to \mathbb{R}^p$, $\dot{\Psi}_n(\theta_0)$ are $p \times p$ matrices that converges to the $p \times p$ matrix $\dot{\Psi}(\theta_0)$ whose entries are the expectation of the partial derivatives of $\psi_i(\theta_0)$ with respect to θ_j. The invertibility of $P\dot{\psi}(\theta_0)$ is also required.

2.7.3 Asymptotic Normality of MLEs

Recall that, for a sequence of random vectors $X_1, \ldots, X_n$ with common distribution function $f(\theta)$, the maximum likelihood estimator of a parameter θ is defined as

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(X_i, \theta) \tag{2.19}$$

That is, $\hat{\theta}_n$ is an M-estimator. To show consistency of maximum likelihood estimators it is sometimes convenient to write $M_n(\theta)$ is the form

$$M_n(\theta) = n^{-1} \sum_{i=1}^{n} \log \frac{f(X_i, \theta)}{f(X_i, \theta_0)} = \mathbb{P}_n \log \frac{f(\theta)}{f(\theta_0)} \tag{2.20}$$

Note that, $\hat{\theta}_n$ satisfying (2.19) also maximizes (2.20) since $M_n(\theta) = n^{-1} \sum_i \log f(X_i, \theta) - c$ where $c = n^{-1} \sum_i \log f(X_i, \theta_0)$ which is constant with respect to θ. $M(\theta)$ is then

$$M(\theta) = P \log \frac{f(\theta)}{f(\theta_0)}.$$

Since, by the Weak Law of Large Numbers (WLLN), $M_n(\theta) \xrightarrow{p} M(\theta)$, it is expected that the MLE $\hat{\theta}_n$ to converge in probability to the maximizer of $M(\theta)$. One condition for the true parameter θ_0 to be a maximizer of $M(\theta)$ is for $f(\theta)$ to be *identifiable*, this means that for all $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$, $f(\theta_1) \neq f(\theta_2)$. Along with the conditions in Theorem 6, the MLE, $\hat{\theta}_n$, is a consistent estimator of θ_0.

Moreover, MLEs are also the solution of the equation

$$\Psi_n(\theta) = \sum_{i=1}^{n} \psi(X_i, \theta) = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta}(X_i, \theta) = 0$$

or, solution of the system of equations

$$\sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta_j}(X_i, \theta) = 0 \quad \text{for} \quad j = 1, ..., p$$

if θ is a p-dimensional parameter vector. Hence MLEs are also Z-estimators provided $\dot{\ell}(\theta) = \partial \ell(X_i, \theta)/\partial \theta$ exists. By the WLLN $\Psi_n(\theta) \xrightarrow{p} \Psi(\theta)$ where $\Psi(\theta) = P\dot{\ell}(\theta)$ provided the expectation and variance-covariance matrix of $\dot{\ell}(\theta)$ exist. Then, under other conditions stated in Theorem 5, it is expected that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to

$$\mathcal{N}_p\big(0, (P\dot{\psi}(\theta_0))^{-1} P\psi(\theta_0)\psi^T(\theta_0)(P\dot{\psi}(\theta_0))^{-1})\big).$$

Recall that, for the information matrix $\mathcal{I}(\theta_0)$, $\mathcal{I}(\theta_0) = P\psi(\theta_0)\psi^T(\theta_0)$ and $\mathcal{I}(\theta_0) = -P\dot{\psi}(\theta_0)$, and therefore $(P\dot{\psi}(\theta_0))^{-1} P\psi(\theta_0)\psi^T(\theta_0)(P\dot{\psi}(\theta_0))^{-1})$ reduces to $\mathcal{I}^{-1}(\theta_0)$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}_p(0, \mathcal{I}^{-1}(\theta_0)).$$

Chapter 3

Proposed Methods

A method for hypothesis testing based on Maximum Likelihood and Profile Likelihood frameworks is proposed. The main idea is to split a p-dimensional parameter into two parts and test the null hypothesis that one of them is the zero vector, or in the more general case, is a constant vector. The statistic proposed for this test is found by computing the MLE of one part under H_0 and use it to compute the MLE of the other part. The exact procedure is described in the next section.

3.1 The Efficient Wald Test

Suppose a true parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ is of interest. Consider partitioning $\boldsymbol{\theta}$ in two parts $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ where $\boldsymbol{\theta}_1 \in \mathbb{R}^q$ and $\boldsymbol{\theta}_2 \in \mathbb{R}^{p-q}$. Let $H_0 : \boldsymbol{\theta}_2 = \mathbf{b}$ the null hypothesis to be tested. Let $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ the log-likelihood function. Three traditional tests can be implemented within the maximum likelihood framework: the likelihood ratio test, the Wald's test, and the score test. Modification of the first two will be explored. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \text{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^p} \ell(\boldsymbol{\theta})$ the unrestricted MLE of $\boldsymbol{\theta}$ and

$$\tilde{\boldsymbol{\theta}}_1 = \underset{\boldsymbol{\theta}_1 \in \mathbb{R}^q}{\text{argmax}}\, \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{b}), \tag{3.1}$$

the restricted MLE of $\boldsymbol{\theta}_1$ under H_0. The LRT statistic is given by

$$LRT = 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}}_1, \mathbf{b})\right). \tag{3.2}$$

As shown in the [9], $LRT \xrightarrow{d} \chi^2_{p-q}$ under H_0.

In the proposed method, $\tilde{\boldsymbol{\theta}}_1$ is computed as before and used to compute $\tilde{\boldsymbol{\theta}}_2$ as

$$\tilde{\boldsymbol{\theta}}_2 = \underset{\boldsymbol{\theta}_2 \in \mathbb{R}^{p-q}}{\operatorname{argmax}} \, \ell(\boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2). \tag{3.3}$$

Note that this is the maximizer of the profile likelihood for $\boldsymbol{\theta}_2$ with the only difference that the maximizer of $\boldsymbol{\theta}_1$ is obtained computationally rather than analytically. In the following sections, the asymptotic distribution of $\tilde{\boldsymbol{\theta}}_2$ will be derived. In addition, the modified LR test,

$$LRT^* = 2\left(\ell(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) - \ell(\tilde{\boldsymbol{\theta}}_1, \mathbf{b})\right) \tag{3.4}$$

will be studied in simulation. Another modification of the LRT could be considered

$$LRT^{**} = 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)\right). \tag{3.5}$$

For the Wald test, consider partitioning $\mathcal{I}(\boldsymbol{\theta})$ as

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix} \tag{3.6}$$

where $\mathcal{I}_{ij}$ is the submatrix of $\mathcal{I}$ defined by

$$\mathcal{I}_{ij} = \mathcal{I}_{ij}(\boldsymbol{\theta}) = -\mathrm{E}\left(\frac{\partial^2 \ell(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j^T}\right), \quad \text{for} \quad i, j = 1, 2.$$

Then, the Wald test is defined as

$$W = (\hat{\boldsymbol{\theta}}_2 - \mathbf{b})^T (\mathcal{I}_{\mathbf{22}} - \mathcal{I}_{\mathbf{21}}\mathcal{I}_{\mathbf{11}}^{-1}\mathcal{I}_{\mathbf{12}})(\hat{\boldsymbol{\theta}}_2 - \mathbf{b}) \tag{3.7}$$

which asymptotic distribution is a χ^2_{p-q} as shown in Appendix A. The Efficient Wald test is finally defined as

$$W^* = (\tilde{\boldsymbol{\theta}}_2 - \mathbf{b})^T (\mathcal{I}_{\mathbf{22}}^{-1} - \mathcal{I}_{\mathbf{22}}^{-1}\mathcal{I}_{\mathbf{21}}\mathcal{I}_{\mathbf{11}}^{-1}\mathcal{I}_{\mathbf{12}}\mathcal{I}_{\mathbf{22}}^{-1})^{-1} (\tilde{\boldsymbol{\theta}}_2 - \mathbf{b}). \tag{3.8}$$

3.2 Asymptotic Properties of $\tilde{\boldsymbol{\theta}}_2$

In previous sections has been shown that the MLE of $\boldsymbol{\theta}$ is a consistent estimator of $\boldsymbol{\theta}$, under some mild conditions. Since $\tilde{\boldsymbol{\theta}}_1$ is the MLE of $\boldsymbol{\theta}_1$, $\tilde{\boldsymbol{\theta}}_1 \xrightarrow{p} \boldsymbol{\theta}_1$. It is enough to prove that if $\tilde{\boldsymbol{\theta}}_2$ is a consistent estimator of $\boldsymbol{\theta}_2 = \mathbf{b}$, then, by Theorem 2, $(\tilde{\boldsymbol{\theta}}_1^T, \tilde{\boldsymbol{\theta}}_2^T)^T \xrightarrow{p} (\boldsymbol{\theta}_1^T, \mathbf{b}^T)^T$. Next result shows that $\tilde{\boldsymbol{\theta}}_2 \xrightarrow{p} \mathbf{b}$ under the same conditions stated in Theorem 6.

3.2.1 Consistency of $\tilde{\boldsymbol{\theta}}_2$

To set up, consider the framework of M-estimation. Let $M_n(\boldsymbol{\theta}) = \mathbb{P}_n m(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta}) = Pm(\boldsymbol{\theta})$, where $m(\boldsymbol{\theta}, \cdot)$ is a function involved in the objective function to optimize and $\mathbb{P}_n$ and P are usual notations in empirical processes. Without loss of generality, it is of interest to test $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$. Thus $\tilde{\boldsymbol{\theta}}_1 = \text{argmax}_{\boldsymbol{\theta}_1} M_n(\boldsymbol{\theta}_1, \mathbf{0})$ and $\tilde{\boldsymbol{\theta}}_2 = \text{argmax}_{\boldsymbol{\theta}_2} M_n(\tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$. Let $\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta} \in \Theta} M(\boldsymbol{\theta})$ denote the true parameter, which is a global maximizer of $M(\cdot)$. Under the null, $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^{*T}, \mathbf{0}^T)^T$.

Theorem 6. *Suppose that*

(i) $\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow{p} 0$;

(ii) $\sup_{\boldsymbol{\theta}} \{M(\boldsymbol{\theta}) : d(\boldsymbol{\theta}, \boldsymbol{\theta}^*) > \delta\} < M(\boldsymbol{\theta}^*)$ *for all* $\delta > 0$.

Then $\tilde{\boldsymbol{\theta}}_2 \xrightarrow{p} \mathbf{0}$ *under the null* H_0.

Proof. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1^T, \tilde{\boldsymbol{\theta}}_2^T)^T$. It suffices to show $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$. By condition (ii), for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\begin{aligned}
\Pr\left(d(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq \epsilon\right) &\leq \Pr\left(M(\boldsymbol{\theta}^*) - M(\tilde{\boldsymbol{\theta}}) \geq \delta\right) \\
&= \Pr\left(M(\boldsymbol{\theta}^*) - M_n(\boldsymbol{\theta}^*) + M_n(\boldsymbol{\theta}^*) - M_n(\tilde{\boldsymbol{\theta}}) + M_n(\tilde{\boldsymbol{\theta}}) - M(\tilde{\boldsymbol{\theta}}) \geq \delta\right) \\
&\leq \Pr\left(M(\boldsymbol{\theta}^*) - M_n(\boldsymbol{\theta}^*) \geq \delta/3\right) + \Pr\left(M_n(\boldsymbol{\theta}^*) - M_n(\tilde{\boldsymbol{\theta}}) \geq \delta/3\right) + \\
&\quad \Pr\left(M_n(\tilde{\boldsymbol{\theta}}) - M(\tilde{\boldsymbol{\theta}}) \geq \delta/3\right).
\end{aligned}$$

Condition (i) implies that the first and third terms go to zero while the second probability also goes to zero since $\tilde{\boldsymbol{\theta}}$ is an approximate maximizer of $M_n(\cdot)$ in the sense that

$$M_n(\tilde{\boldsymbol{\theta}}) = M_n(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) \geq M_n(\tilde{\boldsymbol{\theta}}_1, \mathbf{0}) \geq M_n(\boldsymbol{\theta}_1^*, \mathbf{0}) \geq M_n(\boldsymbol{\theta}^*) - o_p(1).$$

□

Condition (i) is essentially a ULLN (Uniform Law of Large Numbers) while condition (ii) is an identifiability condition where approximately maximizing $M(\cdot)$ can unambiguously specify the true parameter $\boldsymbol{\theta}^*$.

3.2.2 Asymptotic Normality of $\tilde{\boldsymbol{\theta}}_2$

Consider the estimator $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1^T, \tilde{\boldsymbol{\theta}}_2^T)^T$ in the GLM setting within the likelihood framework. The following technical conditions are assumed. These are essentially regularity conditions that guarantee asymptotic normality of the ordinary MLE in GLM.

(i) The observations $(\boldsymbol{x}_i, y_i)$ are independent and identically distributed with probability density $f(\boldsymbol{x}, y, \boldsymbol{\theta})$ with respect to some measure μ. The function $f(\boldsymbol{x}, y, \boldsymbol{\theta})$ has a common support and the model is identifiable.

(ii) The first and second logarithmic derivatives of f satisfying the equations

$$E_{\boldsymbol{\theta}}\left[\frac{\partial \log f(\boldsymbol{x}, y, \boldsymbol{\theta})}{\partial \theta_j}\right] = 0 \tag{3.9}$$

and

$$E_{\boldsymbol{\theta}}\left[\frac{\partial \log f(\boldsymbol{x}, y, \boldsymbol{\theta})}{\partial \beta_j}\,\frac{\partial \log f(\boldsymbol{x}, y, \boldsymbol{\theta})}{\partial \beta_k}\right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f(\boldsymbol{x}, y, \boldsymbol{\theta})}{\partial \beta_j \partial \beta_k}\right] \tag{3.10}$$

for $j, k = 1, \ldots, p$. The expected Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \log f(\boldsymbol{x}, y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f(\boldsymbol{x}, y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right]$$

is finite and positive definite at the true $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$.

(iii) There exists an open subset or neighborhood $\mathcal{N}(\boldsymbol{\theta}^\star)$ of Ω that contains the true parameter point $\boldsymbol{\theta}^\star$ such that, for almost all $(\boldsymbol{x}, y)$, the density $f(\boldsymbol{x}, y, \boldsymbol{\theta})$ admits all third derivatives $\partial f(\boldsymbol{x}, y, \boldsymbol{\theta})/\partial\theta_j\partial\theta_k\partial\theta_l$ for all $\boldsymbol{\theta} \in \Omega$. Furthermore, there exist functions $M_{jkl}(\boldsymbol{x}, y)$ such that

$$\left|\frac{\partial^3 \log f(\boldsymbol{x}, y, \boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_l}\right| \leq M_{jkl}(\boldsymbol{x}, y),$$

for all $\boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}^\star)$, where $E_{\boldsymbol{\theta}^\star}\left[M_{jkl}(\boldsymbol{x}, y)\right] < \infty$ for $j, k, l = 1, \ldots, p$.

Let $\boldsymbol{\theta}$ being partition as $(\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. Let $\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{x}_i, y_i; \boldsymbol{\theta})$ be the likelihood function, $\ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta})$ be the log-likelihood function, and $H_0 : \boldsymbol{\theta}_2 = \mathbf{b}$. Let $\dot{\boldsymbol{\ell}}_1$ and $\dot{\boldsymbol{\ell}}_2$ be

the partial derivatives of $\boldsymbol{\ell}$ with respect to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. Consider the case where $\boldsymbol{\theta}$ is 2-dimensional, i.e. $p = 2$ and therefore $q = 1$ and $p - q = 1$, to avoid third derivative arrays. In such case, $\boldsymbol{\theta}_1 = \theta_1$ and $\boldsymbol{\theta}_2 = \theta_2$. Recall that $\tilde{\theta}_1$ was defined as the maximizer of $\ell(\theta_1, b)$ with respect to θ_1 by fixing $\theta_2 = b$ under H_0. Hence, it must satisfy $\dot{\ell}_1(\tilde{\theta}_1, b) = 0$ and by Taylor expansion of $\dot{\ell}(\tilde{\theta}_1, b)$ at θ_1

$$0 = \dot{\ell}_1(\tilde{\theta}_1, b) = \dot{\ell}_1(\theta_1, b) + \ddot{\ell}_{11}(\theta_1, b)(\tilde{\theta}_1 - \theta_1) + \frac{1}{2}\dddot{\ell}_{111}(\theta_1', b)(\tilde{\theta}_1 - \theta_1)^2$$

for some number θ_1' between θ_1 and $\tilde{\theta}_1$. Rearranging last equality yields to

$$\tilde{\theta}_1 - \theta_1 = -\frac{\dot{\ell}_1(\theta_1, b)}{\ddot{\ell}_{11}(\theta_1, b)} - \frac{1}{2}\frac{\dddot{\ell}_{111}(\theta_1', b)}{\ddot{\ell}_{11}(\theta_1, b)}(\tilde{\theta}_1 - \theta_1)^2. \tag{3.11}$$

Under regularity conditions, $\sqrt{n}\dot{\ell}_1(\theta_1, b) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{11})$. Moreover $-\ddot{\ell}_{11}(\theta_1, b) \xrightarrow{p} \mathcal{I}_{11}$. The second term in (3.11) is $o_p(|\tilde{\theta}_1 - \theta_1|)$ by assumption. Therefore, by Slutsky's theorem

$$\tilde{\theta}_1 \xrightarrow{d} \mathcal{N}(\theta_1, \mathcal{I}_{11}^{-1})$$

where $\mathcal{I}_{11}^{-1} = -P\ddot{\ell}_{11}$ evaluated at (θ_1, b).

Now, consider $\tilde{\theta}_2$. Recall that $\tilde{\theta}_2 = \text{argmax}_{\theta_2}\, \ell(\tilde{\theta}_1, \theta_2)$ which must satisfy $0 = \dot{\ell}_2(\tilde{\theta}_1, \tilde{\theta}_2)$. Applying Taylor expansion of $\ell(\tilde{\theta}_1, \tilde{\theta}_2)$ at $\theta_2 = b$ by fixing $\tilde{\theta}_1$ yields to

$$0 = \dot{\ell}_2(\tilde{\theta}_1, \tilde{\theta}_2) = \dot{\ell}_2(\tilde{\theta}_1, b) + \ddot{\ell}_{22}(\tilde{\theta}_1, b)(\tilde{\theta}_2 - b) + \frac{1}{2}\dddot{\ell}_{222}(\tilde{\theta}_1, \theta_2')(\tilde{\theta}_2 - b)^2$$

for some θ_2' between $\tilde{\theta}_2$ and b . Now, expanding $\dot{\ell}_2(\tilde{\theta}_1, b)$ at θ_1 and plugging it at previous equation one gets

$$\begin{aligned} 0 = \dot{\ell}_2(\theta_1, b) + \ddot{\ell}_{21}(\theta_1, b)(\tilde{\theta}_1 - \theta_1) + \frac{1}{2}\dddot{\ell}_{211}(\theta_1'', b)(\tilde{\theta}_1 - \theta_1)^2 \\ + \ddot{\ell}_{22}(\tilde{\theta}_1, b)(\tilde{\theta}_2 - b) + \frac{1}{2}\dddot{\ell}_{222}(\tilde{\theta}_1, \theta_2')(\tilde{\theta}_2 - b)^2 \end{aligned}$$

for some θ_1'' between $\tilde{\theta}_1$ and θ_1. Rearranging terms, it follows that

$$\begin{aligned} \tilde{\theta}_2 - b = -\frac{\dot{\ell}_2(\theta_1, b)}{\ddot{\ell}_{22}(\tilde{\theta}_1, b)} - \frac{\ddot{\ell}_{21}(\theta_1, b)}{\ddot{\ell}_{22}(\tilde{\theta}_1, b)}(\tilde{\theta}_1 - \theta_1) \\ - \frac{\dddot{\ell}_{211}(\theta_1'', b)}{2\ddot{\ell}_{22}(\tilde{\theta}_1, b)}(\tilde{\theta}_1 - \theta_1)^2 - \frac{\dddot{\ell}_{222}(\tilde{\theta}_1, \theta_2')}{2\ddot{\ell}_{22}(\tilde{\theta}_1, b)}(\tilde{\theta}_2 - \theta_2)^2 \end{aligned} \tag{3.12}$$

Notice that last two terms are $o_p(||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||)$ and hence can be ignored. However, the first two terms are correlated. Bringing (3.11) into (3.12) leads to

$$\begin{aligned}\tilde{\theta}_2 - b &= -\frac{\dot{\ell}_2(\theta_1, b)}{\ddot{\ell}_{22}(\tilde{\theta}_1, b)} + \frac{\ddot{\ell}_{21}(\theta_1, b)}{\ddot{\ell}_{22}(\tilde{\theta}_1, b)} \frac{\dot{\ell}_1(\theta_1, b)}{\ddot{\ell}_{11}(\theta_1, b)} + o_p||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}|| \\ &= \begin{bmatrix} \ddot{\ell}_{22}^{-1}(\tilde{\theta}_1, b)\ddot{\ell}_{21}(\theta_1, b)\ddot{\ell}_{11}^{-1}(\theta_1, b) & -\ddot{\ell}_{22}^{-1}(\tilde{\theta}_1, b) \end{bmatrix} \begin{bmatrix} \dot{\ell}_1(\theta_1, b) \\ \dot{\ell}_2(\theta_1, b) \end{bmatrix} + o_p||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||\end{aligned}$$

Under regularity conditions $\dot{\boldsymbol{\ell}}(\theta_1, b) \xrightarrow{d} \mathcal{N}_2(\mathbf{0}, -P\ddot{\boldsymbol{\ell}}(\theta_1, b))$, and therefore

$$\tilde{\theta}_2 \xrightarrow{d} \mathcal{N}(b, V)$$

with (asymptotic) variance

$$\begin{aligned}V &= \begin{bmatrix} -\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\mathcal{I}_{11}^{-1} & \mathcal{I}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{bmatrix} \begin{bmatrix} -\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\mathcal{I}_{11}^{-1} \\ \mathcal{I}_{22}^{-1} \end{bmatrix} \\ &= \mathcal{I}_{22}^{-1} - \mathcal{I}_{22}^{-1}\mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\end{aligned}$$

The following theorem summarized the result presented.

Theorem 7 (Asymptotic Normality of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$)**.** *Let $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_2$ be defined as in (3.1) and (3.3), respectively. Let $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{b}$ be the null hypothesis and $\mathcal{I}$ be defined as in (3.6). Under conditions stated previously*

$$(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} \mathcal{N}_q(\mathbf{0}, \mathcal{I}_{11}^{-1})$$

and

$$(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{b}) \xrightarrow{d} \mathcal{N}_{p-q}(\mathbf{0}, \mathcal{I}_{\mathbf{22}}^{-1} - \mathcal{I}_{\mathbf{22}}^{-1}\mathcal{I}_{\mathbf{21}}\mathcal{I}_{\mathbf{11}}^{-1}\mathcal{I}_{\mathbf{12}}\mathcal{I}_{\mathbf{22}}^{-1})$$

3.3 Several Applications

3.3.1 Significance Test

The new method that was proposed has a considerable number of applications, starting with the usual applications in GLM. One could be interested in testing the null hypothesis

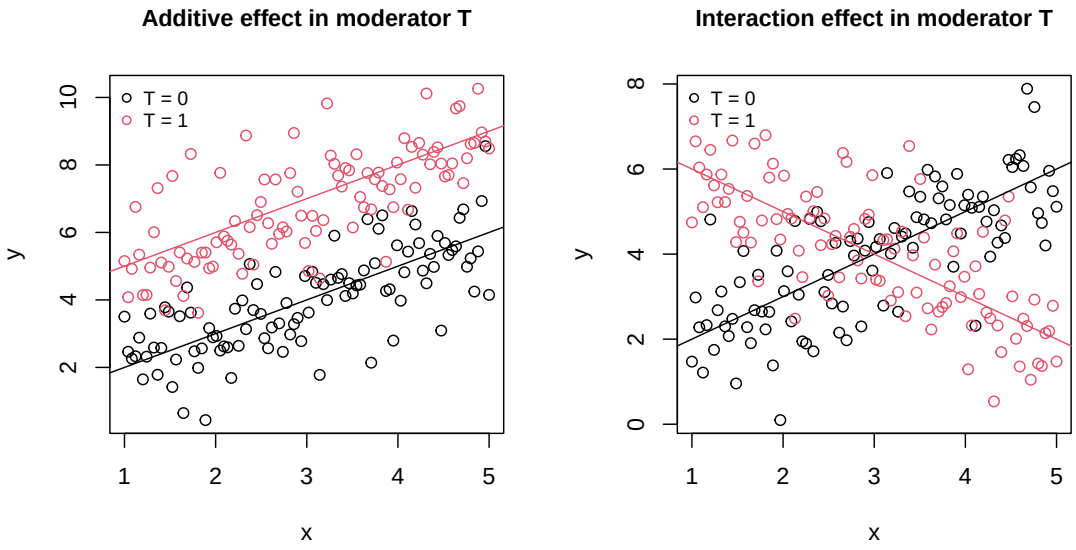

Figure 3.1: Graphical representation of additive (left plot) and interaction (right plot) effect induced by the binary moderator T for a single predictor X

that a set of predictors are not simultaneously significant, i.e., if the model

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_p X_p$$

is fitted, one wants to test $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$ where $0 \leq q < p$, or, equivalently, if $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. In such case, one would like to know whether the model

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$$

fits at least as well as the full model. Either LRT^* and $Wald^*$, defined in (3.4) and (3.8), respectively, can be used to perform the test.

3.3.2 Moderation Analysis

Another application occurs in *moderation analysis*. A moderator between the set of predictors $\mathbf{X} = [X_1, X_2, \ldots, X_p]^T$ and the response variable Y is a third variable T which can be either numerical or categorical. In moderation analysis, it is of interest to evaluate the

effect of the different levels of a mediator in the response variable. One could be interested in fitting either the *additive main-effect model*

$$y = \beta_0 + \beta_1 T + \boldsymbol{\beta}^T \mathbf{X} + \epsilon, \tag{3.13}$$

or the *interaction model*

$$y = \beta_0 + \beta_1 T + \boldsymbol{\beta}^T \mathbf{X} + \boldsymbol{\gamma}^T \mathbf{X}' T + \epsilon. \tag{3.14}$$

where $\mathbf{X}'$ is any subset of predictors.

The simplest case occurs when the moderator T is a dummy variable, i.e., a categorical variable with only two different levels: 0 or 1. In such a case, (3.13) becomes

$$y = \begin{cases} \beta_0 + \boldsymbol{\beta}^T \mathbf{X} + \epsilon, & \text{if } T = 0 \\ (\beta_0 + \beta_1) + \boldsymbol{\beta}^T \mathbf{X} + \epsilon, & \text{if } T = 1 \end{cases}.$$

Notice that, in both cases, the model becomes a linear model and the graphical representation for both cases is a hyperplane with same normal vector but possibly different intercept (see left plot in Figure 3.1). In this case it is of interest to test $H_0 : \beta_1 = 0$ to establish an additive main-effect.

Equation (3.14), on the other hand, can be expressed as

$$y = \begin{cases} \beta_0 + \boldsymbol{\beta}^T \mathbf{X} + \epsilon & \text{if } T = 0 \\ (\beta_0 + \beta_1) + \boldsymbol{\beta}^T \mathbf{X} + \boldsymbol{\gamma}^T \mathbf{X}' + \epsilon, & \text{if } T = 1 \end{cases}$$

If $\mathbf{X}' = \mathbf{X}$, the model for $T = 1$ becomes $y = (\beta_0 + \beta_1) + (\boldsymbol{\beta} + \boldsymbol{\gamma})^T \mathbf{X} + \epsilon$, that is, its graphical representation for both when $T = 0$ and $T = 1$ is a hyperplane with possibly different intercept and possibly different normal vector (see right plot in Figure 3.1). It is of interest to test $H_0 : \boldsymbol{\gamma} = 0$ to discard or confirm an interaction effect.

3.3.3 Over-dispersion

One of the most common problem one has to face when fitting a linear model is heteroscedasticity. In linear regression a set of assumptions needs to meet including the as-

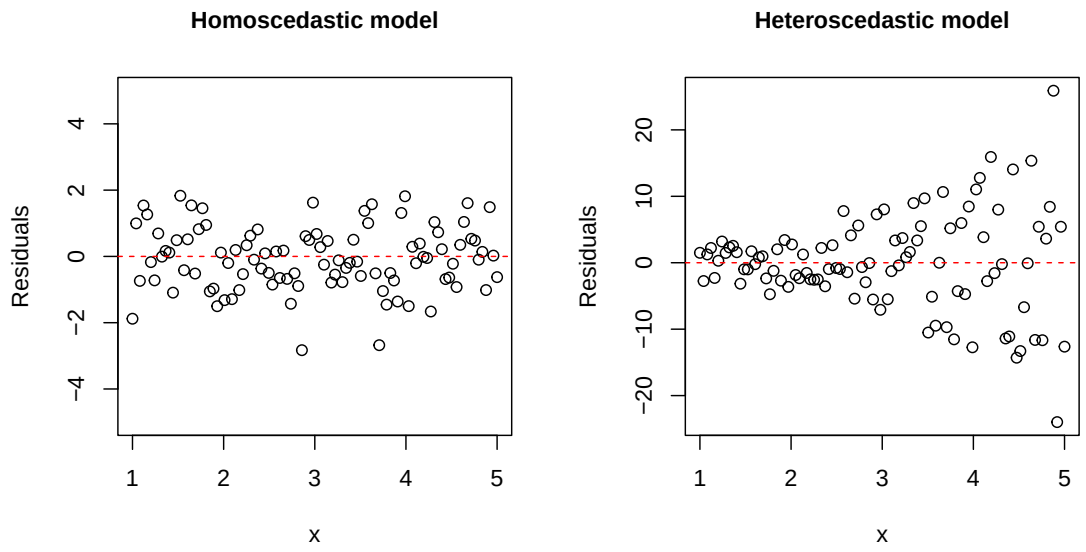

Figure 3.2: Graphical detection of homoscedasticity (left plot) and heteroscedasticity (right plot) in a linear model throughout the residuals $e_i = \hat{y}_i - y_i$

sumption of constant variance. Recall that a linear regression model has the form

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

When σ^2 is not constant, homoscedasticity assumption is violated and inferences regarding the model may not be valid. This problem is also known as *over-dispersion* and occurs when the observed variance in the model is higher than the theoretical variance and seems to vary among observations.

One way to visually detect over-dispersion or heteroscedasticity is to plot the (standardized) residuals, $e = \hat{y} - y$, obtained by fitting the linear model, versus a single predictor x, the response variable y, or the fitted values $\hat{y}$. Figure 3.2 shows both cases when condition of homoscedasticity is met (left plot) and when it is violated (right plot). One of the most known test for heteroscedasticity is the Breusch-Pagan test which assumes a heteroscedastic lineal model of the form

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, h(\mathbf{z}_i^T \boldsymbol{\gamma})) \tag{3.15}$$

where h is a twice differentiable real-valued function which does not depend on the index i, $\boldsymbol{\gamma} \in \mathbb{R}^k$ is not related to $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\mathbf{z}$ is a k-dimensional vector of regressors which first component is assumed to be one and remaining components could be replaced by some predictors in $\mathbf{x}$. The null hypothesis to be tested for heteroscedasticity is

$$H_0 : \gamma_2 = \cdots = \gamma_k = 0.$$

If the null hypothesis is true, $h(\mathbf{z}_i^T\boldsymbol{\gamma}) = h(\gamma_1)$ is constant and then $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = h(\gamma_1)$. Usual choices for h are $h(x) = \exp(x)$ or $h(x) = x^m$ for a known integer m. If $h(\cdot)$ is chosen to be $\exp(\cdot)$, (3.15) becomes

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, \exp(\mathbf{z}_i^T\boldsymbol{\gamma}))$$

Letting $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)^T$ with $\boldsymbol{\gamma}_1 = \gamma_1$ and $\boldsymbol{\gamma}_2 = [\gamma_2, \gamma_3, \ldots, \gamma_k]^T$, one is interested in testing $H_0 : \boldsymbol{\gamma}_2 = \mathbf{0}$.

The following steps must be performed to conduct the Breusch-Pagan test of heteroscedasticity: (i) fit the linear model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, (ii) compute the residuals $e_i = \hat{y}_i - y_i$, (iii) fit the auxiliary linear model $e_i^2 = \mathbf{z}_i^T\boldsymbol{\gamma} + \eta_i$, (iv) get the coefficient of determination, R^2, from previous model, (v) compute the test statistic $LM = nR^2$ where n is the sample size, and (vi) reject $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ if $LM > \chi^2_{k-1}(\alpha)$.

Steps (iv) - (vi) can be omitted and instead to use the proposed statistics to test the null hypothesis $H_0 : \boldsymbol{\gamma}_2 = \mathbf{0}$, where $\boldsymbol{\gamma}_2$ is defined as above. If either LRT^* or $Wald^*$ exceeds $\chi^2_{k-1}(\alpha)$, H_0 can be rejected and heteroscedasticity can be assumed to be present in the model.

Chapter 4

Simulation Studies

Three models have been considered for the simulation. All belonging to the generalized linear models; normal (or Gaussian), Binomial, and Poisson. Data is generated artificially in the following way. Design matrix $\mathbf{X}$ is created assuming that its columns have a multivariate normal distribution with mean zero and variance-covariance $\boldsymbol{\Sigma}$ whose components are defined by $\sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, p$ according to an autoregressive model of first order with ρ being the correlation coefficient. The linear predictor is then computed by the product $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the true parameter vector. The response variable $\mathbf{Y}$ is finally generated according to one of the three models and using that $\mathrm{E}(\mathbf{Y}) = g^{-1}(\boldsymbol{\eta})$ where g is the link function. See Table 2.1.

During the simulation the following values are computed: the LRT, Wald, proposed LRT (mLRT), and proposed Wald (mWald) statistics as well as the time it takes for each of the test to be executed. The empirical sample distribution, size and power of the proposed tests will be investigated and compared with the original LR and Wald tests. The computational time of the proposed tests is also compared with their original counterparts.

4.1 Empirical Sample Distributions

For investigating the sample distribution of the modified tests, three scenarios have been considered. The true parameter vector is set to be of the form

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \quad \text{for} \quad \theta_0 \in \mathbb{R}, \quad \boldsymbol{\theta}_1 \in \mathbb{R}^q, \quad \text{and} \quad \boldsymbol{\theta}_2 \in \mathbb{R}^{p-q} \tag{4.1}$$

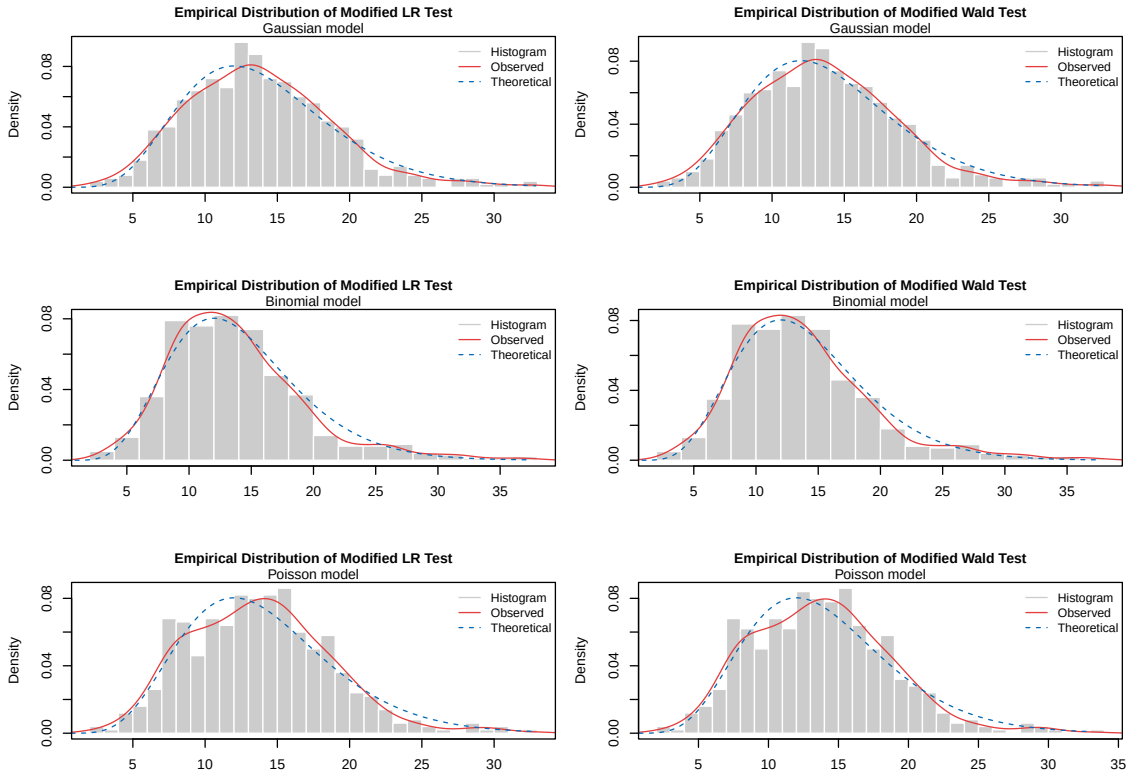

Figure 4.1: Empirical null distribution of the proposed tests in scenario (i) compared with the theoretical asymptotic χ^2_{15-1} distribution.

where all the components of $\boldsymbol{\theta}_1$ equal to one, and all components of $\boldsymbol{\theta}_2$ equal zero; the standard deviation for the Gaussian model is set to one, i.e., $\sigma = 1$, and the correlation coefficient ρ to 0.5. $N = 500$ samples of size $n = 1000$ have been generated artificially with $p = 15$ variables and (i) $q = 1$, (ii) $q = 5$, and (iii) $q = 10$ variables which associated coefficients are non-zero. In all scenarios, the three models are considered. Graphs of the empirical null distributions can be seen in Figures 4.1, 4.2, and 4.3.

Figures show the empirical distribution (histogram and red solid line) and theoretical χ^2_{p-q} distribution (blue dashed line) of the proposed Likelihood Ratio test (left plots) and the proposed Wald test (right plots) of $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$ in the three models. According to what was shown in last chapter, all the empirical distributions should match the theoretical χ^2_{14},

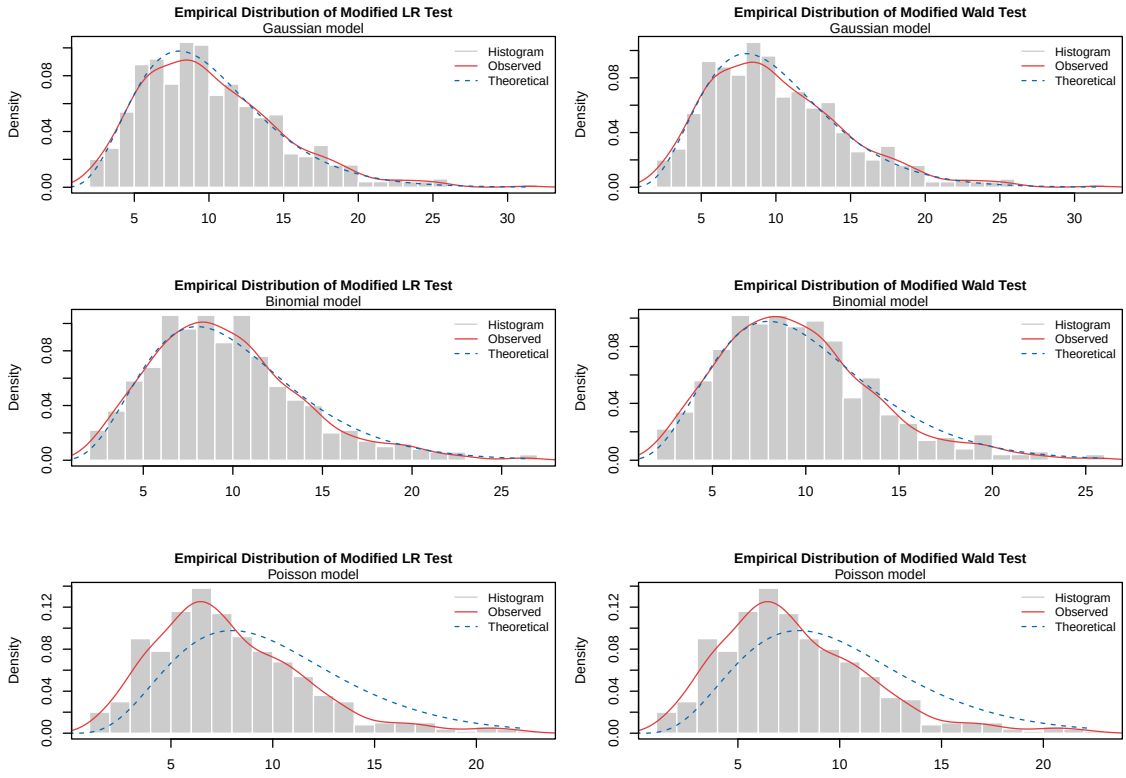

Figure 4.2: Empirical null distribution of the proposed tests in scenario (ii) compared with the theoretical asymptotic χ^2_{15-5} distribution.

χ^2_{10} and χ^2_5 distributions for scenarios (i), (ii), and (iii), respectively.

In Figure 4.1, it can be seen that all empirical sample distributions seems to match the χ^2_{14}. Also, the empirical sample distributions for both the Gaussian and Binomial models in Figures 4.2 and 4.3 coincide with the χ^2_{10} and χ^2_5 distributions. But it also can be observed that the empirical sample distribution for both the proposed LRT (mLRT) and proposed Wald test (mWALD) in the Poisson model mismatch the theoretical χ^2_{10} and χ^2_5 distributions. This mismatch seems to increase as the number of variables with non-zero coefficient associated to them, q, increases. Although it could be hypothesized that, under H_0, the proposed tests still follows an asymptotic χ^2 distribution but with a lower degrees of freedom.

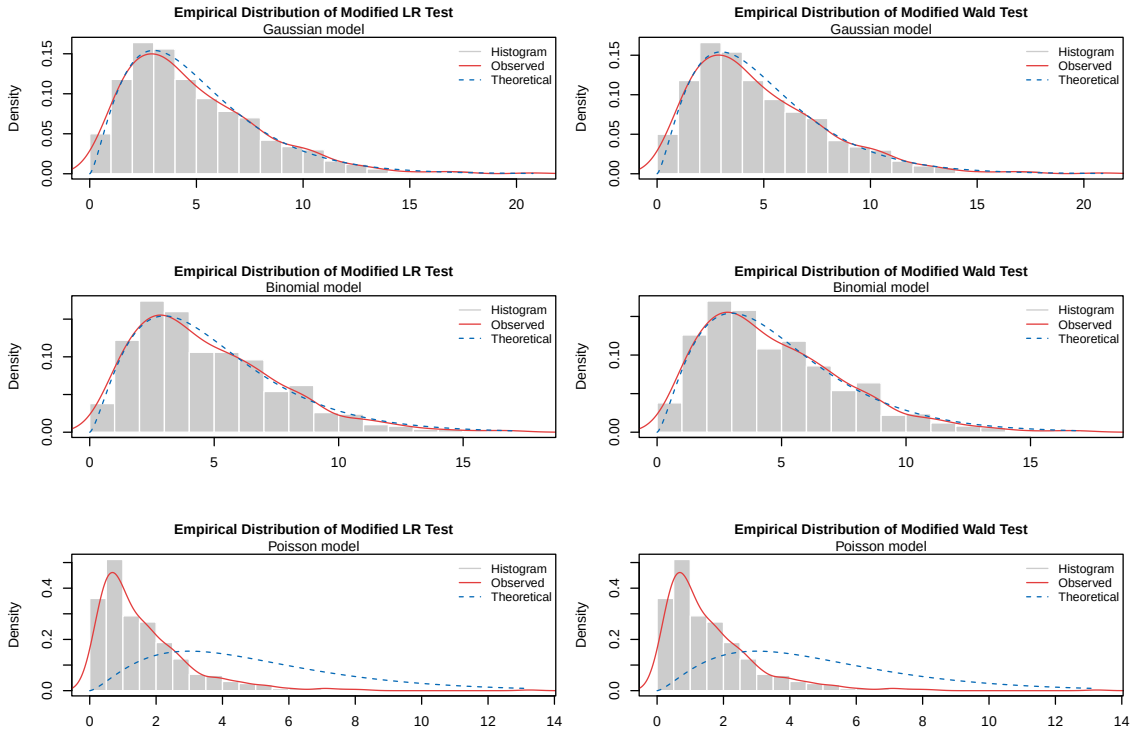

Figure 4.3: Empirical null distribution of the proposed tests in scenario (iii) compared with the theoretical asymptotic χ^2_{15-5} distribution.

Due to these mismatches in the sample distribution for the proposed tests in the Poisson model the following analysis regarding this model could be not valid. Even though, empirical power, size and computational time will be investigated for all three models.

4.2 Empirical Size and Power

The empirical sizes for all scenarios are displayed in Figures 4.4, 4.5, and 4.6. Different significance levels have been used: $\alpha = 0.01, 0.02, \ldots, 0.10$. The red dashed line indicates the desired size of the tests.

Note that the empirical sizes of the proposed tests in all scenarios and models, except for Poisson model in scenario (ii) and (iii), are similar compared with the empirical sizes

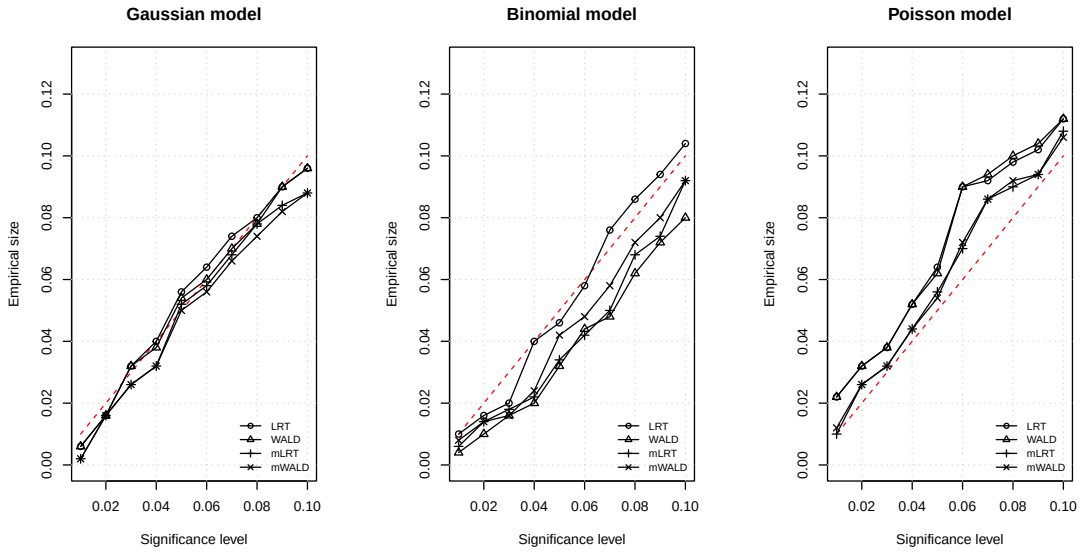

Figure 4.4: Empirical size comparison in scenario (i). Theoretical asymptotic χ^2_{15-1} distribution.

of the original tests. All of them seems to be around the red dashed line which indicates the desired behavior. For the Poisson model, the empirical sizes of the proposed tests are considerably lower than the empirical sizes of the original ones. This is because the mismatch of the empirical sample distribution of the test which was previously discussed.

To investigate the power of the tests in the three scenarios several $(p-q)$-dimensional $\boldsymbol{\theta}_2$ have been generated using the formula

$$\boldsymbol{\theta}_2 = \frac{c}{\sqrt{p-q}}\mathbf{1}_{p-q} \tag{4.2}$$

where c is a positive constant which varies according to the scenario and model, and $\mathbf{1}_{p-q}$ is the $(p-q)$-dimensional vector which components are all one. See values in the x-axis of graphs in Figures 4.7, 4.8, and 4.9. The significance level is set to $\alpha = 0.05$. Power is compared with the euclidean distance, defined by

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2 = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \quad \text{where} \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p, \tag{4.3}$$

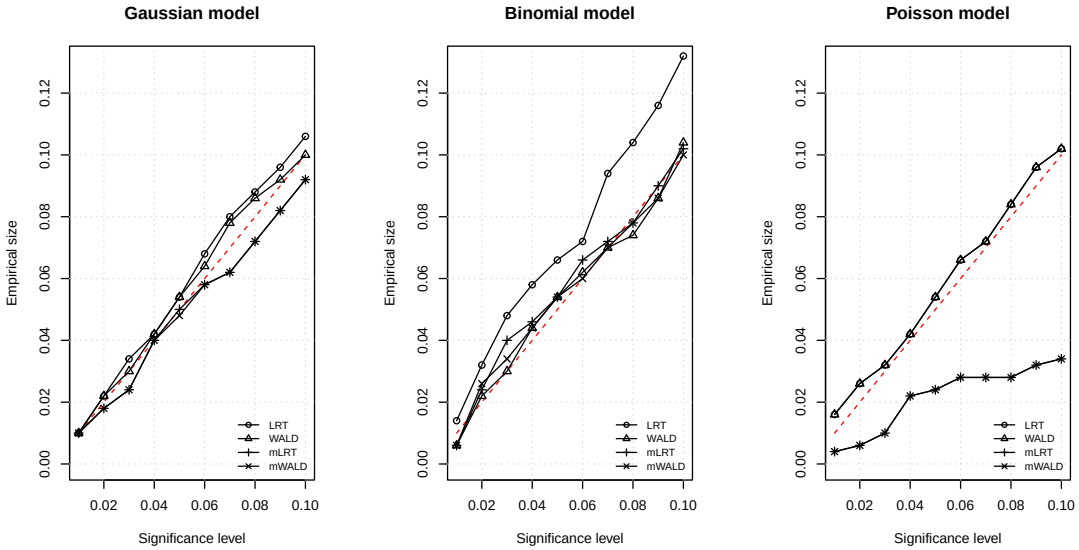

Figure 4.5: Empirical size comparison in scenario (ii). Theoretical asymptotic χ^2_{15-5} distribution.

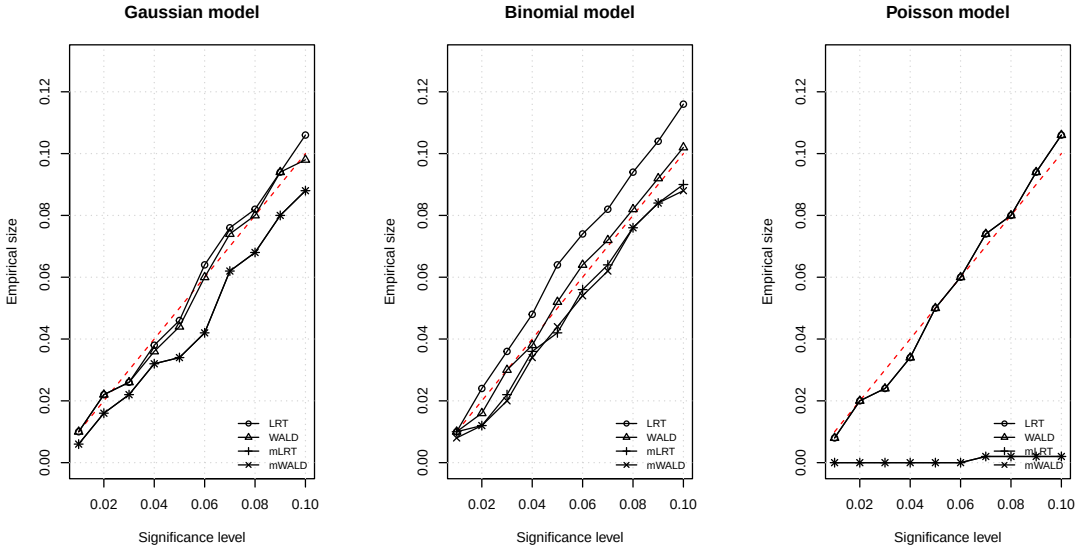

Figure 4.6: Empirical size comparison in scenario (i). Theoretical asymptotic χ^2_{15-10} distribution.

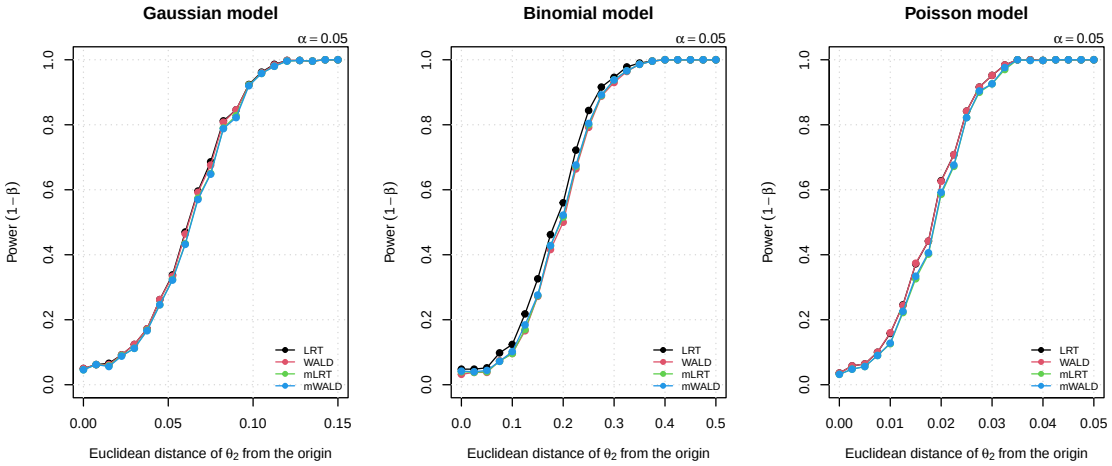

Figure 4.7: Empirical Power Comparison in scenario (i)

of $\boldsymbol{\theta}_2$ from $\mathbf{0} \in \mathbb{R}^{p-q}$ which is equal to the positive constant c chosen:

$$d_2(\boldsymbol{\theta}_2, \mathbf{0}) = ||\boldsymbol{\theta}_2||_2 = \frac{c}{\sqrt{p-q}}\sqrt{p-q} = c.$$

according to (4.2) and (4.3).

It can be noticed that the power of the proposed tests in the different scenarios and models is extremely similar to the original ones, except for the power of the tests in the scenarios (ii) and (iii) for the Poisson model which is not surprising since the χ^2_{p-q} distribution is used to compute the power.

Also, comparisons among the tests of the power with varying sample size are shown in Figures 4.10, 4.11, and 4.12. All scenarios and models are considered but in case of the binomial model, higher sample sizes had to be set to avoid what is known as the complete separation problem which could inflate extremely the estimated coefficients. The significance level α is set to 0.05, $\boldsymbol{\theta}_2$ is chosen to be of the form specified in (4.2) with c being a fixed positive constant for each model. In case of the Gaussian model $c = 0.4$ and $n = 20, 30, \ldots, 150$ have been chosen; in the binomial model $c = 2.0$ and $n = 100, 150, \ldots, 1000$ have been set; and in the Poisson model $c = 0.1$ and $n = 40, 50, \ldots, 150$ was selected.

What can be observed from Figure 4.10 is that the power of the proposed tests for

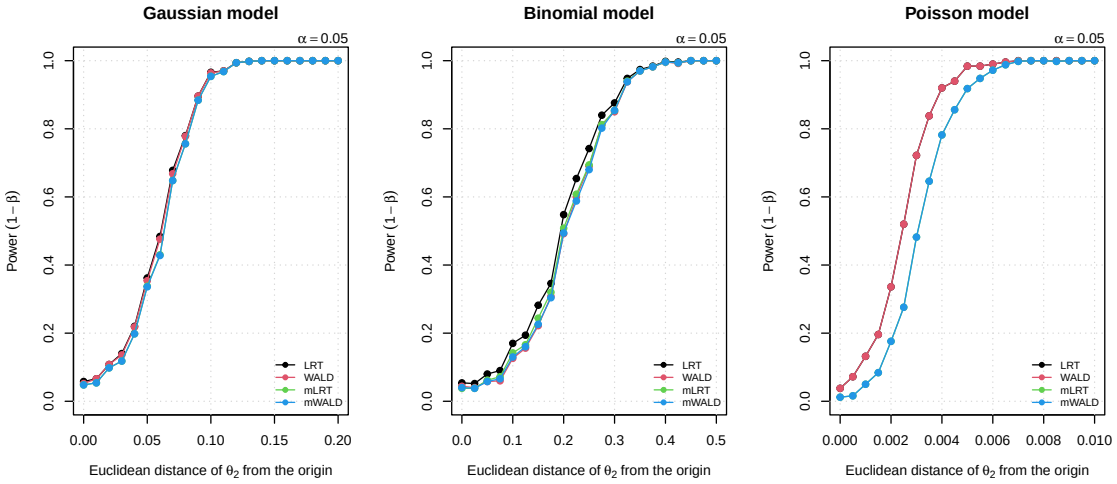

Figure 4.8: Empirical Power Comparison in scenario (ii)

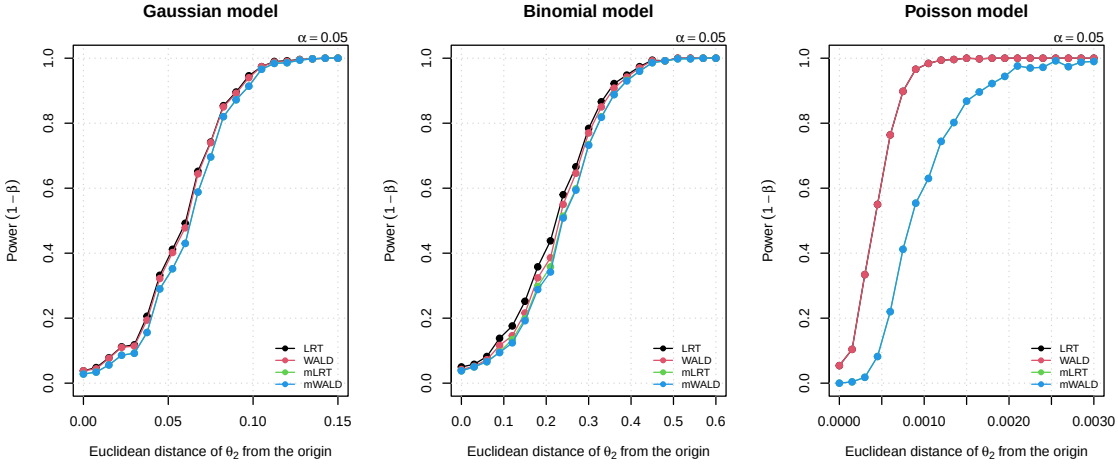

Figure 4.9: Empirical Power Comparison in scenario (iii)

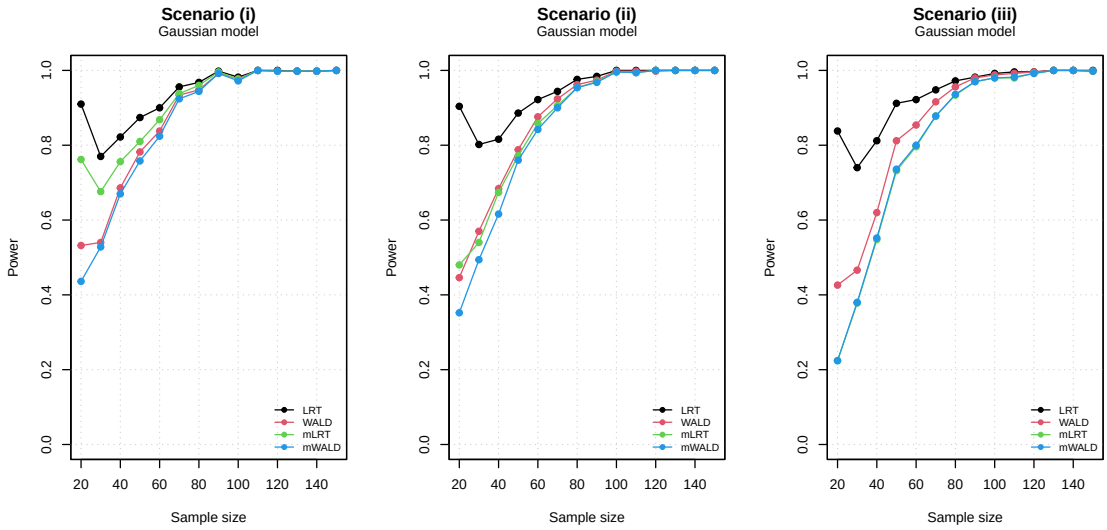

Figure 4.10: Empirical Power vs. Sample Size in Gaussian Model

small sample sizes differs considerably from the original tests in scenario (i) where $q = 1$. In scenario (ii) all tests but LRT seems to be similar even for small sample sizes. In all the scenarios, LRT shows a higher power than the rest but for moderate sample size, the power among the four tests is almost equal. The comparison of power for different sample sized in the binomial model in scenario (i) seems to be significantly different for sample sizes smaller than 300. As long as q is varying according to the scenario, the power among the tests in this model gets closer and closer. In case of the Poisson model the power among the tests in the scenario (i) is extremely similar for all sample sizes. In contrast, the power of the proposed tests in the scenarios (ii) and (iii) differ considerably from the original tests.

4.3 Computational Time Comparison

For time running performance four different sample sizes are considered: $n = 30, 50, 100,$ 1000 and the true parameter vector is selected according to (4.1). Initial settings in the three scenarios mentioned at the beginning of the chapter are used. Table 4.1 shows the computational time for the four tests in the three scenarios and different values of n.

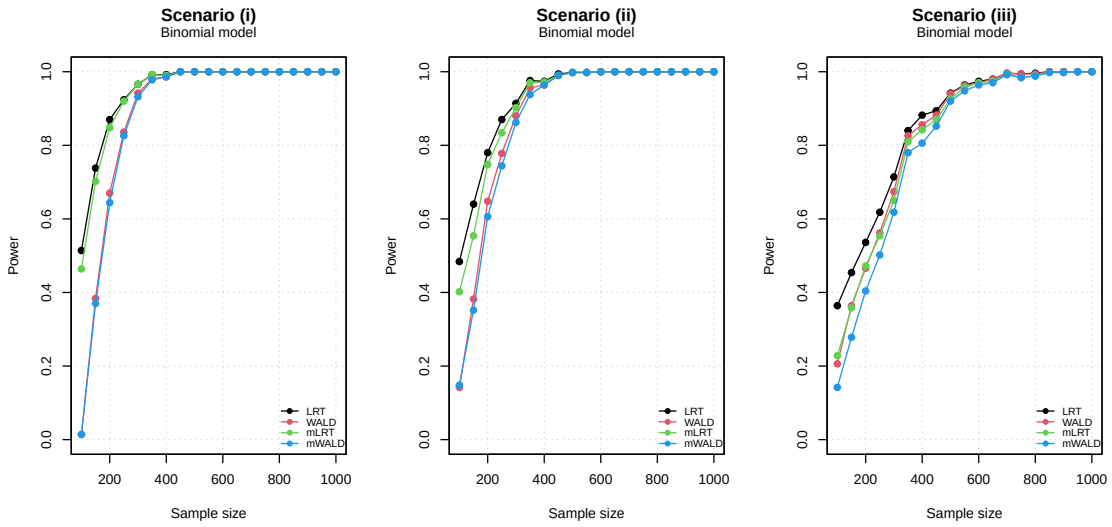

Figure 4.11: Empirical Power vs. Sample Size in Binomial Model

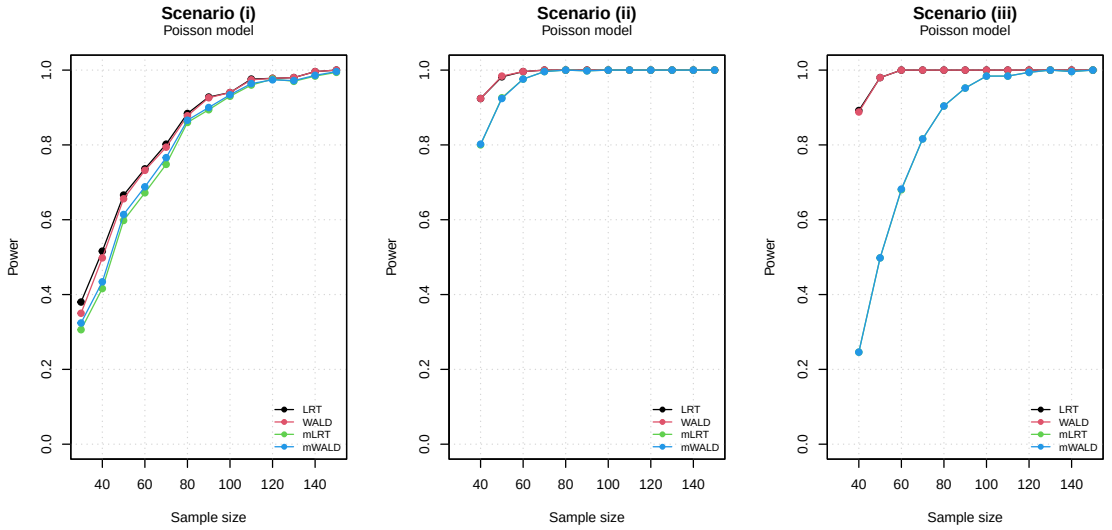

Figure 4.12: Empirical Power vs. Sample Size in Poisson Model

Note the similarity in execution time among the tests. One can observed that the time the proposed LRT in the Binomial and Poisson models takes to execute is almost always smaller than the time taken by the original LRT. On the other hand, the Wald test seems to be slightly faster than the proposed one in almost all the cases and models. This could be justified by the way the Wald test and the proposed Wald are computed. When computing the Wald test only the full model

$$g(\mu) = \theta_0 + \theta_1 X_1 + \cdots + \theta_q X_q + \cdots + \theta_p X_p \tag{4.4}$$

is fitted, but when the proposed Wald test is calculated, the two models:

$$g(\mu) = \theta_0 + \theta_1 X_1 + \cdots + \theta_q X_q, \tag{4.5}$$

$$g(\mu) = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \cdots + \hat{\theta}_q X_q + \theta_{q+1} X_{q+1} + \cdots + \theta_p X_p \tag{4.6}$$

need to be fitted. Here $\hat{\theta}_i$ in (4.6) is the ith estimated coefficient obtained when fitting the reduced model in (4.5).

The estimation of the coefficients in the three models has been done using the `glm` function available in `R` which internally implements the Newton-Raphson method. One way to improve the execution time in the Wald tests is to implement them from the beginning as well as rewrite in a proper way the updating formula used in the Newton-Raphson method (see Equation 2.4). From the optimization perspective, the proposed methods breaks an optimization problem involving p parameters into two smaller ones which requires the estimation of q and $p - q$ parameters. This approach could not be significantly more (computationally) efficient than the original methods in lower-dimensional cases but when a high-dimensional problem is faced the execution time could be considerably improved.

To see the point described in the previous paragraph, the computational time of the four tests is examined. $N = 100$ samples of size $n = 1000$ are generated for each of the models with $p = 100$ predictor variables of which $q = 5, 10, \ldots, 95$ has a non-zero associated coefficient. Results are shown in Figure 4.13.

Table 4.1: Computational Time for the Four Tests and the Three Models in the Three Scenarios with Different Sample Size

		Gaussian				**Binomial**				**Poisson**			
q	n	LRT	Wald	mLRT	mWald	LRT	Wald	mLRT	mWald	LRT	Wald	mLRT	mWald
1	30	0.28	0.27	0.22	0.36	0.35	0.31	0.37	0.36	0.33	0.20	0.35	0.26
	50	0.17	0.30	0.25	0.32	0.49	0.26	0.32	0.39	0.40	0.21	0.29	0.29
	100	0.25	0.27	0.28	0.25	0.40	0.12	0.35	0.38	0.44	0.22	0.30	0.32
	1000	0.52	0.36	0.51	0.42	0.96	0.70	0.82	0.96	0.78	0.57	0.94	0.84
5	30	0.29	0.14	0.33	0.31	0.36	0.35	0.37	0.37	0.38	0.28	0.24	0.31
	50	0.32	0.24	0.21	0.30	0.55	0.39	0.46	0.40	0.37	0.22	0.30	0.30
	100	0.54	0.31	0.24	0.24	0.36	0.23	0.40	0.40	0.36	0.22	0.39	0.37
	1000	0.43	0.27	0.45	0.51	1.17	1.00	0.86	1.03	0.97	0.55	0.69	0.76
10	30	0.32	0.19	0.32	0.22	0.83	0.49	0.36	0.29	0.43	0.20	0.34	0.26
	50	0.34	0.15	0.30	0.28	0.60	0.40	0.49	0.37	0.44	0.22	0.26	0.36
	100	0.23	0.32	0.25	0.30	0.75	0.37	0.67	0.66	0.38	0.23	0.32	0.43
	1000	0.64	0.30	0.51	0.30	1.51	0.65	1.08	1.23	0.67	0.50	0.81	0.75

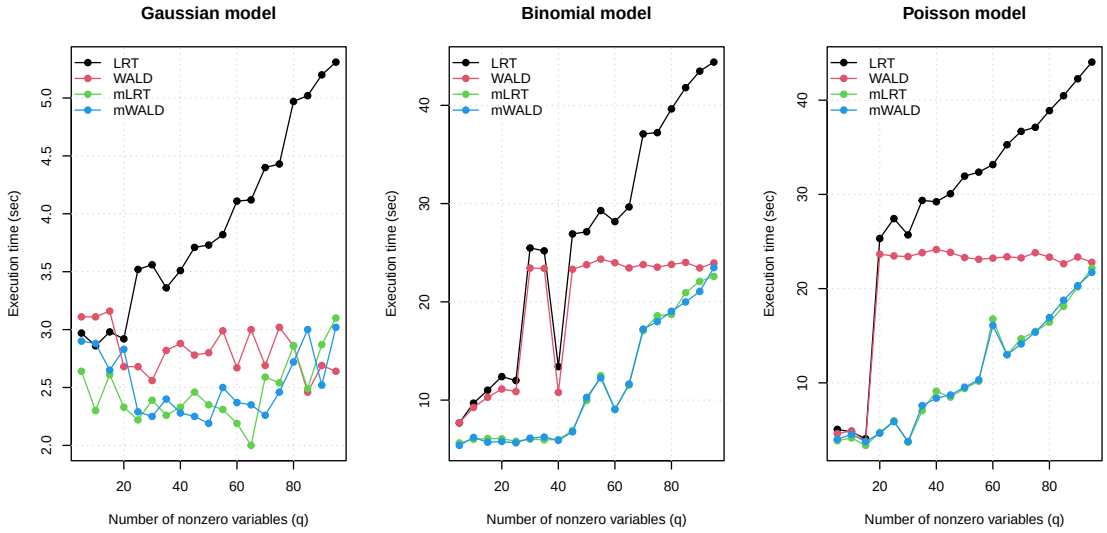

Figure 4.13: Execution Time Comparison

Note that in all the models, the execution time in the original LR test increases considerably compared with the rest of the tests. Wald, the proposed LR, and the proposed Wald tests seems to behave similarly for the Gaussian model. As expected, the execution time increases as q increases. Also note the considerable increase in time when $q = 30$ and $q = 10$ in the original LRT and Wald tests for the Binomial and Poisson models. The execution time for the proposed tests seems to be quite similar to each other even when two fits are used to calculate the value of the proposed Wald test compared with the proposed LRT which requires only one. The more considerable change occurs in the original LRT which increases drastically compared with the proposed one. Something surprising is the behavior of the execution time for the Wald test which remains constant after a certain value of q in both the Binomial and the Poisson models. Even though, the proposed methods seems to be more (computationally) efficient when a high-dimensional problem is faced.

Chapter 5

Real Data Examples

To illustrate the application of the proposed methods, three real datasets will be analyzed: `BostonHousing`, `HepatitisC`, and `DoctorVisits` datasets, for linear, logistic, and Poisson regression, respectively.

5.1 The `BostonHousing` dataset

The `BostonHousing` dataset is available in the `mlbench` `R` package which was taken from the UCI Repository of Machine Leaning Datasets. It has been originally published by Harrison and Rubinfeld (1978) in the paper titled *Hedonic prices and the demand for clear air* and used in the book *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity* written by Belsley, Kuh, and Welsch (1980). The dataset contains information taken from the 1970 census concerning housing in Boston. It contains one response `medv`, 13 predictors, and 506 instances. Table 5.1 describes the attributes present in the data.

The partial output of the linear model fit of `medv` on the 13 regressors is shown in R Code 5.1. A full output can be found in Appendix B in R Code B.1. It could be of interest, by looking at the p-values of the estimated coefficients, to test the simultaneous significance of `indus` and `age`. In Table 5.2 the test values and p values for each of the four tests (LRT, Wald and the proposed ones) are displayed. p-values were obtained using the χ^2 distribution with 2 degrees of freedom.

Observe that all test values are similar to each other and therefore the p-values. All four tests strongly suggest not to reject the null hypothesis $H_0 : \beta_{\texttt{indus}} = \beta_{\texttt{age}} = 0$, and therefore, the reduced model could be considered a better fit.

Table 5.1: Description of attributes in `BostonHousing` dataset.

Name	**Description**
`crim`	Per capita crime rate by town
`zn`	proportion of residential land zoned for lots over 25,000 sq.ft
`indus`	Proportion of non-retail business acres per town
`chas`	Charles River dummy variable (1-if tract bounds river; 0-otherwise)
`nox`	Nitric oxides concentration (parts per 10 million)
`rm`	Average number of rooms per dwelling
`age`	Proportion of owner-occupied units built prior to 1940
`dis`	weighted distances to five Boston employment centres
`rad`	Index of accessibility to radial highways
`tax`	Full-value property-tax rate per USD 10,000
`ptratio`	Pupil-teacher ratio by town
`b`	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
`lstat`	Percentage of lower status of the population
`medv`	Median value of owner-occupied homes in USD 1000's (target)

Table 5.2: Test values and p-values of the four tests for testing

$$H_0 : \beta_{\texttt{indus}} = \beta_{\texttt{age}} = 0$$

	Ordinary		Proposed	
	LRT	Wald	LRT	Wald
Test value	0.1177937	0.1145480	0.1172407	0.1142418
p-value	0.9428040	0.9443353	0.9430647	0.9444799

R code 5.1: Partial output of linear model fit applied to `BostonHousing` dataset

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas1        2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
b            9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
```

Table 5.3: Description of attributes in `HepatitisC` dataset

Name	Description	Name	Description
`Category`	Hepatitis C category (target)	`BIL`	Bilirubin
`Age`	Age of individual in years	`CHE`	Choline esterase
`Sex`	Sex of individual (1-female, 0-male)	`CHOL`	Cholesterol
`ALB`	Albumin	`CREA`	Creatinine
`ALP`	Alkaline phosphatase	`GGT`	γ-glutamyl-transferase
`ALT`	Alanine amino-transferase	`PROT`	Protein
`AST`	Aspartate amino-transferase		

5.2 The `HepatitisC` dataset

For logistic regression, the `HepatitisC` dataset is used. This dataset was created and used by Lichtinghagen, Klawonn, and Hoffmann in the paper *Using machine learning techniques to generate laboratory diagnostic pathways—a case study* published in 2018. It is available in the UCI Machine Learning Repository and contains 11 numerical attributes, 10 of them being obtained via blood test: `Age`, `ALB`, `ALP`, `ALT`, `AST`, `BIL`, `CHE`, `CHOL`, `CREA`, `GGT`, and `PROT` (see Table 5.3 for a complete description of the predictors); two categorical attributes: `Sex` and `Category` which values corresponds to one of the five different levels: "0=Blood Donor", "Os=suspect Blood Donor", "1=Hepatitis", "2=Fibrosis","3=Cirrhosis" and 615 instances. The five levels in the response variable `Category` have been collapsed so that 0 corresponds to "0=Blood Donor" or "0s=suspect Blood Donor" levels and 1 corresponds to "1=Hepatitis", "2=Fibrosis", or "3=Cirrhosis" levels. The dataset also contains some missing values in variables `ALB`, `ALP`, `ALT`, `CHOL`, and `PROT`. Imputation has been implemented on the dataset using the `R` function `missForest` available in the package with the same name.

R Code 5.2 shows the partial output of the fit of `Category` on the 12 regressors. As can

be observed in the partial output, `Age`, `ALB`, and `CHE` seem to be non-significant. Complete output can be found in Appendix B in R Code B.2.

R code 5.2: Partial output of logistic model fit applied to `HepatitisC` dataset

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.445193   3.928088  -3.677 0.000236 ***
Age          -0.006511   0.024304  -0.268 0.788772
Sex           1.344974   0.608607   2.210 0.027111 *
ALB          -0.122350   0.063592  -1.924 0.054355 .
ALP          -0.070402   0.012666  -5.558 2.72e-08 ***
ALT          -0.020408   0.009422  -2.166 0.030317 *
AST           0.095025   0.019294   4.925 8.43e-07 ***
BIL           0.086677   0.028910   2.998 0.002716 **
CHE           0.125628   0.121698   1.032 0.301933
CHOL         -0.690593   0.258862  -2.668 0.007635 **
CREA          0.024286   0.005208   4.663 3.11e-06 ***
GGT           0.030608   0.006211   4.928 8.30e-07 ***
PROT          0.225520   0.058047   3.885 0.000102 ***
```

The null hypothesis $H_0 : \beta_{\texttt{Age}} = \beta_{\texttt{ALB}} = \beta_{\texttt{CHE}} = 0$ has been tested using the four methods and the χ^2 distribution with 3 degrees of freedom. Results are shown in Table 5.4. Notice again that all test values and p-values are pretty similar to each other. The four tests suggest to consider the reduced model.

5.3 The `DoctorVisits` dataset

A final model will be fitted using the `DoctorVisits` dataset. The dataset is available in the `R` package `AER`. It was originally published by the Jornal of Applied Econometrics Data

Table 5.4: Test values and p-values of the four tests for testing

$$H_0 : \beta_{\texttt{Age}} = \beta_{\texttt{ALB}} = \beta_{\texttt{CHE}} = 0$$

	Ordinary		Proposed	
	LRT	Wald	LRT	Wald
Test value	4.061120	3.915635	4.058021	4.062995
p-value	0.2549394	0.2707193	0.2552667	0.2547417

Archive and has been used by Cameron and Trevedi (1986) in the paper titled *Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests*. It contains 12 attributes and 5190 instances collected from the 1977-1978 Australian Health Survey. Table 5.5 describes the attributes of the data.

The response variable `visits` has been regressed on the 11 predictor variables assuming a Poisson model. Partial output can be found in R Code 5.3. In Appendix B can be found the complete output at R Code B.3. One could be interested on testing the null hypothesis that the coefficients of `age`, `private`, `freerepat`, `nchronic`, and `lchronic` are simultaneously equal to zero.

Notice the evident difference in the test values for the proposed tests as when as the difference in the p-values shown in Table 5.6. In all the tests, a χ^2 distribution with 5 degrees of freedom has been used to compute the p-values. The ordinary tests suggest to reject the null hypothesis

$$H_0 : \beta_{\texttt{age}} = \beta_{\texttt{private}} = \beta_{\texttt{freerepat}} = \beta_{\texttt{nchronic}} = \beta_{\texttt{lchronic}} = 0$$

with a p-value around 0.003. On the other hand, the proposed tests suggest not to reject H_0 if a significance level of $\alpha = 0.01$ is considered. This is not surprising though, since in Section 4, the null empirical distribution for the Poisson model seemed not to follow the χ^2_{p-q} asymptotic distribution but a χ^2 asymptotic distribution with lower degrees of freedom.

Table 5.5: Description of attributes in `DoctorVisits` dataset.

Name	**Description**
`visits`	Number of doctor visits in past 2 weeks (target)
`gender`	Factor indicating gender (0-male, 1-female)
`age`	Age in years divided by 100
`income`	Annual income in tens of thousands of dollars
`illness`	Number of illnesses in past 2 weeks
`reduced`	Number of days of reduced activity in past 2 weeks due to illness or injury
`health`	General health questionnaire score using Goldberg's method
`private`	Factor indicating whether the individual have private health insurance
`freepoor`	Factor indicating whether the individual have free government health insurance due to low income
`freerepat`	Factor indicating whether the individual have free government health insurance due to old age, disability or veteran status
`nchronic`	Factor indicating whether there is a chronic condition not limiting activity
`lchronic`	Factor indicating whether there is a chronic condition limiting activity

Table 5.6: Test values and p-values of the four tests for testing

$H_0 : \beta_{\texttt{age}} = \beta_{\texttt{private}} = \beta_{\texttt{freerepat}} = \beta_{\texttt{nchronic}} = \beta_{\texttt{lchronic}} = 0$

	Ordinary		Proposed	
	LRT	Wald	LRT	Wald
Test value	18.02394	17.63733	12.24816	12.32561
p-value	0.002916555	0.003437014	0.031541208	0.030588536

R code 5.3: Partial output of Poisson model fit applied to `DoctorVisits` dataset

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.097821   0.101554 -20.657  < 2e-16 ***
gender       0.156490   0.056139   2.788  0.00531 **
age          0.279123   0.165981   1.682  0.09264 .
income      -0.187416   0.085478  -2.193  0.02834 *
illness      0.186156   0.018263  10.193  < 2e-16 ***
reduced      0.126690   0.005031  25.184  < 2e-16 ***
health       0.030683   0.010074   3.046  0.00232 **
private      0.126498   0.071552   1.768  0.07707 .
freepoor    -0.438462   0.179799  -2.439  0.01474 *
freerepat    0.083640   0.092070   0.908  0.36365
nchronic     0.117300   0.066545   1.763  0.07795 .
lchronic     0.150717   0.082260   1.832  0.06692 .
```

Chapter 6

Discussion and Conclusions

In this document, modified versions of the ordinary Wald and LR tests for testing the null hypothesis $H_0 : \boldsymbol{\theta}_2 = \mathbf{b}$ were presented. They were based on the profile likelihood framework which consisted in partitioning the parameter $\boldsymbol{\theta}$ into two subparameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The subparameter $\boldsymbol{\theta}_1$ was estimated by maximizing the log-likelihood function under H_0, that is, maximizing $\ell(\boldsymbol{\theta}_1, \mathbf{b})$ with respect to $\boldsymbol{\theta}_1$. This estimation, say $\tilde{\boldsymbol{\theta}}_1$, was used to finally estimate $\boldsymbol{\theta}_2$ by maximizing the log-likelihood function evaluated at $\boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1$, i.e. maximizing $\ell(\tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$. Then, the modified Wald test was defined by

$$W^* = (\tilde{\boldsymbol{\theta}}_2 - \mathbf{b})^T (\mathcal{I}^{-1}_{\mathbf{22}} - \mathcal{I}^{-1}_{\mathbf{22}} \mathcal{I}_{\mathbf{21}} \mathcal{I}^{-1}_{\mathbf{11}} \mathcal{I}_{\mathbf{12}} \mathcal{I}^{-1}_{\mathbf{22}})^{-1} (\tilde{\boldsymbol{\theta}}_2 - \mathbf{b})$$

and the modified LRT by

$$LRT^* = 2[\ell(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) - \ell(\tilde{\boldsymbol{\theta}}_1, \mathbf{b})].$$

In the simulation, three models have been tested under the GLM framework: the normal, binomial, and Poisson models. Different scenarios were considered where the main focus was varying the number of variables with non-zero associated coefficients, q. It could be observed that the empirical distribution of the modified tests follow an asymptotic χ^2_{p-q} distribution for the normal and binomial model. Also, it was observed that the empirical power and size of the proposed tests behave similar to the ordinary tests. Regarding the computational time performance, one could observed that when the total number of variables p is small the execution time among the ordinary and the proposed tests is similar to each other. However, when p increases the execution time for both ordinary Wald and LR tests increases drastically compared with the proposed ones.

One of the main advantages of the modified tests is the execution time. It could be observed that the execution time of the proposed tests was considerably lower compared to the ordinary tests when p is high. Moreover, the asymptotic properties, the power and size of the new tests were similar to the ordinary ones. Another advantage is that both tests can be used even if p is slightly bigger than n as long as neither q nor $p-q$ exceed n which cannot be done in the ordinary tests. One disadvantage of the proposed Wald test is that it requires to fit two models instead of fitting one as is in the case of the ordinary Wald test. Ordinary Wald test performed slightly faster than the proposed one. Unfortunately, the proposed test did not performed well in the Poisson model compared to the ordinary versions. These issues encountered regarding the proposed tests conduct to the following future work:

1. As was presented in the simulation studies, one can be interested in investigating the problem faced with the null asymptotic distribution in the Poisson model. In that chapter was hypothesized that the proposed Wald and LR tests for the Poisson model in the GLM framework followed an χ^2 asymptotic distribution without a formal proof.

2. The proposed tests has been supposed to be applied to datasets with $p \ll n$ assuming that p is fixed and n goes to infinity. It could be of interest to study the case where $p \geq n$.

References

[1] O. E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Monographs on Statistics and Applied Probability 52. Springer US, 1994.

[2] Rabi Bhattacharya, Lizhen Lin, and Victor Patrangenaru. *A Course in Mathematical Statistics and Large Sample Theory*. Springer Texts in Statistics. Springer-Verlag New York, 1 edition, 2016.

[3] T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47:1287–1294, 1979.

[4] A. Buse. The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, 36:153–157, 1982.

[5] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. CRC, 4th edition, 2018.

[6] Michael H. Kutner, Christopher J Nachtsheim, and John Neter. *Applied Linear Regression Models*. Ingram, 2004.

[7] P. McCullagh and John A. Nelder. *Generalized linear models*. Chapman Hall/CRC Monographs on Statistics Applied Probability. Chapman and Hall/CRC, 2 edition, 1989.

[8] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[9] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62, 1938.

[10] Jin Zhang. Consistency of mle, lse and m-estimation under mild conditions. *Statistical Papers*, 2017.

[11] Z. Zhang and L. Wang. *Advanced statistics using R.* ISDSA Press, 2017.

[12] Zhu Zhongyi and Wei Bocheng. Asymptotic properties of mle in exponential family nonlinear models. *Journal of Systems Science and Complexity*, 10:193–201, 1997.

Appendix A

Theorems and Proofs

Theorem 8. *If* $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, then* $(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_p$*.*

Theorem 9 (Asymptotic Distribution of Wald Test)**.** *Let* $\hat{\boldsymbol{\theta}}$ *be an* M*-estimator of* $\boldsymbol{\theta}_0$ *and* $\psi(\boldsymbol{\theta}) = \ell'(\boldsymbol{\theta})$*. Suppose conditions stated in Theorems 6 and 5*

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \chi^2_p$$

Proof. Assume all conditions in Theorems 6 and 5 are satisfied. Then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N_p\left(\mathbf{0}, (P\psi'(\boldsymbol{\theta}_0))^{-1} P\psi(\boldsymbol{\theta}_0)\psi^T(\boldsymbol{\theta}_0)(P\psi'(\boldsymbol{\theta}_0))^{-1}\right)$$

Recall that $P\psi(\boldsymbol{\theta}_0)\psi^T(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0)$ and that $P\psi'(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0)$ therefore $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_0))$ and

$$n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \chi^2_p$$

□

Theorem 10. *Let* $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *be partitioned as*

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$$

where $\boldsymbol{X}_1$ *is* $q \times 1$ *and* $\boldsymbol{X}_2$ *is* $(p-q) \times 1$*. Moreover, if* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *are also partitioned as*

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

with $\boldsymbol{\mu}_1 \in \mathbb{R}^q$*,* $\boldsymbol{\mu}_2 \in \mathbb{R}^{p-q}$*,* $\boldsymbol{\Sigma}_{11}$*, and* $\boldsymbol{\Sigma}_{22}$ *being* $q \times q$ *and* $(p-q) \times (p-q)$ *matrices, respectively. Then* $\boldsymbol{X}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ *and* $\boldsymbol{X}_2 \sim \mathcal{N}_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$*.*

Theorem 11. *Let $\boldsymbol{A}$ be a $k \times p$ matrix of rank k. If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{AX} \sim \mathcal{N}_k(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$.*

Theorem 12. *Let $\hat{\boldsymbol{\theta}}$ be an M-estimator of $\boldsymbol{\theta} \in \mathbb{R}^p$. Consider partitioning both $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ as*

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\boldsymbol{\theta}}_1 \\ \hat{\boldsymbol{\theta}}_2 \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}$$

where $\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1 \in \mathbb{R}^q$. Let $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{b}$ be the null hypothesis to be tested for some constant vector $\boldsymbol{b}$. Suppose conditions in Theorems 6 and 5 are satisfied. Then the Wald test, defined by

$$W = n(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{b})^T \boldsymbol{V}_{22}^{-1} (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{b})$$

where $\boldsymbol{V}_{22}$ is the $(p-q) \times (p-q)$ submatrix of the variance-covariance matrix $Var(\boldsymbol{\theta}) = \boldsymbol{V}$ partitioned as

$$\boldsymbol{V} = \left[\begin{array}{c:c} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \hdashline \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{array}\right],$$

has a χ^2 asymptotic distribution with $p - q$ degrees of freedom.

Proof. Suppose conditions in the theorems mentioned are satisfied, then as shown in Theorem 9, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}))$.

Let $\mathbf{A}$ be a $(p-q) \times p$ matrix of the form

$$\mathbf{A} = \left[\begin{array}{c:c} \mathbf{0} & \mathbf{I} \end{array}\right]$$

where $\mathbf{I}$ is the $p - q$ identity matrix. Notice that the rank of $\mathbf{A}$ is $p - q$. Then, $\mathbf{A}\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_2$, $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\theta}_2 = \mathbf{b}$ under H_0, and, by Theorem 11,

$$\sqrt{n}(\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{A}\boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_{p-q}(\mathbf{0}, \mathbf{A}\mathcal{I}^{-1}(\boldsymbol{\theta})\mathbf{A}^T)$$

Notice that $\mathcal{I}^{-1}(\boldsymbol{\theta}) = n\text{Var}(\boldsymbol{\theta}) = n\mathbf{V}$. Therefore

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \mathbf{b}) \xrightarrow{d} \mathcal{N}_{p-q}(\mathbf{0}, n\mathbf{A}\mathbf{V}\mathbf{A}^T)$$

Or, equivalently

$$(\hat{\boldsymbol{\theta}}_2 - \mathbf{b})^T \mathbf{V}_{22}^{-1} (\hat{\boldsymbol{\theta}}_2 - \mathbf{b}) \xrightarrow{d} \chi^2_{p-q}$$

□

Theorem 13 (Asymptotic Distribution of LRT)**.** *Let $\ell(\boldsymbol{\theta})$ be the log-likelihood of a random sample $X_1, ..., X_n$. Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. Suppose $\ell(\boldsymbol{\theta}) \in C^3$ and that all conditions stated in Theorems 6 and 5 are met. Then*

$$LRT = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)] \xrightarrow{d} \chi^2_p$$

Proof. Since all conditions in Theorems 6 and 5 are satisfied, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_0))$. Suppose also that $\ell \in C^3$ and consider the Taylor expansion of ℓ at $\boldsymbol{\theta}_0$ around $\hat{\boldsymbol{\theta}}$

$$\ell(\boldsymbol{\theta}_0) = \ell(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \nabla \ell(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \mathbf{H}_\ell(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) + o_p(1)$$

where $\nabla \ell(\hat{\boldsymbol{\theta}})$ is the gradient of ℓ at $\hat{\boldsymbol{\theta}}$ and $\mathbf{H}_\ell(\hat{\boldsymbol{\theta}})$ is the *Hessian matrix* of ℓ evaluated at $\hat{\boldsymbol{\theta}}$ which by definition is the matrix of partial derivatives of ℓ at $\hat{\boldsymbol{\theta}}$. Notice that $\nabla \ell(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and that $\mathbf{H}_\ell(\hat{\boldsymbol{\theta}}) = -\mathcal{I}(\boldsymbol{\theta}_0) + o_p(1)$, therefore

$$2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)] = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1) \xrightarrow{d} \chi^2_p$$

□

Appendix B

R Code and Outputs

R code B.1: Complete output of lineal model of `medv` regressed on the predictors in `BostonHousing` dataset

```
Call:
glm(formula = medv ~ ., family = gaussian(), data = BostonHousing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-15.595   -2.730   -0.518    1.777   26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas1        2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
```

```
ptratio      -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
b             9.312e-03  2.686e-03   3.467 0.000573 ***
lstat        -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
      .    0.1          1

(Dispersion parameter for gaussian family taken to be 22.51785)

    Null deviance: 42716  on 505  degrees of freedom
Residual deviance: 11079  on 492  degrees of freedom
AIC: 3027.6

Number of Fisher Scoring iterations: 2
```

R code B.2: Complete output of logistic model of `Category` on the predictors in `HepatitisC` dataset

```
Call:
glm(formula = Category ~ ., family = binomial(), data = HepatitisC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9143  -0.1860  -0.0906  -0.0402   3.8509

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.445193   3.928088  -3.677 0.000236 ***
Age          -0.006511   0.024304  -0.268 0.788772
Sex           1.344974   0.608607   2.210 0.027111 *
```

```
ALB          -0.122350   0.063592  -1.924 0.054355 .
ALP          -0.070402   0.012666  -5.558 2.72e-08 ***
ALT          -0.020408   0.009422  -2.166 0.030317 *
AST           0.095025   0.019294   4.925 8.43e-07 ***
BIL           0.086677   0.028910   2.998 0.002716 **
CHE           0.125628   0.121698   1.032 0.301933
CHOL         -0.690593   0.258862  -2.668 0.007635 **
CREA          0.024286   0.005208   4.663 3.11e-06 ***
GGT           0.030608   0.006211   4.928 8.30e-07 ***
PROT          0.225520   0.058047   3.885 0.000102 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
      .    0.1          1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 456.08  on 614  degrees of freedom
Residual deviance: 128.44  on 602  degrees of freedom
AIC: 154.44

Number of Fisher Scoring iterations: 8
```

R code B.3: Complete output of Poisson model of `visits` on the predictors in `DoctorVisits` dataset

```
Call:
glm(formula = visits ~ ., family = poisson(), data = DoctorVisits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9502  -0.6858  -0.5747  -0.4852   5.7055

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.097821   0.101554 -20.657  < 2e-16 ***
gender       0.156490   0.056139   2.788  0.00531 **
age          0.279123   0.165981   1.682  0.09264 .
income      -0.187416   0.085478  -2.193  0.02834 *
illness      0.186156   0.018263  10.193  < 2e-16 ***
reduced      0.126690   0.005031  25.184  < 2e-16 ***
health       0.030683   0.010074   3.046  0.00232 **
private      0.126498   0.071552   1.768  0.07707 .
freepoor    -0.438462   0.179799  -2.439  0.01474 *
freerepat    0.083640   0.092070   0.908  0.36365
nchronic     0.117300   0.066545   1.763  0.07795 .
lchronic     0.150717   0.082260   1.832  0.06692 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
      .    0.1          1

(Dispersion parameter for poisson family taken to be 1)

```

```
    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 4380.1  on 5178  degrees of freedom
AIC: 6735.7

Number of Fisher Scoring iterations: 6
```

Curriculum Vitae

Denisse Urenda Castañeda was born in February 04, 1994 in Ciudad Juárez, México. She graduated from Universidad Autónoma de Ciudad Juárez in 2017 with a bachelor's degree in Mathematics. In fall of 2017 she entered to The University of Texas at El Paso to pursue a master's degree in Mathematics. For the rest of the master's degree she worked as professor in Universidad Autónoma de Ciudad Juárez, as a Teaching Assistant, and as a tutor in the Math Tutoring Center at UTEP.

In the fall of 2020, she started a new graduate program at The University of Texas at El Paso to pursue a master's in statistics. She started working as Research Assistant and she continues working as professor.

Phone number: +52 653 170 9160

E-mail: denisse.urenda@gmail.com

LinkedIn: https://www.linkedin.com/in/denisse-urenda-677736205/