UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Technical University of Catalonia

Barcelona School of Informatics

Detecting heterogeneity in Generalized Linear Modeling

Master thesis by:
Yaroslav Hernández Potiomkin

Advisor:
Dr. Tomàs Aluja Banet

Master in Innovation and Research in Informatics

Data Mining and Business Intelligence

February, 2017

Abstract

In classical model fitting techinques, such as traditional Multiple Linear Regression models (MLR) or Generalized Linear Models (GLM), the assumption is that the individuals come from homogeneous population. However, this condition may be not necessarily met, as there may be many factors that influence the behaviour of the individuals and therefore, biasing the model estimations.

For instance, let us consider that we want to study the salaries among a certain set of individuals that come from relatively defined professional sector. The first approach would be to collect all possible modeling variables and fit the model. But it may happen that this could lead us to inaccurate estimations, since the salaries can be driven differently according to gender, region, ethnicity, among others. These variables are called *segmentation* variables and their number may grow very fast. In this case arises a combinatorial problem giving many possibilities of how to group those individuals.

Our main goal in this work, is to go deeper in this kind of problems, and present an automatic solution to detect homogeneous segments among the heterogeneous population in the GLM context. The PATHMOX methodology is a powerful method proposed by Gastón (2009) [19] to automate the task of finding segments. The statistical tests needed to guide the PATHMOX algorithm and discover the constructs that differentiate those segments, are proposed by Lamberti (2015) [8].

First, we provide several solutions to detect heterogeneity, by means of moderating variables as in Covariance Analysis or by means of comparison of coefficients using parametric or non-parametric approaches, in section 2. Additionally, we present the method to characterize classes or continuous response by taking into account only segmentation variables in section 4. Then, we concentrate on the Generalized Linear Modeling context to define the automatic heterogeneity detection method. Then, we accurately present all the needed hypothesis test procedures in section 3. Finally, we also carry out a quite extensive simulation studies and a real problem application in sections 6 and 7, respectively.

Contents

1 Introduction

The heterogeneity problem arises in many statistical applications. Usually it is known as *population heterogeneity*, where the set of samples come from the mixtures of (for instance normals) distributions [10], each of them forming a cluster. But there is also another perspective which consists in considering the *model heterogeneity* where the difference is based on structural relationships between variables leading to different models. That is, in the first case we notice *descriptive* differences between groups of individuals and in the latter we aim to detect *behavioral* differences wrt some chosen statistical model. In this work we concentrate on the second approach.

The variables that introduce heterogeneous behaviour between the individuals in the population are called *segmentation* variables and the set of variables which are systematically related with the dependent variable y (response) are called *predictor* variables. The segmentation variables can be socio-demographic (age, gender, family size, occupation, education), geographic (world region, metropolitan area, climate), psychographic (lifestyle, personality) and behavioral (kind of profile: compulsive, reincidence, ease-of-use, fashion influenced).

According to the a priori knowledge, the sources of heterogeneity may be categorized as follows

1. Assignable (we know which are the segmentation variables and we have them)
   (a) Known segments (we know the partitions a priori)
   (b) Unknown segments (unknown partitions; number may be exponential)
2. Non-assignable (we do not have the segmentation variables)

In assignable case (1) we are able to partition the data into segments, which sometimes are known ($1.a$) but sometimes not ($1.b$). In the latter ($1.b$), we have to figure out the criterion to find the best split into segments.

In a non-assignable case (2) the partitions can not be defined beforehand. So, we know about the existence of different segments and for that we use the modeling variables.

In this work we concentrate on assignable heterogeneity with unknown segments.

2 Modeling heterogeneity

In this section we present several techniques to deal with heterogeneity. Even though these methods can not provide an automatic heterogeneity detection, they form a good introduction to the problem we aim to face and serve as a baseline.

2.1 Moderating variables

In a model with continuous predictors we can introduce segmentation or moderating variables, categorical by nature, leading to models of *Covariance Analysis* [2]. The general expression of such models is as follows

$$\boldsymbol{y} = L\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where L is composed of boolean columns which denote presence or absence of modality in the same way as in ANOVA analysis and columns that are continuous. We will stress the construction of this matrix in the following. The assumption of the residuals of the model is as usual

$$\varepsilon \sim N(0, \sigma^2) \tag{2}$$

Let's assume that we have one continuous predictor and one categorical variable with 3 modalities. Then, the most general model for individual k and modality i can be expressed as follows

$$y_{ik} = (\mu + \alpha_i) + (\eta + \beta_i)x_{ik} + \varepsilon_{ik} \tag{3}$$

where we can identify the intercept term μ, the principal effects (given by α_i coefficient), the common coefficient for the continuous modeling variable η and the *interaction* coefficient between the modality and the continuous predictor β_i. It is clear then, that the first $(3+1)$ columns, for this particular case of one segmentation variable and one continuous modeling variable, represent the one-way ANOVA; that is, one binary column for each modality and one column of ones representing the intercept term. The last 4 columns are just the element-wise product between the first $(3+1)$ columns and the modeling variable $\boldsymbol{x}$, and they represent the interaction effect between the factor and the predictor and the last column is the global effect of the predictor on response variable $\boldsymbol{y}$.

Therefore, the test whether the interaction effect is significant arises in a very natural way by defining a second reduced (or simplified) model from the first giving the corresponding null hypothesis (related to this simplified model) definition

$$H_0 : \begin{cases} \beta_i = 0 \text{ for } i \in \{1, 2, 3\} \\ \text{Any values for } \mu, \eta \text{ and } \alpha_i \end{cases}$$

The test statistic is defined as follows

$$F = \frac{(SSE_0 - SSE_1)/\big((n - p - 1) - (n - 2p)\big)}{SSE_1/(n - 2p)} \tag{4}$$

where the degrees of freedom of the numerator and denominator have been computed from the full rank models, in which we have one binary column of categorical variable less and so one interaction column between this modality and predictor. In this way, we are able to define simpler and simpler models, which would lead us to *chain* testing.

The details of derivation of the F-statistic from expression (4) will be provided in section 3.2.1.

2.2 Comparison of coefficients

Another classical way of treating heterogeneity is building a model for each segment

$$\begin{aligned} y_A &= \boldsymbol{\beta}'_A \boldsymbol{x} + \varepsilon_A \\ y_B &= \boldsymbol{\beta}'_B \boldsymbol{x} + \varepsilon_B \end{aligned} \tag{5}$$

and then performing a statistical test to find out whether $\boldsymbol{\beta}_A = \boldsymbol{\beta}_B$. The test can be conveniently performed in two ways

- Parametric test
  - Equal variances assumption
  - Non-equal variances
- Non-parametric tests

In the non-parametric approach [21] we have no distributional assumptions, which may be advantageous in several situations. For instance, when we have extreme outliers (the counterexample can be found also in [21]).

2.2.1 Parametric test

The parametric test consists in performing a t-test, where we can assume equal variances and use the pooled variance. Or alternatively, we can use a Welch's unequal variances t-test presented and with direct application in section 7.3. In any case, it

is convenient to keep in mind that this kind of tests is performed as one-at-a-time testing of the given coefficient and not globally.

2.2.2 Non parametric test based on ranks

The first non parametric test we present is the Wilcoxon - Mann - Whitney 2-sample rank sum test, which allows for testing the equality of central tendency of two distributions in case of unpaired data. The idea is to combine the two samples assuming that each observation is not coming from any sample in particular. This is done by sorting both samples jointly and then assigning a rank value (using midranks for ties). The interpretation of the test is whether the population medians of the two groups are the same, or more precisely, whether observations in one population are larger than in the other. The hypothesis test is defined then as follows

$$H_0 : P(x_1 > x_2) = \frac{1}{2}$$

and

$$H_1 : P(x_1 > x_2) \neq \frac{1}{2}$$

being x_1 and x_2 randomly chosen from first and second samples, respectively. The test statistic is written by

$$W = R - \frac{n_1(n_1 + 1)}{2} \tag{6}$$

where R is the sum of ranks in the first sample. Assuming the null hypothesis true we have $\mu_W = \frac{n_1 n_2}{2}$ and $\sigma_W = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$. Therefore, we have that the test statistic

$$z = \frac{W - \mu_W}{\sigma_W} \sim N(0, 1) \tag{7}$$

The *concordance probability* is written in the following equation

$$C = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2} \tag{8}$$

where $\bar{R}$ is the mean of ranks from sample one.

2.2.3 Non parametric permutation test

In the test we are going to present [23], the important assumption is the independence of observations. The permutation test that is based on random assignment, as proposed by Edgington (1987) and Good (2000), can be accomplished by

1. A test statistic is computed for the data (with original experimental labels).
2. The data are permuted repeatedly and randomly. The test statistic is computed for each permutation.
3. The obtained data permutations and the data with original labels constitute the reference set for determining significance.
4. The proportion of data permutations in the reference set that have test statistic values greater than or equal to (sometimes it may be less than or equal to) the value obtained with original labels is the p-value.

It is important to take into consideration that in order to use parametric statistical tables (t, F) is needed the random sampling assumption, that means that all possible samples of size n within some predefined population have the same probability of being drawn.

2.3 Global comparison of models

In this approach we aim to build a statistic that allows us to determine the presence of heterogeneity through hypothesis testing, where H_0 is representing the homogeneity, whereas the alternative H_1 denotes heterogeneous situation:

$$H_0 : \begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}'_A \\ \boldsymbol{x}'_B \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}$$

$$H_1 : \begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}'_A & 0 \\ 0 & \boldsymbol{x}'_B \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}$$

The test statistic involves comparing the sum of squares of the residuals of both models in the F-statistic. This way enables us to deal with large number of segments modifying appropriately the test. Moreover, we do not need to know about the heterogeneity generation mechanism; that is, the sources that originated the heterogeneity.

3 State of the art of Generalized Linear Models

In this section we briefly present Generalized Linear Models formulation. Then we take a closer look at classical hypothesis testing by means of F-statistic. We also present two kinds of GLM models, following a Gaussian family and Binomial family. We also discuss Multiple Linear Regression models in a separate section. Finally, we present methods for comparing two models, focusing on a global comparison and then a more specific one that aims to detect the constructs responsible for such a difference.

3.1 Generalized Linear Models

In this section, we will provide a brief introduction to the Generalized Linear Models (GLM).

The most important assumption in GLM, is the independence or uncorrelation of the observations [5]. The samples presenting autocorrelations, time-series relations are not treated by this kind of models. Regarding the notation, we use capital letters such as Y for describing a theoretical random variable and matrices (depending on the context), while the lower case such as y, for the values that takes this variable. Bold font is used for vectors, for instance $\boldsymbol{y}$.

A *linear predictor* is defined as follows

$$\eta = \boldsymbol{x}'\boldsymbol{\beta}$$

and then the *link function*, g, is introduced to relate the linear predictor with the expected value, μ, of the dependent variable

$$E[Y] = \mu = g^{-1}(\eta) \tag{9}$$

The probability model should come from an exponential family, that is to say, where the univariate probability density function of the dependent variable can be expressed in the following form

$$f_Y(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{10}$$

for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. θ is the so-called canonical parameter and in the Normal case it holds that $\theta = \mu$ and $a(\phi) = \frac{\sigma^2}{\omega}$ (in case when each observation is the mean of several independent readings for the same input value $\boldsymbol{x}$, ω is the number of these readings, and in most of the cases $\omega = 1$).

Moreover, it can be shown that the variance of Y depends on the canonical parameter θ (hence on the mean μ by definition) and the dispersion parameter ϕ

$$Var[Y] = \frac{\partial^2 b(\theta)}{\partial \theta^2} a(\phi) \tag{11}$$

For example, in classical linear regression we have that $E[Y] = \mu = \eta$ having an identity link function. It is important to notice, that in the case that the chosen link function is canonical it holds that

$$\theta = \eta \tag{12}$$

that is, the canonical parameter θ is the linear predictor η.

The estimation of the parameters is performed using the Maximum Likelihood Estimator (MLE). The log-likelihood function is conveniently expressed in terms of $\boldsymbol{\mu}$ (mean value) parameters instead of canonical parameters $\boldsymbol{\theta}$:

$$l(\boldsymbol{\mu}|Y = \boldsymbol{y}) = \sum_i \log f_i(y_i|\mu_i) \tag{13}$$

where f is the density function and $Y = \boldsymbol{y}$ is the data, where the random variable Y takes values $\boldsymbol{y}$.

In order to find the maximum-likelihood estimates, in GLM is used the *iterative weighted least squares*. A nice description and justification can be found in [5].

The goodness of fit criterion used is the *scaled deviance*:

$$D^*(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\theta}}) = 2\big(l(\boldsymbol{\mu}^*|\boldsymbol{y}) - l(\boldsymbol{\mu}|\boldsymbol{y})\big) = \sum_i 2\omega_i \frac{y_i(\hat{\theta}_i^* - \hat{\theta}_i) - b(\hat{\theta}_i^*) + b(\hat{\theta}_i)}{\phi} \tag{14}$$

where again $\boldsymbol{y}$ is the data, ω_i are in most cases 1 (only one observation for each predictor input value and not the mean of ω_i observations), ϕ in Normal case is just the σ^2, the $\hat{\theta}_i^*$ are the canonical parameters estimates by MLE, that is where the log-likelihood achieves the highest value (in Normal case they are just the observations y_i) and the $\hat{\theta}_i$ are the canonical parameters estimates by MLE under the model we want to check. Usually, the θ_i^* is considered as the saturated model and the θ_i is the reduced model.

It is important to notice that (14) holds when the ϕ dispersion parameter is the same for both models. And it has sense as the true irreducible variability is supposed to be the same (in bias-variance trade-off expression).

Moreover, it turns out that the scaled deviance, $D^*(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\theta}})$, coincides with the Likelihood Ratio expression that we will present in the section 3.4.2.

3.2 Case of continuous data

The response variable of continuous type is usually modeled by GLM models with identity link function. In this case the normality assumption is not considered but the independence of observations yes it is. Hence, the prediction is expressed as follows

$$\mu_i(\theta_i) = E[y_i] = g^{-1}(\eta_i) \tag{15}$$

and when *identity* canonical link function is chosen, we have that

$$g(\mu_i) = \mu_i = \eta_i = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \theta_i{}^1 \tag{16}$$

The exponential family form of the normal distribution (from Normal theory for continuous data) can be found in [5]. This form allows to derive the mean and the variance of Y response variable using the following relations

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{17}$$

and

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = 0 \tag{18}$$

The log-likelihood function form with known σ can be written as

$$l(\mu_i|y_i) = -\frac{1}{2}\log(2\pi\sigma^2) - (y_i - \mu_i)^2/(2\sigma^2) \tag{19}$$

Therefore, the scaled deviance can be derived as follows

$$D^*(y_i|\hat{\boldsymbol{\theta}}_i{}^*, \hat{\boldsymbol{\theta}}_i) = 2\big\{l(y_i|y_i) - l(\mu_i|y_i)\big\} = (y_i - \mu_i)^2/\sigma^2 \tag{20}$$

From the expression (20) it is clear that deviance for Normal response data is the residual sum of squares.

[1] When the link is not canonical, the last equality does not hold.

3.2.1 Derivation of hypothesis tests for classical regression models

If we consider a generic linear model formulation

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{21}$$

and the same model expressed after introducing an estimator function for the parameters vector $\boldsymbol{\beta}$

$$\boldsymbol{y} = X\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is our vector of data of dimension $n \times 1$, X is our design matrix of dimension $n \times p$ where each column represents one independent variable or predictor and the first is the vector of 1's for the intercept. $\hat{\boldsymbol{\beta}}$ is the vector $(p \times 1)$ of parameter estimations and finally, $\boldsymbol{\epsilon}$ is the vector $(n \times 1)$ of residuals and it is orthogonal to the space $\mathbb{R}_X \in \mathbb{R}^n$ generated by the predictors.

It is important to notice the relation between $\boldsymbol{\varepsilon}$ and $\boldsymbol{\epsilon}$, where the former is the population errors vector whereas the second one is the estimated one through the model. Usually, they are very near and formally

$$\boldsymbol{\epsilon} = Q\boldsymbol{\varepsilon}$$

where $Q = I_n - H$ (being H a so-called hat matrix). The matrix Q is the orthogonal projector wrt the orthogonal space $\mathbb{R}_X$ (space generated by X; that is, all the possible linear combinations of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$) over the vector of residuals $\boldsymbol{\epsilon}$. So, for instance, if we take the true error vector $\boldsymbol{\varepsilon}$

$$\begin{aligned} Q\boldsymbol{\varepsilon} &= (I_n - X(X'X)^{-1}X')\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon} - X(X'X)^{-1}X'\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon} - (X(X'X)^{-1}X'\boldsymbol{y} - X(X'X)^{-1}X'X\boldsymbol{\beta}) \\ &= \boldsymbol{\varepsilon} - (X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}) \\ &= \boldsymbol{y} - X\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{\epsilon} \end{aligned} \tag{22}$$

And similarly we have that

$$Q\boldsymbol{y} = (I_n - X(X'X)^{-1}X')\boldsymbol{y} = y - X\hat{\boldsymbol{\beta}} = \boldsymbol{\epsilon} \tag{23}$$

The following Lemmas and their corresponding proofs can be found in [2, 8]. In this work we tried to extend some parts of the proofs for ease-of-read.

Lemma 1. *Let $\boldsymbol{\varepsilon}$ be a normally distributed vector in $\mathbb{R}^n$ with $E[\boldsymbol{\varepsilon}] = 0$ and $Var[\boldsymbol{\varepsilon}] = \sigma^2 I_n$*

1. *Let Q be an $n \times n$ symmetric and idempotent[2] matrix. Then, the quadratic form $\frac{\boldsymbol{\varepsilon}' Q \boldsymbol{\varepsilon}}{\sigma^2}$ follows a χ^2 distribution with ν d.o.f. (where ν is the rank of the matrix Q).*
2. *Let L be a matrix s.t. $LQ = 0$. Then, the vectors $L\boldsymbol{\varepsilon}$ and $Q\boldsymbol{\varepsilon}$ follow an independent normal distributions. In particular, the vector $L\boldsymbol{\varepsilon}$ and the variable $\frac{\boldsymbol{\varepsilon}' Q \boldsymbol{\varepsilon}}{\sigma^2}$ are independent.*

Proof. Given that the matrix Q is symmetric, there exists Λ and H s.t. $Q = H'\Lambda H$ that can be easily checked through a SVD [4], where H is the matrix where the rows are orthonormal eigenvectors of Q ($H'H = HH' = I$), and Λ is the diagonal matrix of the eigenvalues. Moreover, as Q is an idempotent matrix, the eigenvalues are either 0 or 1 (it can be seen using the th. *EOMP* - Eigenvalues Of Matrix Powers [3]). If we consider $\boldsymbol{v} = H\boldsymbol{\varepsilon}$, then the linear transformation of $\boldsymbol{\varepsilon}$, $\boldsymbol{v}$ follows a normal distribution

$$\boldsymbol{v} \sim N(E[\boldsymbol{v}], Var[\boldsymbol{v}])$$

with $E[\boldsymbol{v}] = HE[\boldsymbol{\varepsilon}] = 0$ and $Var[\boldsymbol{v}] = E[H\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'H'] \overset{linearity}{=} HE[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']H' = H\sigma^2 I H' = \sigma^2 I$

Hence, the reduced component $\boldsymbol{v}/\sigma$ follows a standardized and independent normal distribution. We can rewrite the quadratic form $\boldsymbol{\varepsilon}' Q \boldsymbol{\varepsilon}/\sigma^2$ as

$$\boldsymbol{\varepsilon}' Q \boldsymbol{\varepsilon}/\sigma^2 = \boldsymbol{\varepsilon}' H' \Lambda H \boldsymbol{\varepsilon}/\sigma^2 = \boldsymbol{v}' \Lambda \boldsymbol{v}/\sigma^2 = \sum_i \boldsymbol{v}_i^2/\sigma^2$$

It is important to notice that the last sum includes all terms of the eigenvalues equal to 1 since the other values are zero, having that $tr(Q) = \sum_i \lambda_i = \sum_i \mathbb{I}_{\lambda_i == 0} = rank(Q) = d.o.f.(\boldsymbol{\varepsilon}' Q \boldsymbol{\varepsilon}/\sigma^2)$. Thus, since χ^2 with p d.o.f. is defined as the sum of squares of p standardized normal variables, the first point of the lemma is verified.

Regarding the second point of the lemma, it can be seen as follows. The vectors $L\boldsymbol{\varepsilon}$ and $Q\boldsymbol{\varepsilon}$ are two linear transformations of $\boldsymbol{\varepsilon}$ following a normal distribution. Under this assumption, if the covariance $Cov(L\boldsymbol{\varepsilon}, Q\boldsymbol{\varepsilon})$ is zero these vectors turn out to be independent

$$Cov(L\boldsymbol{\varepsilon}, Q\boldsymbol{\varepsilon}) = E[(L\boldsymbol{\varepsilon})(Q\boldsymbol{\varepsilon})'] \overset{linearity}{=} \sigma^2 LQ = 0$$

The second point of the lemma is verified. Then, $\boldsymbol{\varepsilon}' Q \boldsymbol{\varepsilon} = (Q\boldsymbol{\varepsilon})'(Q\boldsymbol{\varepsilon})$ is independent of $L\boldsymbol{\varepsilon}$.

[2]The matrix M is idempotent iff $MM = M$

□

Lemma 2. *Let $\boldsymbol{\varepsilon}$ be a vector normally distributed in $\mathbb{R}^n$ with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $Var[\boldsymbol{\varepsilon}] = \sigma^2 I$. Given two matrices, X and X_0, where X_0 is defined as $X_0 = XA$ (for any matrix A), we can define the following matrices Q and Q_0 as follows*

$$Q = I_n - X(X'X)^{-1}X'$$
$$Q_0 = I_n - X_0(X_0'X_0)^{-1}X_0'$$

It can be shown that the quotient

$$F = \frac{(\boldsymbol{\varepsilon}'Q_0\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'Q\boldsymbol{\varepsilon})/(\nu_0 - \nu)}{\boldsymbol{\varepsilon}'Q\boldsymbol{\varepsilon}/\nu} \tag{24}$$

follows an F distribution with $\nu_0 - \nu$ and ν d.o.f., where

- *ν is the rank of Q*
- *ν_0 is the rank of Q_0*

Proof. A geometric representation (figure 1) allows to see the intuition of the proof. Let $\mathbb{R}_{X_0}$ be a sub-space generated from the columns of X_0, the distance between $\boldsymbol{y}$ and $\mathbb{R}_X$ to be the longitude of $\boldsymbol{\epsilon} = Q\boldsymbol{\varepsilon}$ (the vector of sample errors)

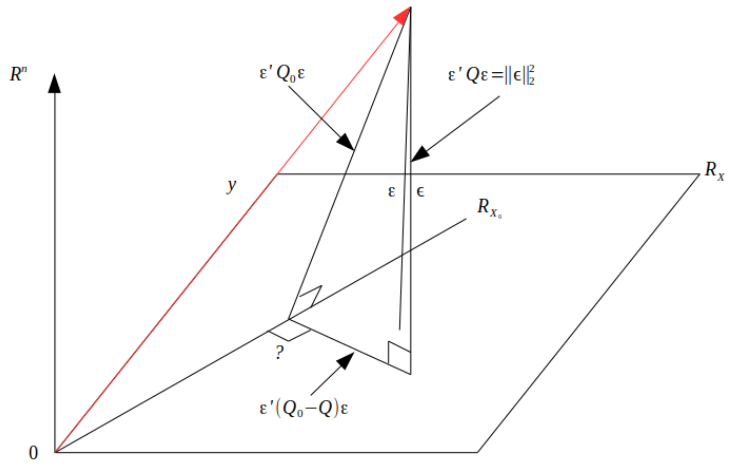

Figure 1: Projection of $\boldsymbol{y}$ onto $\mathbb{R}_X$ and $\mathbb{R}_{X_0}$.

$$\sum \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \boldsymbol{\varepsilon}'Q\boldsymbol{\varepsilon} \tag{25}$$

On the other hand, the relationship $X_0 = XA$ means that the space $\mathbb{R}_{X_0}$ is generated by the columns of X_0 and it is supposed to be strictly a sub-space of $\mathbb{R}_X$. The

orthogonal projections are represented in figure 1, where $\mathbb{R}_X$ is the plane and $\mathbb{R}_{X_0}$ is the line (with one dimension less).

From the lemma 1, $\boldsymbol{\varepsilon}'Q\boldsymbol{\varepsilon}/\sigma^2$ follows a χ^2 with $rank(Q)$ d.o.f. The numerator term becomes

$$\boldsymbol{\varepsilon}'Q_0\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'Q\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'(Q_0 - Q)\boldsymbol{\varepsilon}$$

Figure 1 and the theorem of the three perpendiculars (see Appendix A.2 for the intuition behind this theorem) suggests that $Q_0 - Q$ is an orthogonal projector. Given that both Q and Q_0 are symmetric, $Q_0 - Q$ is symmetric too. The idempotence still should be verified

$$(Q_0 - Q)^2 = Q_0 - Q_0Q - QQ_0 + Q \stackrel{?}{=} Q_0 - Q$$

Direct calculation shows that

$$QQ_0 = (I_n - X(X'X)^{-1}X')(I_n - X_0(X_0'X_0)^{-1}X_0') = Q$$

Therefore, $QQ_0 = Q_0Q = Q$ and $(Q_0 - Q)^2 = Q_0 - Q$, the matrix is idempotent and its rank is

$$rank(Q_0 - Q) = tr(Q_0 - Q)$$

The $tr(Q_0 - Q)$ can be seen using the additive ($tr(A+B) = tr(A)+tr(B)$) and cyclic ($tr(AB) = tr(BA)$) properties of the traces. Thus, $tr(Q) = tr(I_n) - tr(I_{2p}) = n - 2p$, $tr(Q_0) = n - p$ and

$$rank(Q_0 - Q) = tr(Q_0 - Q) = p$$

where p is the number of parameters including the intercept. Furthermore, the intercept should be always considered even if the data has been previously centered as in the posteriors splits the mean does not necessarily remains zero [1] and the intercept could be of high importance. Consider, for instance, the salaries among men and women. It may happen, that the only predictor *years of studies* infuences equally for both, but for some reasons one of the genders perceives greater salary. So, the models would be just parallel, but the intercept clearly different.

Fisher distribution is defined as the ratio between two χ^2 independent variables divided by their d.o.f. and using lemma 1, the numerator of equation (24) follows a χ^2 distribution with $tr(Q_0) - tr(Q)$ d.o.f. Now we only require to verify if the numerator and denominator of equation (24) are independent. For this purpose, we should only check if their covariance is equal to zero

$$Cov((Q_0 - Q)\boldsymbol{\varepsilon}, Q\boldsymbol{\varepsilon}) = E[(Q_0 - Q)\boldsymbol{\varepsilon}(Q\boldsymbol{\varepsilon})'] = (Q_0 - Q)E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']Q = \sigma^2(Q_0 - Q)Q = 0$$

The lemma is proven. □

3.2.2 Multiple linear regression

A *multivariate linear regression model* [9] assumes the response to be a linear function of predictors

$$E[Y|X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{26}$$

However, it is very important to notice that the linear models are understood to be linear wrt the coefficients, the betas, disregarding whether the predictors enter in a non-linear way (for instance, $E[Y|\boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$). Although being so obvious, but in many cases it may lead to confusion when treating with a highly complex models.

Returning to the MLR model, the assumptions about the error term, $\boldsymbol{\varepsilon}$, are the following

- Independence
- Normality
- Homoscedasticity (σ^2)

The solution consists in a p-dimensional hyperplane in a $p + 1$ dimensional space. The estimated coefficients are found by minimizing the least squares errors function or maximizing the log-likelihood function giving the following expression

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y} \tag{27}$$

Therefore, we can observe the following properties

1. The estimated coefficients are linear functions of the responses
2. The estimator functions provide an unbiased estimations $E[\hat{\boldsymbol{\beta}}] = (X'X)^{-1}X'E[\boldsymbol{y}] = (X'X)^{-1}X'X\boldsymbol{\beta} = \boldsymbol{\beta}$
3. $\hat{\boldsymbol{\beta}}$ are normally distributed (when $\boldsymbol{y}$ do so)

The sample variance of $\hat{\beta}_i$ is given as follows

$$s^2(\hat{\beta}_i) = s^2(X'X)^{-1}_{ii} \tag{28}$$

where s^2 is an unbiased estimator of the σ^2. The standard error is used to test the significance of the coefficient using a t-test statistic.

The unseen response is predicted as follows

$$\tilde{\mu} = \tilde{y} = E[Y|\tilde{\boldsymbol{x}}] = \tilde{\boldsymbol{x}}'\hat{\boldsymbol{\beta}} \tag{29}$$

The corresponding variance is computed as follows

$$Var(\tilde{y}) = Var(\tilde{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}) = \tilde{\boldsymbol{x}}'Var(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{x}} = s^2\tilde{\boldsymbol{x}}'(X'X)^{-1}\tilde{\boldsymbol{x}} \tag{30}$$

It is used in the calculation of confidence interval for the mean $\tilde{\mu} = \tilde{y} = E[Y|\tilde{\boldsymbol{x}}]$

$$\tilde{\boldsymbol{x}}'\boldsymbol{\beta} \pm t_{\alpha/2}s\sqrt{\tilde{\boldsymbol{x}}'(X'X)^{-1}\tilde{\boldsymbol{x}}} \tag{31}$$

and in the prediction interval for the individual response $\tilde{y}_*$

$$\tilde{\boldsymbol{x}}'\boldsymbol{\beta} \pm t_{\alpha/2}s\sqrt{1 + \tilde{\boldsymbol{x}}'(X'X)^{-1}\tilde{\boldsymbol{x}}} \tag{32}$$

The goodness-of-fit criterion (GOF) R^2, coefficient of multiple determination, measures the proportion of variance explained by the model

$$R^2 = \frac{\text{Regression sum of squares}}{\text{Sum of squares of observations}} = \frac{SS_R}{s^2_{\boldsymbol{y}}} = \frac{(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})'(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})}{(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})'(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) + \boldsymbol{\epsilon}'\epsilon} \tag{33}$$

However, it is very convenient to use an adjusted R^2 which takes into account the number of predictors

$$R^2_{adj} \overset{\text{def.}}{=} 1 - \frac{s^2_{\boldsymbol{\epsilon}}/(n - \#params)}{s^2_{\boldsymbol{y}}/(n-1)} = 1 - \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})'(\boldsymbol{y} - \hat{\boldsymbol{y}})/(n - \#params)}{s^2_{\boldsymbol{y}}/(n-1)} \tag{34}$$

which can also be rewritten as follows

$$R^2_{adj} = 1 - \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/(n - \#params)}{\{(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})'(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) + \boldsymbol{\epsilon}'\boldsymbol{\epsilon}\}/(n-1)} \tag{35}$$

where $\#params$ is the number of estimated parameters (including the intercept).

Another adequacy measure is the *Prediction Error Sum of Squares* (PRESS) proposed by Allen (1971, 1974) [13]. It provides prediction capacity of the model based on leave-one-out sampling. Given n observations, at each iteration i ($i \in 1 \dots n$), we build a regression model on the data (X_{-i}, $\boldsymbol{y}_{-i}$); all observations except i, and

evaluate it on validation data $(\boldsymbol{x}_i,\, y_i)$. These n models provide a response $\hat{y}_{(i)}$ and the corresponding PRESS residual $e_{(i)} = y_i - \hat{y}_i$. The PRESS statistic is then defined in the following.

$$PRESS = \sum_{i}^{n} e_{(i)}^2 \tag{36}$$

Moreover, it can be calculated by using one model fitted to all data at a time having that (see the proof in Appendix A.3)

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \tag{37}$$

where h_{ii} are the diagonal elements of the hat matrix. The PRESS residuals are the weighted ordinary residuals, where the weights are related to the leverage, h_{ii}, of the observations. The higher the value of h_{ii} the more influent is the point i. The variance of the PRESS residuals is the following [13 Chap. 4.9.3]

$$Var\big(e_{(i)}\big) = Var\left(\frac{e_i}{1 - h_{ii}}\right) = \frac{\sigma^2}{1 - h_{ii}} \tag{38}$$

PRESS can be used to compute an approximate R^2 for prediction

$$R^2_{\text{pred}} = 1 - \frac{PRESS}{s^2_{\boldsymbol{y}}} \tag{39}$$

that is the proportion of variability explained by the model in predicting new observations.

In order to choose the relevant variables, it can be applied a *forward selection* or *backward elimination* which use an F-statistic.

For categorical variables it can be introduced a set of dummy variables, that indicate the presence or the absence of the category (level) of the variable.

A more detailed description of this kind of models is provided in [9].

3.3 Case of binary data

There are two approaches considering binary data. The first one is using ungrouped data where each observation $y_i \sim B(m_i = 1, \pi_i)$ or as grouped data where $y_i \sim B(m_i > 1, \pi_i)$. In the latter case, the responses are considered as follows $y_1/m_1, \ldots, y_n/m_n$. The two approaches are useful to consider for the asymptotic approximations.

Hence, the prediction is expressed as follows

$$\mu_i(\theta_i) = E[y_i]/m_i = \pi_i = g^{-1}(\eta_i) \tag{40}$$

and when *logit* canonical link function is chosen, we have that

$$g(\mu_i) = g(\pi_i) = \eta_i = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \theta_i{}^3 \tag{41}$$

which is the linear predictor that has infinite support; that is, the inverse of link function has the support on entire real axis, so it can be evaluated for any real input value. In order to obtain the expected value (of the response variable), which is the probability of $y_i = 1$ given that $m_i = 1$, is applied the inverse of the link function as follows

$$\mu_i = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}} \tag{42}$$

The interpretation in such kind of models is very natural. For instance, a unit increment in x_i increments a *log of the odds* in β_i units.

The interpretation in the probabilistic scale, that is for $\mu_i = \pi_i$ is more difficult and depends on how far is the value of π_i from 0.5.

The Binomial distribution in the exponential form can be derived applying logarithms and exponential function to the original discrete Binomial probability mass function

$$\begin{aligned} p(y_i|\pi_i) &= \exp\left\{ \log(\binom{m_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{m_i - y_i}) \right\} \\ &= \exp\left\{ \log\binom{m_i}{y_i} + \log \pi_i^{y_i} + \log(1-\pi_i)^{m_i - y_i} \right\} \\ &= \binom{m_i}{y_i} \exp\left\{ y_i \log\frac{\pi_i}{1-\pi_i} + m_i \log(1-\pi_i) \right\} \end{aligned} \tag{43}$$

The log-likelihood function is defined as follows

[3] When the link is not canonical, the last equality does not hold.

$$
\begin{aligned}
l(\boldsymbol{\mu} = \boldsymbol{\pi}|\boldsymbol{y}) &= \log \prod_i p(y_i|\pi_i) \\
&= \log \prod_i \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} \\
&= \sum_i \big(y_i \log \pi_i + (m_i - y_i) \log(1-\pi_i) \big) \\
&= \sum_i \big(y_i \log \frac{\pi_i}{1-\pi_i} + m_i \log(1-\pi_i) \big)
\end{aligned}
\tag{44}
$$

where the term $\sum \log \binom{m_i}{y_i}$ has been omitted.

Using the log-likelihood expression above we can define the *scaled deviance*

$$
D^*(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\theta}}) = 2l(\boldsymbol{\pi}^*|\boldsymbol{y}) - 2l(\boldsymbol{\pi}|\boldsymbol{y}) = 2 \sum_i \left\{ y_i \log \frac{y_i}{\mu_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i - \mu_i} \right\} \tag{45}
$$

where in the case of ungrouped data we have that $\mu_i = \pi_i$ and $m_i = 1$ and $y_i \in \{0, 1\}$. If we substitute the μ_i values by their corresponding estimates $\hat{\mu}_i$, we would obtain the scaled deviance for the current model. The asymptotic approximation of this statistic is presented in the following sections.

3.3.1 A matter of sampling

A very interesting property of the logit link function in binary data is that either performing a *prospective* or *retrospective* sampling of the data to train our model, only the intercept term in the linear predictor would be different. This property indicates the appropriateness of this link function for any chosen sampling criterion.

This is very convenient to consider when we treat with highly unbalanced problems and forced to perform an undersampling to obtain more useful and unbiased models.

For instance, when we have ungrouped data and consider an undersampling approach for the convenience of our problem, which is very common in clinical studies when the number of negatives is much smaller than of those positives. A dummy variable Z represents whether an individual is sampled from the dataset. Then, the proportions of the sampling is defined beforehand depending on the data. Usually, they are set in such a way that the classes are balanced, so the $p(Z = 1|Y)$ defines these proportions. Using a Bayes's theorem we have

$$
p(Y = 1|Z = 1, \boldsymbol{x}) = \frac{p(Z = 1|Y = 1, \boldsymbol{x}) p(Y = 1|\boldsymbol{x})}{\sum_{y \in \{0,1\}} p(Z = 1|Y = y, \boldsymbol{x}) p(Y = y|\boldsymbol{x})} \tag{46}
$$

Then, using 42 we have

$$p(Y = 1|Z = 1, \boldsymbol{x}) = \frac{\exp\{\beta_0^* + \sum_{i=1}^{p} \beta_i x_i\}}{1 + \exp\{\beta_0^* + \sum_{i=1}^{p} \beta_i x_i\}} \tag{47}$$

where $\beta_0^* = \beta_0 + \log \frac{p(Z=1|Y=1)}{p(Z=1|Y=0)}$, that in case of balanced classes, when no sampling is required, we obtain the expression as in 42.

3.4 Comparison of two models

In this section we present the comparison of two models in two different scenarios: Multiple Linear Regression and classification with Binomial family (Logistic Regression or GLM with *logit* link function). The first case is covered by means of F-statistic and the second one uses Λ-statistic as it appears to be a widely accepted discrepancy of fit measure in GLM.

Additionally to the global F/Λ-statistic, we also specify the coefficient statistic in order to determine the responsible coefficients for that difference, if any, hopefully detected by the global statistic. This adds an additional power to the PATHMOX method, presented further on in section 5.

3.4.1 Testing two models in Multiple Linear Regression

As it was presented in section 3.2.1, in order to compare two MLR models we can use the F-statistic approach.

Let's assume we are splitting our population at each step into two segments by some segmentation variable already predefined in the design phase. At the time instant t of the process we have two segments

$$\left[X_A^{(t)}, \boldsymbol{y}_A^{(t)}\right] \quad \text{and} \quad \left[X_B^{(t)}, \boldsymbol{y}_B^{(t)}\right]$$

where for the further discussion we omit the time index to simplify the notation, but in reality it plays an important role as the dataset is reduced at each step and therefore becomes a natural identifier of the whole process.

Moreover, it is important to take into account that after previous $t-1$ splits we do not know whether our data is centered or not, even if at time step $t = 1$ it was. Hence, it is convenient to include the **intercept** term in the model.

3.4.1.1 Global F-test in MLR

In Multiple Linear Models, in order to determine whether the difference of the two models is not due to a randomness, the following hypothesis test allows to see whether *the coefficients of two segments A and B are equal at the same time*:

$$H_0 : \underset{n\times 1}{\begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix}} = \underset{n\times p}{\begin{bmatrix} X_A \\ X_B \end{bmatrix}} \cdot \underset{p\times 1}{[\boldsymbol{\beta}]} + \underset{n\times 1}{\begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}} \tag{48}$$

and the alternative hypothesis is as follows

$$H_1 : \underset{n\times 1}{\begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix}} = \underset{n\times 2p}{\begin{bmatrix} X_A & 0 \\ 0 & X_B \end{bmatrix}} \cdot \underset{2p\times 1}{\begin{bmatrix} \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}} \tag{49}$$

In order to be able to apply the lemma 2 from section 3.2.1 and to calculate the F-statistic we define the equation

$$X_0 = XA$$

as follows

$$\underset{n\times p}{\begin{bmatrix} X_A \\ X_B \end{bmatrix}} = \underset{n\times 2p}{\begin{bmatrix} X_A & 0 \\ 0 & X_B \end{bmatrix}} \underset{2p\times p}{\begin{bmatrix} I_{p\times p} \\ I_{p\times p} \end{bmatrix}}$$

where $I_{p\times p}$ is the identity matrix of order p. As the matrices fulfill the conditions asked in lemmas 1 and 2, we can build the F-statistic

$$F = \frac{\left(SS_{H_0} - SS_{H_1}\right)/p}{SS_{H_1}/(n-2p)} \tag{50}$$

with p d.o.f. in the numerator and $n-2p$ d.o.f. in the denominator.

3.4.1.2 F-coefficient test in MLR

A test based on F-statistic to see which coefficient is responsible for the difference between two models can be expressed again in terms of lemmas 1 and 2. The test hypothesis is *the coefficient i out of p is equal disregarding the segment, while the other $p-1$ coefficients can vary freely.*

Without loss of generality let us assume we want to test the β_1 coefficient. Hence, the null hypothesis will be expressed as follows

$$H_0 : \underset{n\times 1}{\begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix}} = \underset{n\times(2p-1)}{\begin{bmatrix} \mathbf{1}_A & \boldsymbol{x}_{A1} & \boldsymbol{x}_{A2} & 0 & 0 \\ 0 & \boldsymbol{x}_{B1} & 0 & \mathbf{1}_B & \boldsymbol{x}_{B2} \end{bmatrix}} \cdot \underset{(2p-1)\times 1}{\begin{bmatrix} \beta_{A0} \\ \beta_1 \\ \beta_{A2} \\ \beta_{B0} \\ \beta_{B2} \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}} \tag{51}$$

and the alternative hypothesis

$$H_1 : \underset{n\times 1}{\begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix}} = \underset{n\times 2p}{\begin{bmatrix} \mathbf{1}_A & \boldsymbol{x}_{A1} & \boldsymbol{x}_{A2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1}_B & \boldsymbol{x}_{B1} & \boldsymbol{x}_{B2} \end{bmatrix}} \cdot \underset{2p\times 1}{\begin{bmatrix} \beta_{A0} \\ \beta_{A1} \\ \beta_{A2} \\ \beta_{B0} \\ \beta_{B1} \\ \beta_{B2} \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}} \tag{52}$$

Again we need to fulfill the conditions by expressing the following equation

$$X_0 = XA$$

as follows

$$\underset{n\times(2p-1)}{\begin{bmatrix} \mathbf{1}_A & \boldsymbol{x}_{A1} & \boldsymbol{x}_{A2} & 0 & 0 \\ 0 & \boldsymbol{x}_{B1} & 0 & \mathbf{1}_B & \boldsymbol{x}_{B2} \end{bmatrix}} = \underset{n\times 2p}{\begin{bmatrix} \mathbf{1}_A & \boldsymbol{x}_{A1} & \boldsymbol{x}_{A2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1}_B & \boldsymbol{x}_{B1} & \boldsymbol{x}_{B2} \end{bmatrix}} \underset{2p\times(2p-1)}{\left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right]} \tag{53}$$

where in the matrix A the upper zone remains always the same, but in the lower zone at some point the coefficient $i-p$ from segment B (row i) is aligned with the coefficient i from segment A by placing 1 on the column $i-p$. The coefficients $j-p$ from segment B such that $j > i$ are shifted to the columns $j-1$.

At this point, the F-statistic can be computed as follows

$$F = \frac{\left(SS_{H_0} - SS_{H_1}\right)/1}{SS_{H_1}/(n-2p)} \tag{54}$$

with 1 and $n-2p$ d.o.f. in the numerator and denominator, respectively.

To be noticed, the denominator is the same as in (50).

3.4.1.3 An alternative F-coefficient test in MLR

In this section we present an alternative way of testing for the responsible coefficient. In section 3.4.1.2 it was exposed a classical approach performing a one-at-a-time test,

by allowing in H_0 vary freely all coefficients except the one under testing. In the alternative hypothesis we have that all the coefficients are freely adjusted for each segment.

Conversely, in this approach that we are going to present, the idea is to hold the same coefficients in both segments for the null hypothesis and allowing to vary freely over segments the only one coefficient in the alternative hypothesis. Assuming that we aim to test whether the predictor X_1 is responsible for the split, this new test can be written in the following way

$$H_0 : \underset{n\times 1}{\begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix}} = \underset{n\times p}{\begin{bmatrix} X_A \\ X_B \end{bmatrix}} \cdot \underset{p\times 1}{[\boldsymbol{\beta}]} + \underset{n\times 1}{\begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}} \tag{55}$$

and the alternative hypothesis is as follows

$$H_1 : \underset{n\times 1}{\begin{bmatrix} \boldsymbol{y}_A \\ \boldsymbol{y}_B \end{bmatrix}} = \underset{n\times(p+1)}{\begin{bmatrix} \mathbf{1}_A & \boldsymbol{x}_{A1} & 0 & \boldsymbol{x}_{A2} \\ \mathbf{1}_B & 0 & \boldsymbol{x}_{B1} & \boldsymbol{x}_{B2} \end{bmatrix}} \cdot \underset{(p+1)\times 1}{\begin{bmatrix} \beta_0 \\ \beta_{A1} \\ \beta_{B1} \\ \beta_2 \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix}} \tag{56}$$

It should be noticed that the test design follows a pretty similar way as the general F-statistic.

3.4.2 Testing two models in Generalized Linear Models

In section 3.1 we have provided a criterion to test a nested generalized linear models, that is the scaled deviance. Its expression in (14) completely matches the (maximum) likelihood ratio test statistic [5, 6]

$$D^*(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_1) = 2\big(l(\boldsymbol{\mu}_2|\boldsymbol{y}) - l(\boldsymbol{\mu}_1|\boldsymbol{y})\big) = -2\log\left(\frac{\mathcal{L}(\boldsymbol{\mu}_1)}{\mathcal{L}(\boldsymbol{\mu}_2)}\right) = LRT = \Lambda \tag{57}$$

where $\boldsymbol{\mu}_i \in \Omega_i$ with $i = 1..2$ and Ω_2 is the parameters space that includes Ω_1 parameters space. The $\hat{\boldsymbol{\mu}}_i$ are the maximum likelihood estimations.

It is important to notice, as it is noted in [6 Chap. 11], the scaled deviance of two nested models where none of them is necessarily a full model, it is the same as if we performed a difference of the deviances of the respective models with the full model for each of them:

$$
\begin{aligned}
D^*(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\theta}}_1) - D^*(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\theta}}_2) &= -2\log\left(\frac{\mathcal{L}(\boldsymbol{\mu}_1)}{\mathcal{L}(\boldsymbol{\mu}^*)}\right) + 2\log\left(\frac{\mathcal{L}(\boldsymbol{\mu}_2)}{\mathcal{L}(\boldsymbol{\mu}^*)}\right) \\
&= -2\log\left(\frac{\mathcal{L}(\boldsymbol{\mu}_1)}{\mathcal{L}(\boldsymbol{\mu}_2)}\right)
\end{aligned} \tag{58}
$$

where $\boldsymbol{\mu}_i \in \Omega_i$ with $i = 1..2$ and Ω_1 is the parameters space that is nested in Ω_2 parameters space. The dispersion parameter ϕ needed for the calculation of scaled deviance can be estimated by ML or the corrected version (dividing by d.o.f.) [11].

There are two options to test the models. Either we assume that the statistic follows assymptotically a χ^2 distribution with $p_2 - p_1$ d.o.f. (the proof can be seen in Appendix A.1); that is the difference in dimension of the nested parameter spaces for the models, or we can perform a *mean scaled deviance* as

$$
\frac{D^*(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_1)}{p_2 - p_1} \sim F - distribution \tag{59}
$$

where lemma 2 is applied. The only problem that perhaps may arise it is in case when parameters are not linearly independent, because in that case the d.o.f. would be less.

3.4.2.1 Global Λ-test in GLM

Let us assume that we split the parent node into two child nodes (segments), each one of them containing n_A and n_B observations, respectively. Then, we compute the two models as follows

$$
g(\boldsymbol{y}_A) = X_A\boldsymbol{\beta}_A
$$

and

$$
g(\boldsymbol{y}_B) = X_B\boldsymbol{\beta}_B
$$

with independent residuals. In order to see that the two models are different we formulate the following hypothesis: *the coefficients of segments A and B are equal in both equations at the same time.* That is the same expressed formally

$$
H_0 : \underset{n\times 1}{\begin{bmatrix} g(\boldsymbol{y}_A) \\ g(\boldsymbol{y}_B) \end{bmatrix}} = \underset{n\times p}{\begin{bmatrix} X_A \\ X_B \end{bmatrix}} \cdot \underset{p\times 1}{[\boldsymbol{\beta}]} \tag{60}
$$

and

$$H_1 : \underset{n\times 1}{\begin{bmatrix} g(\boldsymbol{y}_A) \\ g(\boldsymbol{y}_B) \end{bmatrix}} = \underset{n\times 2p}{\begin{bmatrix} X_A & 0 \\ 0 & X_B \end{bmatrix}} \cdot \underset{2p\times 1}{\begin{bmatrix} \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B \end{bmatrix}} \tag{61}$$

Using a test statistic presented in 3.4.2, it follows a χ^2 distribution with $n - p - n + 2p = p$ d.o.f. A special attention should be paid to the intercept term which is included by setting a column of one's to each X_A and X_B and then, β_0, β_{A0} and β_{B0} are also estimated in the null and alternative hypotheses, respectively.

3.4.2.2 Λ-coefficient test in GLM

In this section we present the Λ-test that aims to determine which of the coefficients is responsible for that difference in the model. Perhaps, we may find that there are more than one predictor involved in that difference. But, in any case the test is performed considering one-at-a-time predictor [2, 8]. For instance, considering only two independent variables and that we aim to test the coefficient β_1, the test would be formulated as follows

$$H_0 : \underset{n\times 1}{\begin{bmatrix} g(\boldsymbol{y}_A) \\ g(\boldsymbol{y}_B) \end{bmatrix}} = \underset{n\times(2p-1)}{\begin{bmatrix} \mathbf{1}_A & \boldsymbol{x}_{A1} & \boldsymbol{x}_{A2} & 0 & 0 \\ 0 & \boldsymbol{x}_{B1} & 0 & \mathbf{1}_B & \boldsymbol{x}_{B2} \end{bmatrix}} \cdot \underset{(2p-1)\times 1}{\begin{bmatrix} \beta_{A0} \\ \beta_1 \\ \beta_{A2} \\ \beta_{B0} \\ \beta_{B2} \end{bmatrix}} \tag{62}$$

and the alternative hypothesis

$$H_1 : \underset{n\times 1}{\begin{bmatrix} g(\boldsymbol{y}_A) \\ g(\boldsymbol{y}_B) \end{bmatrix}} = \underset{n\times 2p}{\begin{bmatrix} \mathbf{1}_A & \boldsymbol{x}_{A1} & \boldsymbol{x}_{A2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1}_B & \boldsymbol{x}_{B1} & \boldsymbol{x}_{B2} \end{bmatrix}} \cdot \underset{2p\times 1}{\begin{bmatrix} \beta_{A0} \\ \beta_{A1} \\ \beta_{A2} \\ \beta_{B0} \\ \beta_{B1} \\ \beta_{B2} \end{bmatrix}} \tag{63}$$

The test statistic follows a χ^2 distribution with $n - 2p + 1 - n + 2p = 1$ d.o.f. as presented in section 3.4.2. It can be easily extended to more than two covariates and iteratively testing for each coefficient at a time.

4 Profiling of segments

In this section we describe the procedure [14] which aims to find statistically significant insights, in order to detect and describe meaningful differences between two groups, once the segmentation has been performed.

Over a population of n individuals, p continuous variables, q categorical and one variable more that indicates those individuals who belongs to a group formed by n_k individuals, where $n_k < n$. The goal is to order the variables, which one by one best characterize the group of n_k individuals.

The variable will be not significant when the n_k values are being found to be chosen randomly over the n observations. But, the more doubtful seems to be this hypothesis, the better will be this variable to characterize the group.

Therefore, the test is proposed in a classical way, the null hypothesis states as *the n_k individuals are sampled randomly and without replacement over the n individuals.* In this case, we assume sampling without replacement as each of the individuals of n_k-th group is one and only one of the n individuals. Assuming the null hypothesis true, one will calculate the probability of observing at least as far as the configuration of values (of the n_k group) already observed. Then, the lower this probability (the p-value), the more we can question the hypothesis of random draw.

In the following we describe the hypothesis test for both, continuous and nominal variables.

4.1 Characterization for continuous variables

The idea is to determine how far is the mean of group n_k, $\bar{x}_k$, from the mean of the whole group of n individuals of variable j, $\bar{x}_j$. The test can be defined as follows

$$H_0 : \bar{x}_k = \mu_j \tag{64}$$

and

$$H_1 : \bar{x}_k \neq \bar{x}_j \tag{65}$$

In order to compute the sample mean of n_k individuals assuming the null hypothesis true, we have drawn n_k values without replacement from n individuals. Hence, we have that the expected value and the variance of this sample mean estimation are as follows

$$E_{H_0}[\bar{x}_k] = E[\bar{x}_j] = E[x_j] = \mu_j \tag{66}$$

and

$$Var_{H_0}(\bar{x}_k) = \frac{Var(x_j)}{n_k} \cdot \text{FPC} = \frac{\sigma_j^2}{n_k} \cdot \frac{n - n_k}{n - 1} \tag{67}$$

where σ_j^2 is the population variance of variable j and FPC is the Finite Population Correction factor. The test statistic is defined in the following way

$$u = \frac{\bar{x}_k - \mu_j}{\sqrt{\frac{\sigma_j^2}{n_k} \cdot \frac{n-n_k}{n-1}}} \tag{68}$$

Given that we do not know the population parameters μ_j neither σ_j^2, we substitute them by their MLE estimators, $\bar{x}_j$ and s_j^2, respectively.

When, n and n_k are big enough, the statistic u follows centered and reduced Gaussian distribution. The associated p-value under the null hypothesis is calculated as

$$p - value = p(|U| > |u|) \tag{69}$$

Moreover, the statistic u is called *valeur-test* and measures the distance between the two means in number of standard deviations.

4.2 Characterization for nominal variables

In this case the idea is to answer whether the proportion of individuals possessing the modality j of some categorical variable in group of n_k individuals is greater than the proportion in the whole population of n individuals.

Similarly as in 4.1, the null hypothesis assumes the random draw without replacement from the population of n individuals, which in other words means equality of proportions $\frac{n_{jk}}{n_k}$ and $\frac{n_j}{n}$. As stated in [14], the alternative hypothesis denotes unilateral assumption of anormally greater proportion of individuals possessing modality j in the group of n_k individuals. But in this work we describe the two-sided alternative hypothesis; in the same way the results are returned in `catdes` **R** function. Formally, the hypothesis test can be expressed as follows:

$$H_0 : \frac{n_{jk}}{n_k} = \frac{n_j}{n} \tag{70}$$

and

$$H_1 : \frac{n_{jk}}{n_k} \neq \frac{n_j}{n} \tag{71}$$

We define the random variable N as *the number of individuals taking the modality j in the group of n_k individuals.*The success is defined, then, as possesing the modality j.

We could define the random variable $\bar{N} = \frac{1}{n_k}N$ as the proportion of individuals having modality j in group of n_k. As can be seen in Appendix A.5, assuming the null hypothesis true, its variance can be defined as $Var_{H_0}(\bar{N}) = \frac{1}{n_k}\frac{n_j}{n}\left(1-\frac{n_j}{n}\right)\frac{n-n_k}{n-1}$. Therefore, we can define the following statistic

$$u' = \frac{n_{jk}/n_k - n_j/n}{\sqrt{Var_{H_0}(\bar{N})}} \tag{72}$$

where we take into account finite population constraint for the standard error, but the variable N in this case follows a Binomial distribution with parameters n_k and $\pi = \frac{n_j}{n}$ assuming H_0 true.

However, we assume sampling without replacement in finite population as in [14]. Hence, the random variable N follows a *Hypergeometric* distribution (see the Appendix section A.5). Therefore, we are interested in measuring

$$P_{H_0}(N \geq n_{jk}) = \sum_{x \geq n_{jk}} \frac{\binom{n_j}{x}\binom{n-n_j}{n_k-x}}{\binom{n}{n_k}} \tag{73}$$

the probability of observing at least as much individuals possessing modality j in group of n_k individuals as n_{jk}, assuming the null hypothesis true. The greater n_{jk}, the lower the expression in (73) and the more doubtful is the null hypothesis of random draw. When this probability is lower than a desired significance level α, we refuse the null hypothesis and state that the modality j makes significant difference among the two groups (those formed by n_k and n individuals, respectively).

In practice, it is very convenient to approximate the computation of such statistic by a centered and reduced Gaussian distribution [22]:

$$u = \frac{2}{\sqrt{n-1}}\left\{\sqrt{n_{jk}+1}\sqrt{n-n_j-n_k+n_{jk}+1} - \sqrt{n_j-n_{jk}}\sqrt{n_k-n_{jk}}\right\} \tag{74}$$

5 PATHMOX methodology

As we have stated in section 1, often we are interested in identifying groups of individuals not known a priori, which in addition may be a product of split by different segmentation variables. The number of this kind of variables may be quite high, which makes a very interesting the possibility of applying some automatic method of finding those segments. This can be achieved by PATHMOX approach [8, 19].

PATHMOX method is based on widely known techniques, such as Automatic Interaction Detection (AID, 1971), CHAID algorithm (1980), CART decision trees (Breiman 1984) and later ID3 and $C4.5$ (Quinlan 1986 and 1993, respectively). Moreover, a more sophisticated methods such as Multivariate Decision Trees have been proposed and combination of linear (CART) and non-linear, wrt the input X variables and weights, separators (MLP) [20], called also as *hybrid* models.

The approach presented in this work is based on Binary trees. The idea is to split each node by some modality of some segmentation variable in such a way that the resulting segments' models, built on defined set of modeling variables, are significantly different, in statistical sense. To achieve this, an algorithm based on Decision Trees methods is provided, which in addition implements a pre-pruning mechanism. That is, the splitting is performed until the stopping criterion is not satisfied (when it is, we stop growing that branch) and there are still some possible splits.

5.1 Particularities of binary trees

Following the Breiman et al. (1984) structure of building such a trees, there are several steps that should be performed. First, we should define a set of possible partitions that improve some functional which measures the goodness of the split (in CART trees it is defined through impurity measures, like *gini* impurity measure, *entropy* measure and *missclassification* measure). This is called the *split* criterion. The second step is to define the stopping rule, which is usually understood as a pre-pruning mechanism that avoids growing the tree on the given branch. Finally, the resulting leaf nodes are defined a model for further analysis. The models may be for prediction and also for exploratory tasks, for instance Principal Component Analysis.

This type of trees defines an upper bound of binary partitions for each type of partitioning variable (see table 1). In our particular problem, the split criterion and one of the stopping rules are implemented through a statistical tests; that is, when the level of heterogeneity in the node is not significant, the node is not splitted and automatically the growth of the tree is stopped. The split is performed by the most significant partition variable, the one that achieves the greatest significance.

Variable type	#binary splits
Binary	1
Nominal	$2^{k-1} - 1$
Ordinal	$k - 1$
Numerical	$k - 1$

Table 1: Number of binary splits for each type of partitioning variable (having k as number of levels or modalities).

5.2 The PATHMOX algorithm

In the context of Generalized Linear Models and Multiple Linear Regression models, we aim to identify a subgroups of population by iteratively partitioning the data by predefined segmentation variables. In order to achieve this in an automatic way, the procedure is based on binary partitioning processes obtaining a binary tree with different models on leaf nodes. The internal nodes may also keep their corresponding models. The main difference with classical Decision Tree methods which implement prediction models, our goal is to determine the subgroups that have significantly different models; that is, the behaviours of these groups are different. Additionally, when two models are found to be different, we also perform a statistical test to determine which predictors have originated the split. Originally, the PATHMOX methodology was proposed by Gastón (2009) [19] in the context of Partial Least Squares - Path Modeling [19]. In this work we present the PATHMOX algorithm in the context of Generalized Linear Models.

Algorithm 1 Pathmox

```
1: procedure PATHMOX(y, X_s, X_m)
2:     Fit global model at the root node
3:     All binary partitions list of the node by segmentation variables x_j ∈ X_s
4:     From most significant split (if any) add nodes to the queue q (as in BFS)
5:     while q.empty() = FALSE do
6:         node = q.top()
7:         q.pop()
8:         if stopping rule = FALSE then
9:             Repeat steps 3 and 4
10:        end if
11:    end while
12: end procedure
```

The stopping rule in this case represents a pre-prunning mechanism that controls for the minimum number of individuals for the split to be allowed and/or the maximum depth of the tree.

6 Simulations

In this section we present several simulation studies. We consider two different data generation mechanisms. One, for the Multiple Linear Regression problems and the second for Generalized Linear Models, in particular for Binomial family, however the generation mechanism can be used jointly with other of *logit* link functions.

The idea is first to check the validity of generated data by observing the sample distribution of the estimates, which follow a Normal distribution (property of Maximum Likelihood estimators [7]). Then, we compare the behaviour of the test statistics, F/Λ-global and F/Λ-coefficient. The first one is to determine if the two models are different enough so as to be doubtful against the null hypothesis. The second statistic allows us to determine which coefficient/s are responsible for that difference between models, hopefully also captured by the first F/Λ-global statistic. We also provide an analysis summarizing the statistics behaviour against the configuration of the data generating parameters. Finally, we present the statistics' values plotted over their theoretical distributions in Appendix B.

6.1 Models

We have considered two possible scenarios. The first refers to Multiple Linear Regression (MLR) and the second scenario is related to the Logistic Regression (LR) or GLM with *logit* link function.

The probability model for MLR is

$$y \sim N(\mu, \sigma^2) \tag{75}$$

the parameter's space

$$\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R} \tag{76}$$

and the linear predictor

$$\eta = g(\mu) = \mu = E[y] = \beta_0 + x_1\beta_1 + x_2\beta_2 \tag{77}$$

The probability model for LR is

$$y \sim Binomial(m, \pi) \tag{78}$$

the parameter's space

$$m \in \mathbb{Z}^+, \pi \in [0, 1] \tag{79}$$

where, in particular, $m = 1$ (ungrouped data), and the linear predictor

$$\eta = g(\mu) = g\Big(E\Big[\frac{y}{m}\Big]\Big) = g(\pi) = \log\frac{\pi}{1-\pi} = \beta_0 + x_1\beta_1 + x_2\beta_2 \tag{80}$$

6.2 Data generation configurations

In the table 2 are provided the conditions of the simulations. The shaded column of parameters corresponding to σ (noise) is only used for MLR models. So, for instance, if we perform a cartesian product, we obtain the total number of simulations, 72, for each parameter's combination. For each combination; that is, for instance, $n = 100$ (Small sample size), $\sigma = 0.5$ (Medium noise), $\boldsymbol{\beta}_A = (3, 0.7, 0.7)'$ and $\boldsymbol{\beta}_B = (3, 0.9, 0.7)'$ (C_3) we perform all statistical analyses over 100 iterations with this specific combination. The configurations C_1 - C_8 correspond to the different combinations of theoretical coefficients in both segments.

n		σ			β_{A0}	β_{A1}	β_{A2}	β_{B0}	β_{B1}	β_{B2}
Small	100	Small	0.1	C_1	3	0.7	0.7	3	0.7	0.7
Medium	400	Medium	0.5	C_2	3	0.7	0.7	7	0.7	0.7
Big	1000	Big	1	C_3	3	0.7	0.7	3	0.9	0.7
				C_4	3	0.7	0.7	3	1.2	0.7
				C_5	3	0.7	0.7	3	1.8	0.7
				C_6	3	0.7	0.7	3	0.9	0.9
				C_7	3	0.7	0.7	3	1.2	1.2
				C_8	3	0.7	0.7	3	1.8	1.8

Table 2: Parameter's configurations for the simulations. The shaded column of parameters (σ) is only available for MLR models.

To be noticed, that for the case of Logistic Regression we only have different configurations for the sample size, n, and for coefficients, $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$. Hence, we obtain the total number of 24 simulation combinations. Again, we perform 100 iterations for each of them.

6.3 Data generation process

In order to generate data for Multiple Linear Regression case and for each configuration of parameters, we performed the following steps

1. Generate error vector for segment A: `rnorm(na, mean = 0, sd = sd)`.
2. Generate error vector for segment B: `rnorm(nb, mean = 0, sd = sd)`.
3. Generate the two variables x_1 and x_2, both following the same distribution $x_1, x_2 \sim Beta(n, 6, 3)$: `rbeta(n, 6, 3)`.

4. Generate the column of 1's and build the data frame.
5. Generate vector of responses for segment A:
   `as.matrix(Xa)%*%as.vector(beta.a)+as.vector(eps.a)`.
6. Generate vector of responses for segment B:
   `as.matrix(Xb)%*%as.vector(beta.b)+as.vector(eps.b)`.

For Logistic Regression case and for each configuration of parameters, the process is slightly different

1. Generate the two variables x_1 and x_2, both following the same distribution $x_1, x_2 \sim N(n, \mu = 0, \sigma = 1)$: `rnorm(n, mean = 0, sd = 1)`.
2. Generate linear predictor for segment A:
   `eta.a <- as.matrix(Xa)%*%as.vector(beta.a)`.
3. Generate linear predictor for segment B:
   `eta.b <- as.matrix(Xb)%*%as.vector(beta.b)`.
4. Generate vector of responses for segment A:
   `ya <- rbinom(n=na, size=1, prob=(exp(eta.a)/(1+exp(eta.a))))`.
5. Generate vector of responses for segment B:
   `yb <- rbinom(n=nb, size=1, prob=(exp(eta.b)/(1+exp(eta.b))))`.

6.4 Adequacy of the generated data

In order to validate the generation process, we have chosen the simulation where the theoretical (generating) coefficients are equal in both segments. That is, varying configuration of parameters such as sample size and noise (only in Multiple Linear Regression, as in Logistic Regression it is implicit), we are able to check the corresponding estimations from the null hypothesis model, which assumes equality of coefficients of both segments.

For the Multiple Linear Regression models, in the table 3 are shown the coefficient estimates jointly with their standard deviation (not the standard error) for each configuration of generating parameters. As it was expected, the lower the noise (irreducible term) and the greater the sample size, the more precise is the estimate.

For the Generalized Linear Model (Logistic Regression) case it happens a similar scenario. The greater the sample size, the better the estimate (table 4).

In the figure 2, we can see how the estimate behaves when the noise (σ^2) is small and the sample size is increasing. In this way, we are able to visually check the sample distribution of the maximum likelihood estimates, which asymptotically follow a Gaussian distribution [7]. Hence, we also plotted the 95% border points ($\pm 1.96 s_{\beta_i}$). It is clear that for greater sample size we obtain a better estimates.

Configuration		$\beta_0 = 3$	$\beta_1 = 0.7$	$\beta_2 = 0.7$
$\sigma = 0.1$	$n = 100$	2.9966 (0.0688)	0.7011 (0.0782)	0.7069 (0.0684)
$\sigma = 0.1$	$n = 400$	3.0028 (0.0338)	0.6962 (0.0367)	0.6997 (0.0351)
$\sigma = 0.1$	$n = 1000$	2.9992 (0.0224)	0.6996 (0.0237)	0.7019 (0.0197)
$\sigma = 0.5$	$n = 100$	2.9915 (0.3252)	0.7171 (0.3852)	0.7104 (0.3279)
$\sigma = 0.5$	$n = 400$	2.9539 (0.1427)	0.7948 (0.1128)	0.6614 (0.1459)
$\sigma = 0.5$	$n = 1000$	3.0076 (0.117)	0.6816 (0.1165)	0.7075 (0.1071)
$\sigma = 1$	$n = 100$	2.8516 (0.5884)	0.7242 (0.5853)	0.9165 (0.6557)
$\sigma = 1$	$n = 400$	3.0627 (0.3429)	0.6543 (0.3618)	0.6549 (0.3455)
$\sigma = 1$	$n = 1000$	3.0083 (0.2146)	0.6849 (0.2203)	0.7062 (0.2178)

Table 3: Simulation estimations of the coefficients for each configuration with their corresponding standard deviation in brackets (MLR).

Configuration	$\beta_0 = 3$	$\beta_1 = 0.7$	$\beta_2 = 0.7$
$n = 100$	3.7022 (2.6471)	0.8919 (0.8611)	0.8241 (0.5499)
$n = 400$	3.0651 (0.2701)	0.7126 (0.2173)	0.7057 (0.2177)
$n = 1000$	3.0156 (0.1623)	0.7015 (0.1477)	0.6882 (0.1266)

Table 4: Simulation estimations of the coefficients for each configuration with their corresponding standard deviation in brackets (LR).

The same happens in figures 3 and 4, but given that the noise is greater the estimates are less precise.

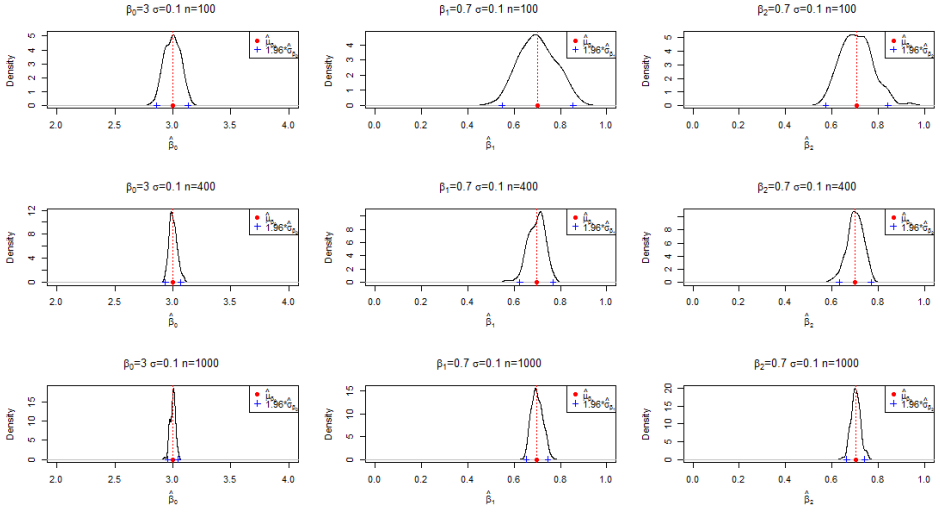

Figure 2: Sample distributions of $\boldsymbol{\beta}$ global parameters for $\sigma = 0.1$ (noise) and different sample sizes $n = 100$, $n = 400$ and 1000, from top to bottom respectively in MLR.

For LR we may observe that the generation process has been defined correctly as

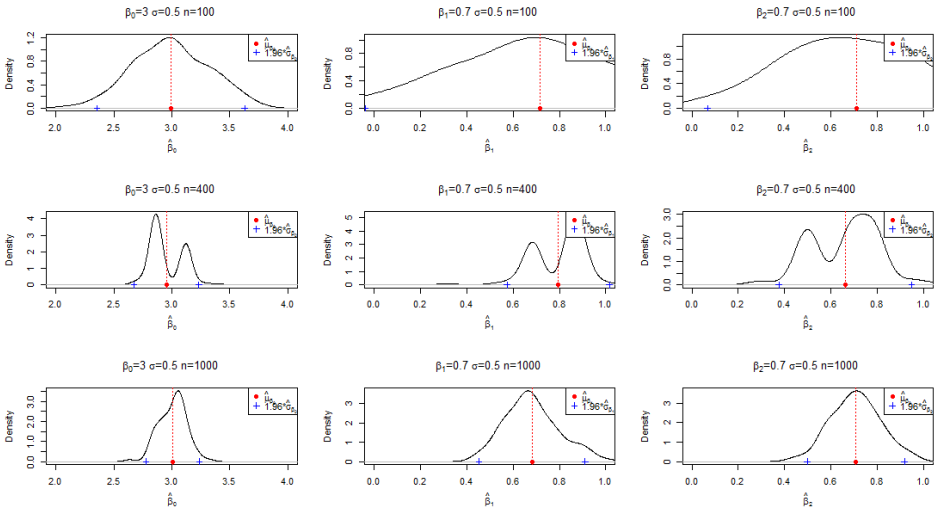

Figure 3: Sample distributions of $\boldsymbol{\beta}$ global parameters for $\sigma = 0.5$ (noise) and different sample sizes $n = 100$, $n = 400$ and 1000, from top to bottom respectively in MLR.

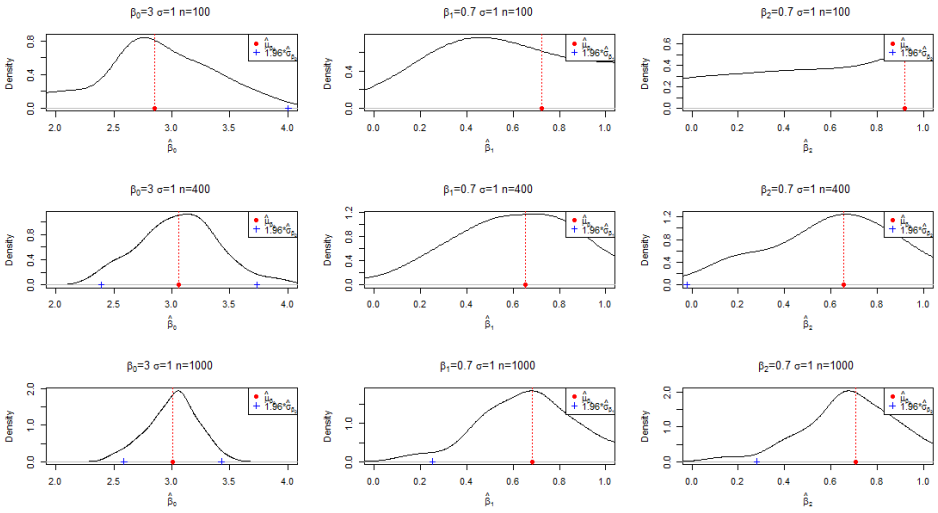

Figure 4: Sample distributions of $\boldsymbol{\beta}$ global parameters for $\sigma = 1$ (noise) and different sample sizes $n = 100$, $n = 400$ and 1000, from top to bottom respectively in MLR.

the coefficient estimates improve with the sample size increase (figure 5).

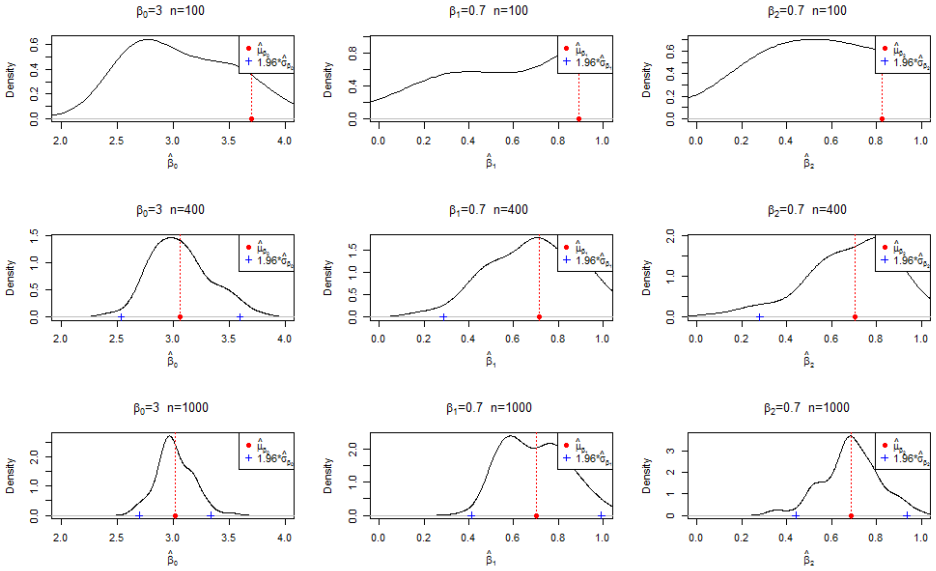

Figure 5: Sample distributions of $\boldsymbol{\beta}$ global parameters for different sample sizes $n = 100$, $n = 400$ and 1000, from top to bottom respectively in LR.

6.5 Hypothesis tests results

In this section we present the results for simulation study. For that, we include levelplots to assess in a very compact way the coherence between F/Λ-global and F/Λ-coefficient statistics. Moreover, we provide marginal mean plots for each of the configurations of the generating parameters, marginalizing over others. This allows us to verify the behaviour of the statistics under different scenarios.

In the figure 6 we can see the binarized levelplots of median p-value of the 100 iterations for each configuration of parameters. That is, for median significant p-values we set zero (magenta) and one otherwise (cyan). We can see that the F-global (top left panel) statistic performs very well and detects the difference between models when the parameters have some difference, even when it is small. It performs specially well when the sample size in each group is strictly greater than 50.

In the top right panel we have the F-coefficient binarized median p-values for the β_0 coefficient. The detection is 100% for any sample size. Moreover, the test does not detect other coefficient's differences, which is very satisfying.

In bottom left and bottom right panels we have results for β_1 and β_2 coefficients. For small noise, both coefficients are detected, except when the sample size is small and at the same time both coefficients in segment B are perturbed slightly. We can also see that the test gradually finds more difficulties when the variance is increased.

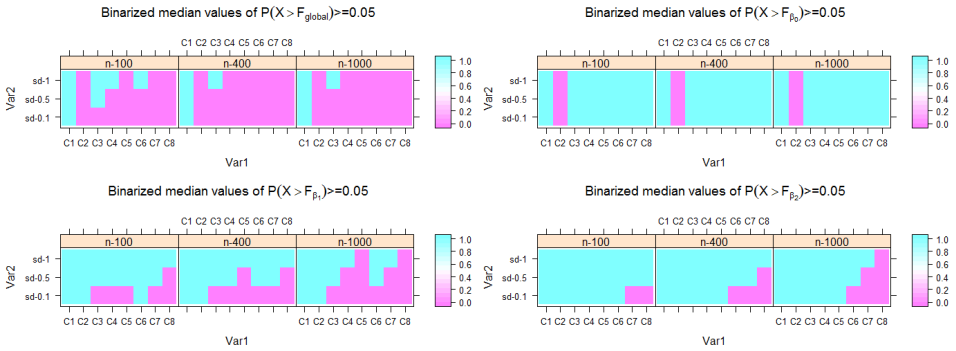

Figure 6: Binarized levelplot for median p-values of F-global and F-coefficient statistics for each combination (see table 2) in MLR.

In the figure 7 we can find the results for binarized levelplot of median p-values in case of Logistic Regression. The global Λ-statistic starts detecting differences between models for sample sizes strictly greater than 50 in each segment and when the difference of coefficients is big enough. The coefficient Λ-statistic behaves in a pretty much the same way. And as it happens in F-statistic (for Multiple Linear Regression case), it does not detect differences attributed to other coefficients than that under testing, which appears to be a very relevant confirmation of the expected behaviour.

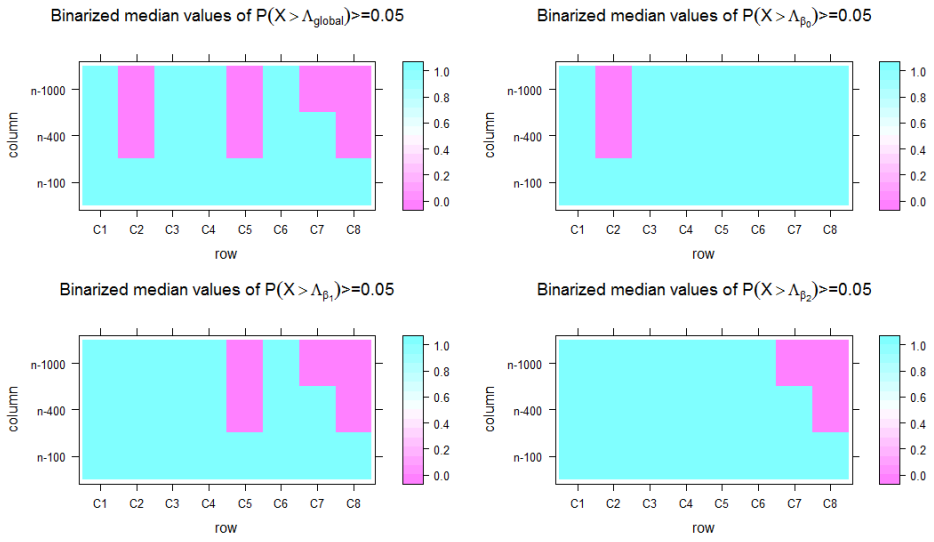

Figure 7: Binarized levelplot for median p-values of Λ-global and Λ-coefficient statistics for each combination (see table 2) in LR.

Further on, we can observe the statistics' behaviour through marginal plots (figures 8 and 9), where we aggregate by each parameter

- n - sample size

- σ^2 - variance of the response (only for MLR, as in LR it is implicit)
- $\boldsymbol{\beta}$ configuration - grade of difference of original coefficients

Therefore, for MLR case we can see that increasing the sample size the global F-statistic p-value (left panels of figure 8) gets lower as it is expected. Alternatively, for larger variance parameter the p-value tends to be higher, which is also a desired behaviour. For each configuration of coefficients (see table 2), we can observe that for configuration 1 we get a very high p-values as there is no difference between original coefficients. Moreover, the test finds more difficulties for configurations 3 and 6 as they correspond to the lowest differences between original coefficients.

The coefficient F-statistic p-values (right panels of figure 8) get lower for increasing sample size, except for the β_0 coefficient, as we only take into account those configurations where the coefficient were involved (see table 2); so that, for β_0 we only take into account p-values of the statistic having configuration C_2, for β_1 having configurations C_3-C_8 and for β_2, C_6-C_8. When aggregating by variance parameter, the correlation is positive as it happens in MLR. Again, for the β_0 coefficient test the p-values are very low for any sample size. Finally, the marginal plot (bottom right) depicts very well the behaviour of the coefficient tests when varying the difference between original coefficients. That is, for the β_0 only configuration C_2 makes the mean p-value to be lower than the significance level. The β_1 and β_2 are detected progressively for those configurations where they are involved, that is; C_3-C_8 for β_1 and C_6-C_8 for β_2.

Similarly, in case of Logistic Regression, both the global Λ-statistic and the coefficient Λ-statistic p-values decrease for larger sample sizes (top panels of figure 9). Regarding the configurations of the coefficient differences, we have that the mean p-values oscilate depending on the degree of difference between original coefficients. Hence, for configurations C_2, C_5 and C_8 we can see that the p-values decrease is noticeable. For the coefficient tests, it depends on the coefficient, so for β_0 the configuration where the difference is detected is C_2 as it was expected. Same happens with other coefficients: for β_1 the configurations are C_3-C_8 having the lowest p-values for C_5 and C_8 where the difference between original coefficients is the largest, whereas for β_2 the involved configurations are C_6-C_8 having that the lowest p-value is reached having the configuration C_8.

6.6 Alternative F-coefficient hypothesis test results

The simulation that has been performed is based on the hypothesis test provided in 3.4.1.3. It has been verified that it reflects the F-global test. For instance, if we take the combination of generation parameters $n = 1000$, $\sigma = 0.5$ and all combinations of $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$ coefficients, we find that whenever there is any difference among the coefficients of the two segments, even if it is very small, the test detects it for every coefficient.

For instance, consider the case when all the coefficients are equal in both segments,

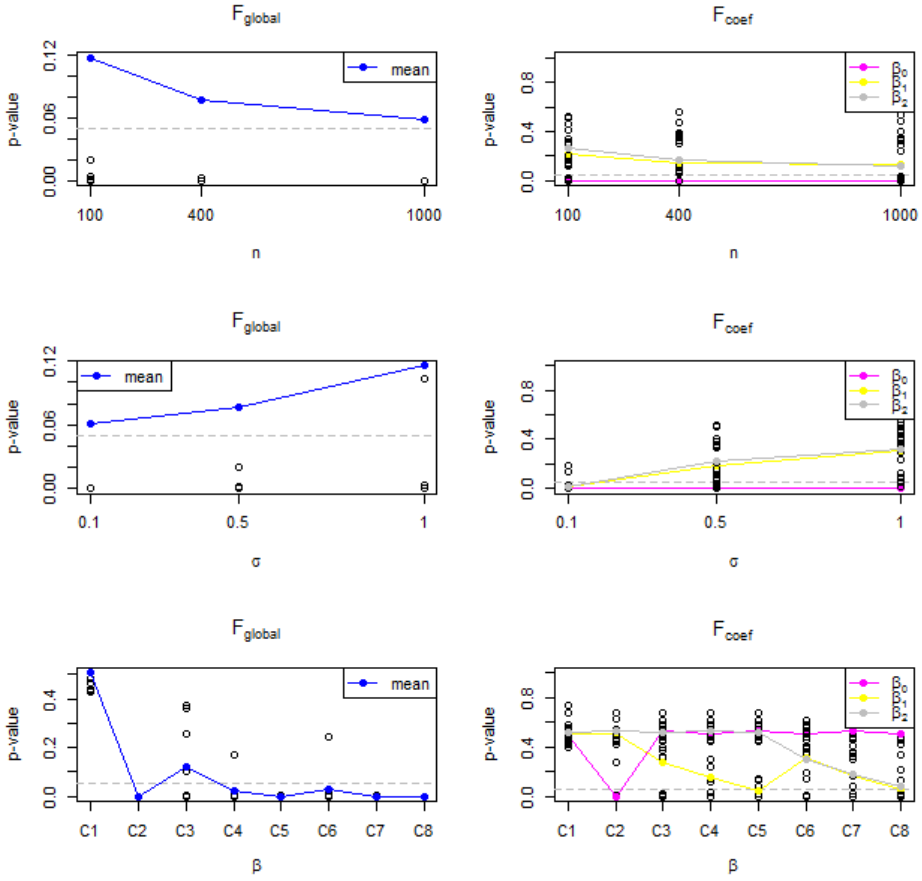

Figure 8: Marginal plots of F-statistics p-values for each parameter configuration (see table 2) in MLR.

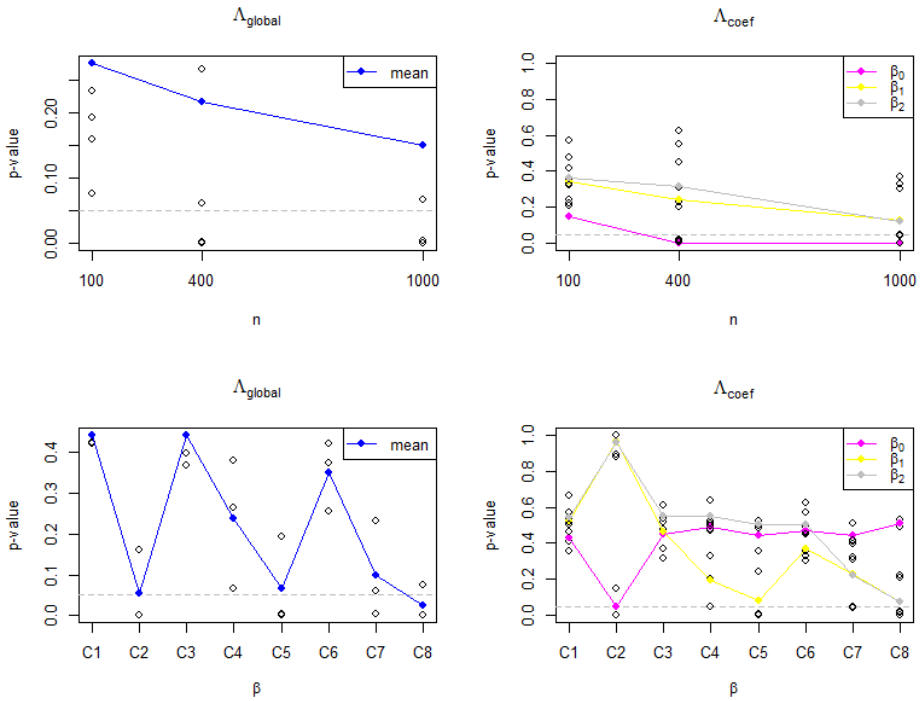

Figure 9: Marginal plots of Λ-statistics p-values for each parameter configuration (see table 2) in LR.

except the β_1; that is $\beta_{A1} \neq \beta_{B1}$. This configuration corresponds to Big n, Medium σ and C_5 from table 2.

The F-global test correctly detects such a difference from the models estimated for both hypothesis in one of the iterations of the simulation:

$$H_0 : \left\{ \hat{y} = 3.09 + 1.13x_1 + 0.69x_2 \right.$$

and

$$H_1 : \begin{cases} \hat{y}_A = 3.21 + 0.61x_1 + 0.49x_2 \\ \hat{y}_B = 2.98 + 1.72x_1 + 0.78x_2 \end{cases}$$

It shows a quite well approximated estimations as we can see that the estimation of the coefficient β_1 in segment B is much bigger and near to the expected value of 1.8.

Conversely, the alternative test for each coefficient, having the others fixed gives the following results. For β_0 coefficient the alternative hypothesis model is (the null hypothesis model remains the same):

$$H_1 : \begin{cases} \hat{y}_A = 2.73 + 1.19x_1 + 0.64x_2 \\ \hat{y}_B = 3.42 + 1.19x_1 + 0.64x_2 \end{cases}$$

For β_1 coefficient is

$$H_1 : \begin{cases} \hat{y}_A = 3.1 + 0.64x_1 + 0.64x_2 \\ \hat{y}_B = 3.1 + 1.7x_1 + 0.64x_2 \end{cases}$$

And for β_2 coefficient is

$$H_1 : \begin{cases} \hat{y}_A = 3.08 + 1.19x_1 + 0.14x_2 \\ \hat{y}_B = 3.08 + 1.19x_1 + 1.14x_2 \end{cases}$$

It turns out that those coefficients from alternative hypothesis model that have been fixed have a very similar estimations to the null hypothesis model, but the coefficient that can vary freely among the two segments is estimated in the way that it aims to reduce the residual variance, originated by the difference among segments, as much as possible. This leads the test statistic to be significant for any coefficient under assessment, as long as there is a difference between segments.

In the figure 10 we can see the result of the median p-values for each test, global and coefficient, and combination of parameters. It depicts the situation described above, whenever there is any difference between two segments, the F-coefficient test gives the same result disregarding which coefficient was responsible for the split.

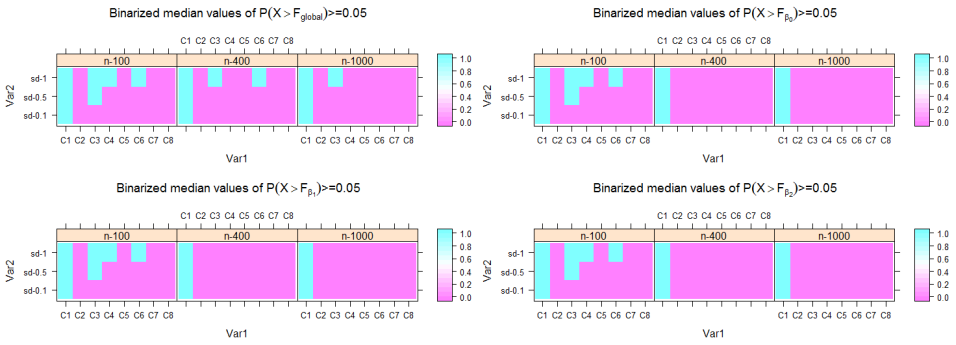

Figure 10: Binarized levelplot for median p-values of F-global statistic and an alternative F-coefficient statistic (see section 3.4.1.3) for each combination (see table 2) in MLR.

7 Application

In this section we present an application of the method presented in previous sections in a real context. We have used the dataset of Alumni Satisfaction Problem gathered from Informatics and Telecommunications schools of Technical University of Catalonia in 2008. This data can be obtained from "genpathmox" **R** package (Lamberti, 2014) available on CRAN. First we present the problem at hands and the available data in 7.1. The models are described in 7.2 and finally the results can be found in sections 7.3 and 7.4.

7.1 The dataset

The Alumni data, FibTele, was gathered from 147 students in 2008. Originally it was collected in order to study their satisfaction with respect to their actual work conditions using PLS-PM methodology [8, 19]. But in our case we aimed to study their salary taking into account *heterogeneity* (segmentation) variables and modeling variables in Generalized Linear Modeling context. Moreover, the modeling variables have been set to continuous scale and the binary modeling variable *Contract* has been binarized; that is, split into $n_{\text{levels}} - 1$ binary variables, one for each level minus one (in this case one binary variable with levels 0 and 1). In the table 5 we can see the segmentation variables and in the table 6 we can see the modeling variables or predictors jointly with the dependent variable *Salary* considered both as continuous and categorical.

Name	Scale	N. levels	Levels description
Career	nominal	3	EI / ETS / TEL
Gender	binary	2	female / male
Studying	binary	2	yes studying / no studying
Startwork	binary	2	after graduating / before graduating
Firmtype	binary	2	private / public

Table 5: Segmentation variables of FibTele dataset.

Name	Scale	Type	Values
Salary	continuous	dependent	16, 21.5, 30, 40, 50 (in thousands)
	categorical	dependent	< 30, > 30 {0,1}
Age	continuous	independent	25.5, 27.5, 29.5, 33 (in years)
Grade	continuous	independent	5.7, 6.7, 7.3, 8.5 (over 10)
Contract	binarized	independent	other-temp / fix {0,1}

Table 6: Modeling variables of FibTele dataset.

7.2 The models

The model considering the response variable *Salary* as continuous (Linear Multiple Regression model)

$$Salary = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Grade + \beta_3 \cdot Contract + \varepsilon \tag{81}$$

where $\varepsilon \sim N(0, \sigma^2)$. And the model considering the response variable as categorical; that is, for the Logistic Regression case

$$E[Salary]/1 = \pi = \frac{\exp\{\beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Grade + \beta_3 \cdot Contract\}}{1 + \exp\{\beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Grade + \beta_3 \cdot Contract\}} \tag{82}$$

where we have defined ungrouped data, hence taking $m = 1$ (denominator in the leftmost term of (82)).

7.3 Results for *Gaussian* family

In order to ensure an appropriate interpretation of results, it is convenient to take into account that the dataset is very small. Therefore, performing splits at each step of PATHMOX algorithm, the resulting segments are even smaller. This can lead to non-significance of the coefficient estimates. Even though, we do go ahead with the comparison of the models through the statistical tests presented in 3.4. The interpretation of results is followed carefully.

In the figure 11, we can see the resulting tree that defines two segments, one with Telecommunications modality and the second with Informatics modality (that includes Informatics Engineering and Technical Systems Engineering). It is a quite expected result given the prior knowledge.

The models in the root and the child nodes for *Gaussian* family are as follows

$$\begin{aligned} \hat{y}_{\text{root}} &= 15.18 + 0.23 \cdot Age + 0.47 \cdot Grade + 7.1 \cdot Contract \\ \hat{y}_{\text{TEL}} &= 67.91 - 0.97 \cdot Age - 1.55 \cdot Grade + 6.71 \cdot Contract \\ \hat{y}_{\text{INF}} &= 4.07 + 0.34 \cdot Age + 1.18 \cdot Grade + 9.05 \cdot Contract \end{aligned} \tag{83}$$

In the models provided in (83), we can observe a special importance related to *Contract* predictor in both segments. The intercept can not have a direct interpretation, but the other two coefficients related to *Age* and *Grade*, respectively, seem to have an opposite influence in the two segments. This will be discussed in the following analyses.

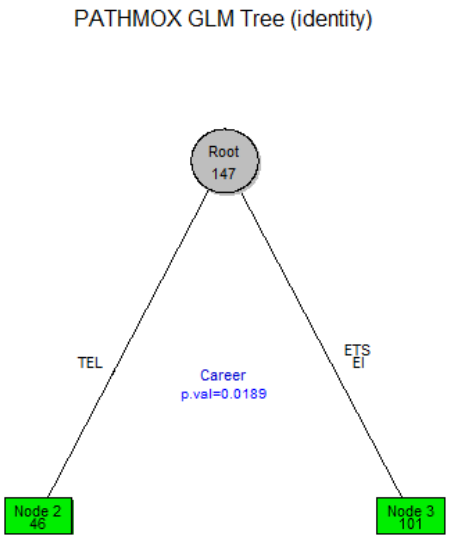

Figure 11: PATHMOX segmentation tree - GLM (*Gaussian* family).

In the table 7 we can see the results of applying both models to the mean profiles of the two segments, Telecos and Informatics. It appears to be that the mean individual in Telecos has greater salary than that of group of Informatics. It is important to underline that the Informatics group includes both degrees: Informatics Engineering and Technical Systems Engineering. Hence, this fact may influence the detected heterogeneity.

Segment	Age	Grade	Contract	Salary
TEL	30.07	6.83	0.6522	32.5
INF	29.05	6.995	0.7921	29.4

Table 7: Mean *Salary* for mean profiles from each segment, Telecos and Informatics for *Gaussian* family.

In the root node we have obtained several candidates (table 8). It is of special interest to pay attention to the first two candidates, by *Career* and by *Firmtype* in the second place, which is near of being significant. As we will find in Logistic Regression application on the same problem, that the two first candidates are the same but not the order among them.

In order to detect the responsible coefficients for the split in the root node we have performed the coefficient F-test presented in section 3.4. It turns out that

Variable	F-statitic	p-value	n_1	n_2	Mod. segm. 1	Mod. segm. 2	Dev. H_1
Career	3.0558	0.0189	46	101	TEL	EI/ETS	9763.02
Firmtype	2.0041	0.0973	116	31	priva	publi	10042.39
Startwork	1.0693	0.3741	53	94	after.grad	befor.grad	10304.46
Studying	1.0605	0.3785	70	77	no.stud	yes.stud	10306.99

Table 8: Candidate splits in the root node for *Gaussian* family.

the intercept is significant and *Age* coefficient is non-significant but the evidence against the null hypothesis is not very strong. This would mean, that by just being a Telecos engineer will lead to have a greater salaries in general. But, as we pointed out above, the group of Informatics includes both degrees (3-long and 5-years-long degrees).

Variable	F-coefficient statistic	p-value
Intercept	4.5107	0.0355
Age	2.8206	0.0953
Grade	1.9008	0.1702
Contract	0.4845	0.4875

Table 9: F-coefficient tests after the split for *Gaussian* family.

In order to assess the estimated coefficients for both segments, Telecos and Informatics, we have plotted the predictors against the dependent variable *Salary* for each group (figure 12). As we can see, for Informatics segment, the three independent variables have positive correlation as it is expected; that is, for greater *Age* and *Grade* and better *Contract* we obtain higher salaries. This does not happen in Telecos group, where seems to be an *Age* specific interval where the professional achieves its best remuneration and then it drops down. Another pattern can be detected for the *Grade*, where in the extremes, with a quite low grades or very high grades, is where the professionals achieve best salaries. The *Contract* has similar relation as in Informatics group. This has sense as the coefficients for this variable are both positive and similar in magnitude in both groups.

The conclusion is that the results obtained in this practical application should be interpreted carefuly given the small sample size and the patterns present in collected data.

Moreover, we have performed a Welch's unequal variances t-test to determine whether the Telecos salaries mean is statistically greater than that of Informatics without taking into account predictor variables as follows

$$\begin{aligned} H_0 &: \mu_{\text{TEL}} - \mu_{\text{INF}} = 0 \\ H_1 &: \mu_{\text{TEL}} - \mu_{\text{INF}} > 0 \end{aligned} \tag{84}$$

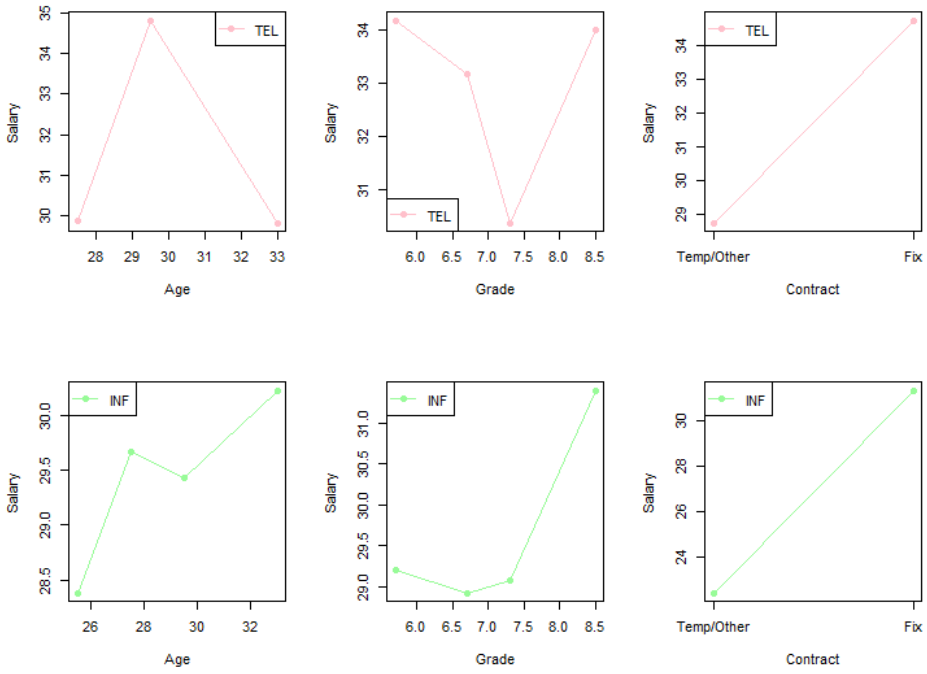

Figure 12: Plot of independent variables (*Age*, *Grade* and *Contract*) against the dependent variable *Salary* for both segments Telecos and Informatics, respectively.

The statistic can be computed as follows

$$t = \frac{\hat{\mu}_{\text{TEL}} - \hat{\mu}_{\text{INF}}}{\sqrt{\frac{s^2_{\text{TEL}}}{n_{\text{TEL}}} + \frac{s^2_{\text{INF}}}{n_{\text{INF}}}}} \tag{85}$$

where the denominator does not correspond to the square root of pooled variance. The pooled variance is used when the true variances are assumed to be equal, $\sigma^2_{\text{TEL}} = \sigma^2_{\text{INF}}$, because then we would have the following statistic

$$t_{\text{pool}} = \frac{\hat{\mu}_{\text{TEL}} - \hat{\mu}_{\text{INF}}}{s_{\text{pool}}\sqrt{\frac{1}{n_{\text{TEL}}} + \frac{1}{n_{\text{INF}}}}} \tag{86}$$

where s^2_{pool} is the pooled variance of the two samples. The definition can be found for instance in [17] and the proof is based on the normality and independence of the samples and then on the definition of the t-distribution.

In the case that we are treating, the statistic in (85); that is, when we do not assume equality of variances, the degrees of freedom are then computed by Welch-Satterthwaite equation [18] as follows

$$\nu = \frac{\left(\frac{s^2_{\text{TEL}}}{n_{\text{TEL}}} + \frac{s^2_{\text{INF}}}{n_{\text{INF}}}\right)^2}{\frac{s^4_{\text{TEL}}}{n^2_{\text{TEL}}\nu_{\text{TEL}}} + \frac{s^4_{\text{INF}}}{n^2_{\text{INF}}\nu_{\text{INF}}}} \tag{87}$$

having that $\nu_{\text{TEL}} = n_{\text{TEL}} - 1$ and $\nu_{\text{INF}} = n_{\text{INF}} - 1$ are the corresponding d.o.f. of the variance estimates.

The result given by **R** is provided in the table 10. From it, we can reject the null hypothesis with significance level $\alpha = 0.05$ and affirm that Telecos *Salary* mean is greater than that of Informatics.

Measurement	Value
t-statistic	1.7539
d.o.f.	67
p-value	0.042

Table 10: Result of Welch's t-test for *Salary* means of Telecos and Informatics samples.

7.4 Results for *Binomial* family

The resulting tree for Logistic Regression scenario for the same data can be seen in figure 13, where the split variable appears to be *Firmtype*.

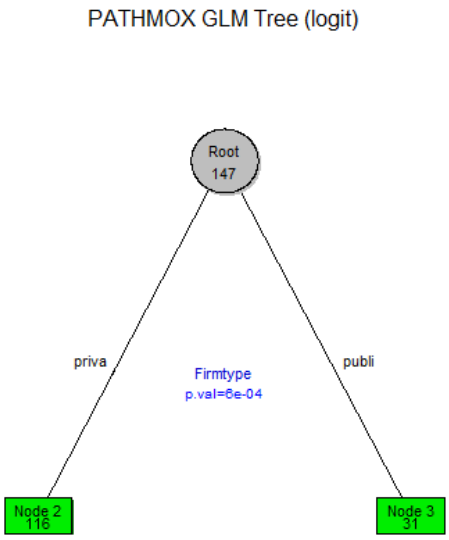

Figure 13: PATHMOX segmentation tree - GLM (Binomial family).

The models in the root and the child nodes for *Binomial* family are as follows

$$\begin{aligned} \hat{\eta}_{\text{root}} &= -4.19 + 0.07 \cdot Age + 0.26 \cdot Grade + 2.11 \cdot Contract \\ \hat{\eta}_{\text{priva}} &= -4.12 + 0.003 \cdot Age + 0.49 \cdot Grade + 2.19 \cdot Contract \\ \hat{\eta}_{\text{publi}} &= -18.7 + 1.27 \cdot Age - 2.49 \cdot Grade + 20.1 \cdot Contract \end{aligned} \tag{88}$$

From the models in (88) we can see that the segment *Private* has similar coefficients to those in the root node model. Alternatively, in the segment *Public* the coefficient of *Contract* plays a much greater role. Moreover, the *Grade*'s coefficient seems to penalize the salaries.

In the table 11 we find the evaluation of both models to the mean individual from each segment. The result is that the probability of having the salary greater than 30K is much greater in *Public* sector. Said this, we should take into account that the sample size is very small, specially for the segment *Public*.

The candidates as split variables in the root node are drawn in table 12, and as we pointed out in MLR scenario for this problem, the two first candidates are the same: *Firmtype* and *Career*, but not in their order. It has sense, as the detection method of heterogeneity is coherent, but still the difference in the order may be attributable

Segment	Age	Grade	Contract	$g^{-1}(\hat{\eta}) = \hat{\mu} = \hat{E}[Salary > 30]/m = \hat{\pi}$
Private	29.3	6.895	0.819	0.7577
Public	29.65	7.126	0.4839	0.9999

Table 11: Mean *Salary* for mean profiles from each segment, Private and Public for *Binomial* family (having $m = 1$ for ungrouped data).

to the binarization of response variable. Additionally, it is important to notice the sample size of the segments to interpret the obtained results.

Variable	Λ-statitic	p-value	n_1	n_2	Mod. segm. 1	Mod. segm. 2	Dev. H_1
Firmtype	19.7057	0.0006	116	31	priva	publi	123.65
Career	10.467	0.0333	46	101	TEL	EI/ETS	132.89
Studying	6.0927	0.1923	70	77	no.stud	yes.stud	137.26
Startwork	4.8329	0.3049	53	94	after.grad	befor.grad	138.52

Table 12: Candidate splits in the root node for *Binomial* family.

We have applied characterization study of both groups presented in section 4 without taking into account modeling variables. It turns out that only the variable *Studying* is significative and the highest salaries group is characterized by individuals that are not currently studying. This result is depicted in the table 13 and the corresponding plot 14.

Group	Variable	Cla/Mod	Mod/Cla	Global	p.value	v.test
$> 30k$	Studying=no.stud	82.8571	53.7037	47.6191	0.0148	2.4377
	Studying=yes.stud	64.9351	46.2963	52.381	0.0148	−2.4377
$< 30k$	Studying=yes.stud	35.0649	69.2308	52.381	0.0148	2.4377
	Studying=no.stud	17.1429	30.7692	47.6191	0.0148	−2.4377

Table 13: Description of each segment by the segmentation variables for *Binomial* family.

When performing a test to detect the variables responsible for the split, we find that all except *Contract* are so. Nevertheless, the evidence of *Contract* coefficient test against the null hypothesis is not so strong. The most significant coefficient is the *Grade*.

Variable	Λ-coefficient statistic	p-value
Intercept	0.583	0.4451
Age	5.7109	0.0169
Grade	7.8407	0.0051
Contract	3.7092	0.0541

Table 14: Λ-coefficient tests after the split for *Binomial* family.

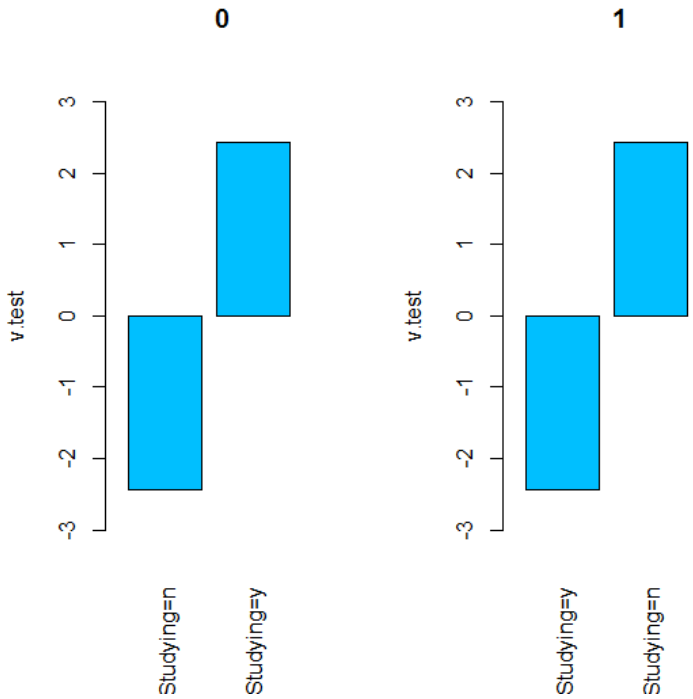

Figure 14: Profiling of *Salary* groups by segmentation variables only.

8 Conclusions and future work

We can make several conclusions and pose several extensions from the results obtained in simulation studies and real application, which involved detecting heterogeneity in Multiple Linear Regression models and in Logistic Regression models (GLM) using two different statistics, the one based on F-distribution and the other on χ^2.

1. The importance of taking into account the heterogeneity when modeling survey data. Alternatively, one can find a global model, a valid model, however it may represent an artificial mean estimation of two groups different by nature.
2. The F/Λ-statistics present coherent (detection under similar scenarios) and consistent behaviour; that is, when improving the quality of the sample in size or variance, their detection appears to be more significant. Nevertheless, the coefficient statistics behave more conservative than the corresponding global statistic, likely due to the greater degrees of freedom.
3. Both coefficient statistics (F and Λ) detect only the involved coefficient difference if there is any, but not the differences attributable to coefficients that are not under testing.
4. Our impression is that the χ^2 based statistic seems to be more prudent than that based on F-distribution.
5. Given the small sample size and hidden patterns present in FibTele data, their results should be interpreted carefully.
6. In real application, first two candidate split variables in root node coincide (but not in the order) in both scenarios: MLR and LR.
7. The PATHMOX approach is a very powerful method to detect heterogeneity, as it uses the information from modeling variables and given its scalability.
8. PATHMOX is very pwerful in the context of many possible segmentations without prior knowledge of which of them are really relevant, the situation that is easily found in Big Data context.
9. An extension of PATHMOX methodology to other models, such as Principal Component Analysis, Partial Least Squares regression, etc.
10. To enhance PATHMOX method including test statistics for each segment in order to provide measures of predictability by k-fold cross validation method.
11. Form a specific **R** package for the extensions just enumerated above.

A
Additional proofs and theoretical results

A.1 Likelihood Ratio Test contrast

The original version of this proof is provided in [7].

We want to see whether LRT statistic is distributed following a χ^2 distribution with $r = p_2 - p_1$ d.o.f. assuming that the null hypothesis is true; that is that the simple model is correct.

Developing $l(\theta)$ near the estimation $\hat{\theta}_{MLE}$, we obtain

$$l(\theta) \simeq l(\hat{\theta}_{MLE}) + \nabla l(\hat{\theta}_{MLE})'(\theta - \hat{\theta}_{MLE}) + \frac{1}{2}(\theta - \hat{\theta}_{MLE})' H(\hat{\theta}_{MLE})(\theta - \hat{\theta}_{MLE}) \quad (89)$$

where $\nabla l(\hat{\theta}_{MLE})$ is the vector of first derivatives which is null vector when evaluated at $\hat{\theta}_{MLE}$ (by definition) and $H(\hat{\theta}_{MLE})$ is the hessian matrix, matrix of second derivatives of the log-likelihood evaluated at the point $\hat{\theta}_{MLE}$. Hence, it can be simplified to give

$$l(\theta) \simeq l(\hat{\theta}_{MLE}) - \frac{1}{2}(\theta - \hat{\theta}_{MLE})'(-H(\hat{\theta}_{MLE}))(\theta - \hat{\theta}_{MLE}) \quad (90)$$

with $-H(\hat{\theta}_{MLE})$ being positive definite. The likelihood ratio test statistic is then defined as

$$LRT = 2(l(\hat{\theta}_{MLE}) - l(\hat{\theta}_0)) \quad (91)$$

where $\hat{\theta}_0$ is the MLE in the restricted space Ω_0 and given that we are assuming the H_0 true; so that the estimator $\hat{\theta}_0$ asymptotically tends to the true θ in the same way as $\hat{\theta}_{MLE}$, so they should be near. Therefore, we can use the Taylor expansion defined in (89) and obtain the simplified expression of the LRT. After few simple algebraic manipulations and cancelling some terms we get

$$LRT \simeq (\hat{\theta}_0 - \hat{\theta}_{MLE})'(-H(\hat{\theta}_{MLE}))(\hat{\theta}_0 - \hat{\theta}_{MLE}) \quad (92)$$

that is the Mahalanobis distance between $\hat{\theta}_0$ and $\hat{\theta}_{MLE}$. Asymptotically, $\boldsymbol{w} = \hat{\theta}_0 - \hat{\theta}_{MLE}$ has normal distribution and its expected value is zero as $\hat{\theta}_{MLE}$ tends to θ_0 assuming the H_0 true. $M = (-H(\hat{\theta}_{MLE}))^{-1}$ is the variance-covariance matrix of $\hat{\theta}_{MLE}$ [7 Chap. 4 Sect. 5.6].

It follows that

$$\boldsymbol{w}' M^{-1} \boldsymbol{w} \sim \chi^2_r \tag{93}$$

follows a chi-square distribution with $r = p_2 - p_1$ d.o.f. (the difference of the both parameters' dimensions).

A.2 Theorem of the three perpendiculars

We present an intuitive proof of the theorem of the three perpendiculars [12].

Theorem A.1. *If PQ is perpendicular to a plane XY and if from Q, the foot of the perpendicular, a straight line QR is drawn perpendicular to any straight line ST in the plane, then PR is also perpendicular to ST.*

Construction: Through Q draw in the plane XY the straight line LM parallel to ST.

Proof. Since LM is parallel to ST and QR perpendicular to ST hence, QR is perpendicular to LM. Again, PQ is perpendicular to the plane XY; hence, it is perpendicular to the line LM. Therefore, LM is perpendicular to both PQ and QR at Q. This implies LM is perpendicular to the plane PQR. Now, ST and LM are parallel and LM is perpendicular to the plane PQR; hence, ST is perpendicular to the plane PQR. Therefore, ST is perpendicular to PR or in other words, PR is perpendicular to ST. □

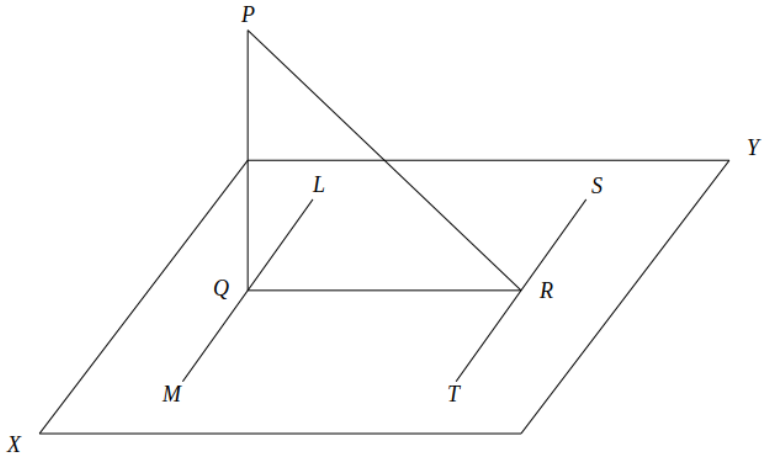

Figure 15: Graphical representation for the Theorem of the three perpendiculars.

A.3 PRESS statistic

The PRESS statistic is the measure of model adequacy and model prediction performance on new observations. For its calculation we need to define PRESS residual as it has been described in section 3.2.2:

$$e_{(i)} = y_i - \hat{y}_{(i)} \tag{94}$$

where $\hat{y}_{(i)}$ is the predicted value from the model fitted without observation i (as by widely used LOO-CV method). Apparently, in order to compute the PRESS statistic, we would need fit n different models omitting the corresponding i-th observation each time and computing the $e_{(i)}$. But, it can be shown that we only need to fit one model taking into account the whole dataset.

The vector of coefficients having omitted the i-th observation can be written as follows

$$\hat{\boldsymbol{\beta}}_{(i)} = \left(X'_{(i)}X_{(i)}\right)^{-1} X'_{(i)}\boldsymbol{y}_{(i)} \tag{95}$$

where sub-index (i) means that the observation i has been omitted (removed temporary).

Hence, the PRESS residual can be expressed as

$$\begin{aligned} e_{(i)} &= y_i - \hat{y}_{(i)} \\ &= y_i - \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_{(i)} \\ &= y_i - \boldsymbol{x}'_i\left(X'_{(i)}X_{(i)}\right)^{-1} X'_{(i)}\boldsymbol{y}_i \end{aligned} \tag{96}$$

Using the result from Appendix A.4, we can rewrite the previous expression as

$$\begin{aligned} e_{(i)} &= y_i - \boldsymbol{x}'_{(i)}\left[\left(X'X\right)^{-1} + \frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i\left(X'X\right)^{-1}}{1-h_{ii}}\right] X'_{(i)}\boldsymbol{y}_{(i)} \\ &= y_i - \boldsymbol{x}'_i\left(X'X\right)^{-1}X'_{(i)}\boldsymbol{y}_{(i)} - \frac{\boldsymbol{x}'_i\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i\left(X'X\right)^{-1}X'_{(i)}\boldsymbol{y}_{(i)}}{1-h_{ii}} \\ &= \frac{(1-h_{ii})y_i - (1-h_{ii})\boldsymbol{x}'_i\left(X'X\right)^{-1}X'_{(i)}\boldsymbol{y}_{(i)} - h_{ii}\boldsymbol{x}'_i\left(X'X\right)^{-1}X'_{(i)}\boldsymbol{y}_{(i)}}{1-h_{ii}} \\ &= \frac{(1-h_{ii})y_i - \boldsymbol{x}'_i\left(X'X\right)^{-1}X'_{(i)}\boldsymbol{y}_{(i)}}{1-h_{ii}} \end{aligned} \tag{97}$$

Having that $X'\boldsymbol{y} = X'_{(i)}\boldsymbol{y}_{(i)} + \boldsymbol{x}_i y_i$, we can manipulate the residual expression as follows

$$\begin{aligned} e_{(i)} &= \frac{(1-h_{ii})y_i - \boldsymbol{x}_i'\big(X'X\big)^{-1}(X'\boldsymbol{y} - \boldsymbol{x}_i y_i)}{1-h_{ii}} \\ &= \frac{(1-h_{ii})y_i - \boldsymbol{x}_i'\big(X'X\big)^{-1}X'\boldsymbol{y} + \boldsymbol{x}_i'\big(X'X\big)^{-1}\boldsymbol{x}_i y_i}{1-h_{ii}} \\ &= \frac{(1-h_{ii})y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + h_{ii}y_i}{1-h_{ii}} \\ &= \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{1-h_{ii}} \\ &= \frac{e_i}{1-h_{ii}} \end{aligned} \tag{98}$$

where the numerator is the ordinary residual from a model fit to all n observations.

A.4 A useful result on inverse of a matrix

We have that

$$X'_{(i)}X_{(i)} = X'X - \boldsymbol{x}_i\boldsymbol{x}_i' \tag{99}$$

where $X_{(i)}$ is the matrix X where observation i has been removed.

The following identity

$$\begin{aligned} \big(X'_{(i)}X_{(i)}\big)^{-1} &= \big(X'X\big)^{-1} + \frac{\big(X'X\big)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\big(X'X\big)^{-1}}{1-\boldsymbol{x}_i'\big(X'X\big)^{-1}\boldsymbol{x}_i} \\ &= \big(X'X\big)^{-1} + \frac{\big(X'X\big)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\big(X'X\big)^{-1}}{1-h_{ii}} \end{aligned} \tag{100}$$

where $h_{ii} = \boldsymbol{x}_i'\big(X'X\big)^{-1}\boldsymbol{x}_i$ can be seen by multiplying the right-hand side by the inverse of the left-hand side.

$$\left[\big(X'X\big)^{-1} + \frac{\big(X'X\big)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\big(X'X\big)^{-1}}{1-\boldsymbol{x}_i'\big(X'X\big)^{-1}\boldsymbol{x}_i}\right]\big(X'X - \boldsymbol{x}_i\boldsymbol{x}_i'\big) = I + A = I \tag{101}$$

where

$$
\begin{aligned}
A &= \frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i} - \left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i' - \frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i}\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i' \\
&= \frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i' - \left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i) - \left(X'X\right)^{-1}\boldsymbol{x}_i\left[\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\right]\boldsymbol{x}_i'}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i} \\
&= \frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i' - \left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i' + \left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left[\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\right] - \left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left[\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\right]}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i} \\
&= 0_{p\times p}
\end{aligned}
\tag{102}
$$

A.5 Basics on Hypergeometric distribution

In this section we describe several basic notions of Hypergeometric distribution [16, Chap. 11 - Finite Sampling Models] related to the hypothesis tests presented in section 4.

Let D be a dichotomous population of individuals each of them taking values 0 or 1. There are r observations taking value 1 and $m-r$, 0, having $|D| = m$. Then, n observations are sampled without replacement from D. The sample is collected through a binary variable X_i

$$
\boldsymbol{X} = (X_1, \ldots, X_n)' \tag{103}
$$

Therefore, we define a random variable Y, which is the number of successes in the sample $\boldsymbol{X}$:

$$
Y = \sum_{i}^{n} X_i \tag{104}
$$

The probability mass function of Y is defined as

$$
P(Y=y) = \frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}}, \quad y \in \{max(0, n-(m-r)), \ldots, min(n,r)\} \tag{105}
$$

Proof. The unordered outcome is uniformly distributed on the set of combinations of size n chosen from the population of size m. The denominator corresponds to the normalization constant; that is, all the possible ways of sampling n observations out of m. The numerator represents all the possible ways of choosing y 1's and $n-y$ 0's from the population of r 1's and $m-r$ 0's, respectively. □

The hypergeometric distribution is unimodal, having that

$$v = \frac{(r+1)(n+1)}{m+2} \tag{106}$$

then,

1. $P(Y = y) > P(Y = y-1)$ iff $y < v$.
2. The mode occurs at $\lfloor v \rfloor$ if v is not an integer, and at v and $v-1$ if v is an integer greater than 0.

The expected value can be derived as follows

$$E[Y] = E\left[\sum_{i}^{n} X_i\right] = \sum_{i}^{n} E[X_i] = n\frac{r}{m} \tag{107}$$

having that $E[X_i] = \frac{r}{m}$. And the variance of the random variable X_i can be derived as follows

$$\begin{aligned} Var(X_i) &= \sum_{x_j \in \Omega = \{0,1\}} P_{x_j}(x_j - E[x_j])^2 \\ &= P_0\left(0 - \frac{r}{m}\right)^2 + P_1\left(1 - \frac{r}{m}\right)^2 \\ &= \left(1 - \frac{r}{m}\right)\left(\frac{r}{m}\right)^2 + \frac{r}{m}\left(1 - \frac{r}{m}\right)^2 \\ &= \frac{r}{m}\left(1 - \frac{r}{m}\right) \end{aligned} \tag{108}$$

where Ω is the sample space of the dychotomous variable X_i, $P_1 = P(X_i = x_i = 1) = \frac{r}{m}$ and $P_0 = P(X_i = x_i = 0) = 1 - P_1 = 1 - \frac{r}{m}$.

The covariance [15, Chap. 30] can be seen as

$$\begin{aligned} Cov(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] \\ &= P(X_i = x_i = 1, X_j = x_j = 1) - \left(\frac{r}{m}\right)^2 \\ &= P(X_i = x_i = 1)P(X_j = x_j = 1 | X_i = x_i = 1) - \left(\frac{r}{m}\right)^2 \\ &= \frac{r}{m}\frac{r-1}{m-1} - \left(\frac{r}{m}\right)^2 \\ &= -\frac{r}{m}\left(1 - \frac{r}{m}\right)\frac{1}{m-1} \end{aligned} \tag{109}$$

where for $E[X_iX_j]$ we have used the definition $E[X_iX_j] = \int\int x_i x_j f(x_i, x_j)\ dx_i dx_j$ (replacing with summation for discrete case). And using expressions (108) and (109), the correlation expression can be derived as

$$Cor(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} = -\frac{1}{m-1} \tag{110}$$

provided that $Var(X_i) \neq 0$ and $Var(X_j) \neq 0$. And finally, the covariance definition for Y r.v. is in the following

$$\begin{aligned} Var(Y) &= Var\left(\sum_i^n X_i\right) \\ &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j) \\ &= n\frac{r}{m}\left(1 - \frac{r}{m}\right)\frac{m-n}{m-1} \end{aligned} \tag{111}$$

Comparing the expression (111) to the case when the sampling is with replacement; that is, when Y follows a Binomial distribution, the difference is the *Finite Population Correction* term:

$$FPC = \frac{m-n}{m-1}$$

A.5.1 Relation to class characterization hypothesis tests

As we described the method of characterization of classes in section 4. The approach requires sampling from finite population, and hence the need of correction factor, which has been derived from Hypergeometric distribution presented in Appendix A.5.

In order to relate the described in Appendix A.5 with the test for continuous and nominal variables from sections 4.1 and 4.2 respectively, we just take into account the notation changes as follows

- $m = n :=$ The size of the population
- $r = n_j :=$ The size of the subset of individuals having modality j
- $n = n_k :=$ The size of the sample without replacement
- $Y = N$

B
Additional simulation results

In the following we present additional plots in order to assess the quality of the detection mechanism, which aims to detect behavioural differences between two segments; that is, comparing their corresponding models.

B.1 Multiple Linear Regression models results

In the context of Multiple Linear Regression models we have applied F based statistics. We plot the test statistics calculated from the data (models) over the true F-distributions to visually see where do they fall. These plots are closely related to the levelplots presented in section 6.5.

In the figure 16 we can see as non of the statistics are significant as it was expected given that there are no differences between original parameters that generated both segments.

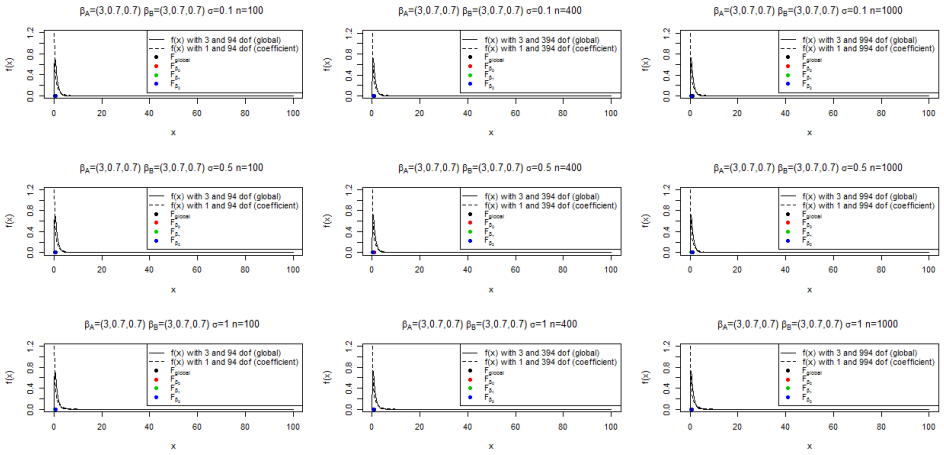

Figure 16: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

In the figure 17 both statistics F-global and F-coefficient corresponding to the β_0 are able to detect in all cases this difference, for any σ and sample size n.

From the figure 18 we can realize as for σ values greater than 0.1 the small difference originated on β_1 coefficient in segment B is not detected by the global test neither by coefficient test.

In the figure 19 we have that now the difference of β_1 coefficient from segment B is bigger and this helps both tests to detect it, but its detection is closely related to the σ and n values.

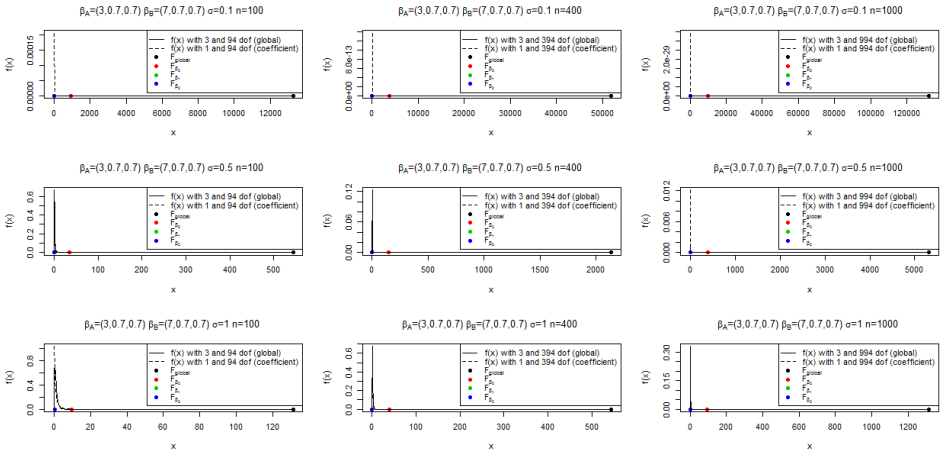

Figure 17: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

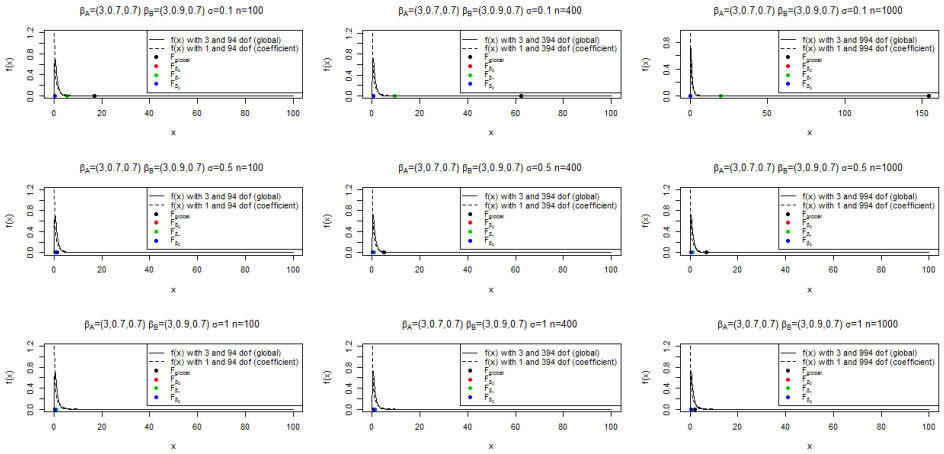

Figure 18: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

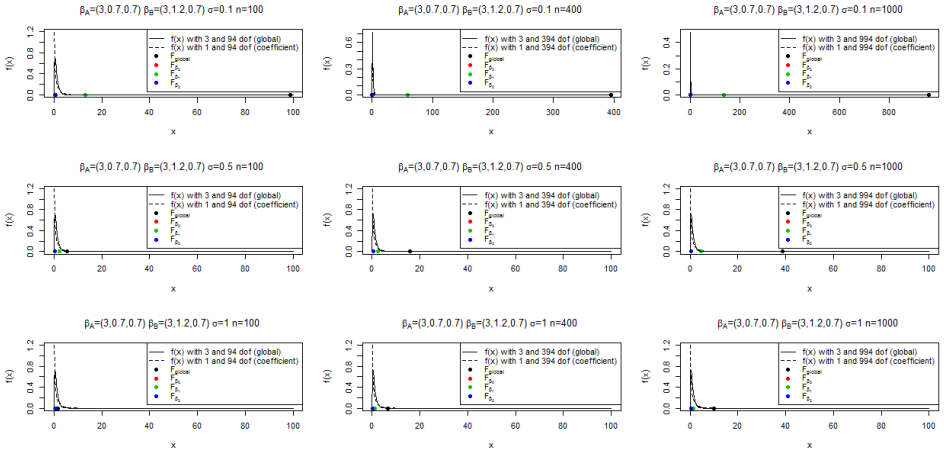

Figure 19: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

In the figure 20 the detection of β_1 coefficient difference from segment B appears to be better than for previous case, specially for the coefficient test.

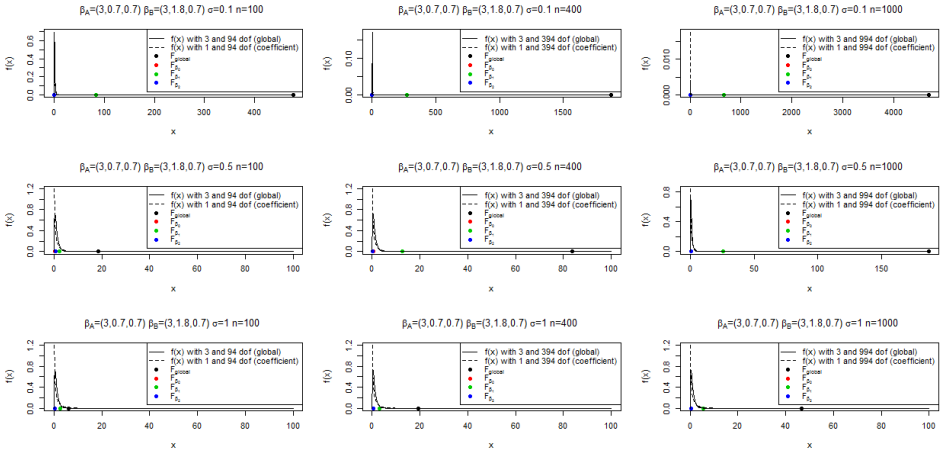

Figure 20: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

In the following plots 21, 22 and 23 the difference is gradually incremented for coefficients β_1 and β_2 from segment B. The results are very similar to those where only one coefficient, β_1 from segment B, has been modified.

B.2 Generalized Linear Models results (*Binomial* family case)

The case of Binomial family model is similar to the previous results for Gaussian data, but the difference is in the statistic used for the detection of differences between

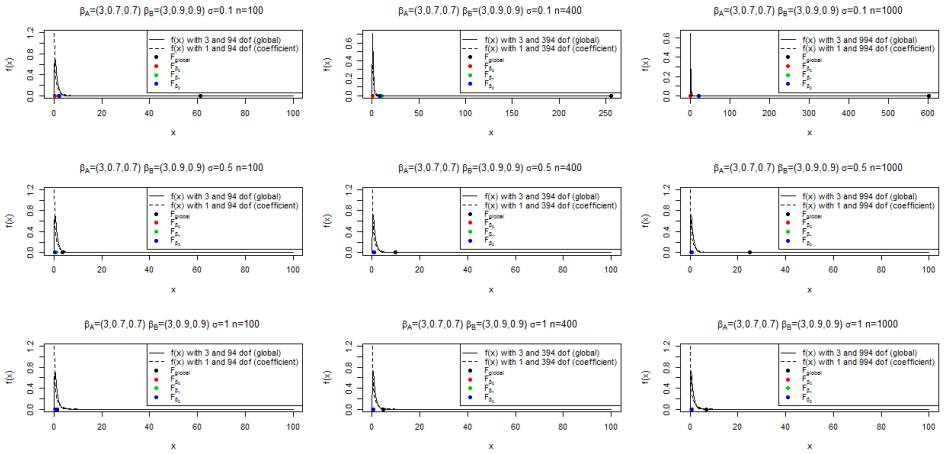

Figure 21: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

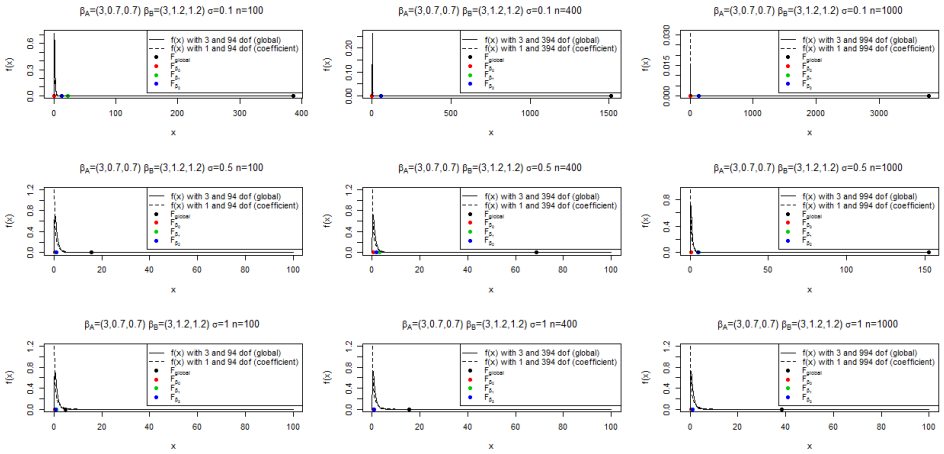

Figure 22: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

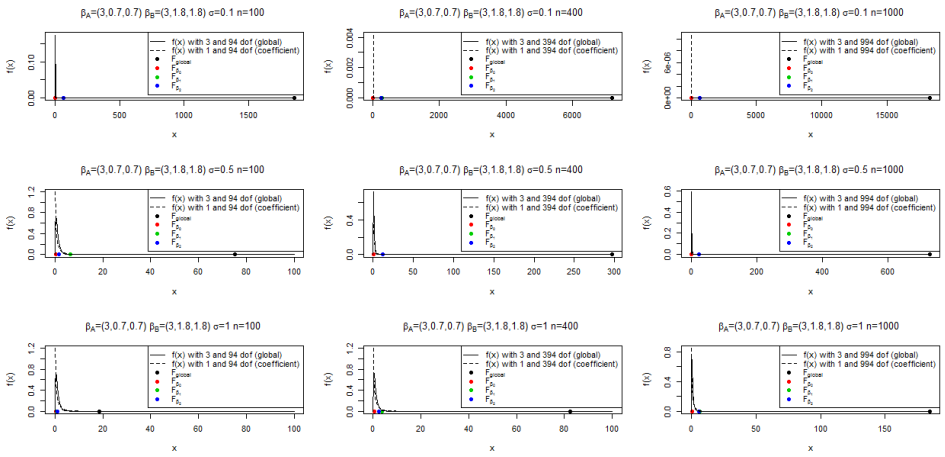

Figure 23: Comparison of F-statistics for global and coefficient hypothesis tests, varying variance and sample size for each coefficient design configuration.

models. In this case we use Λ-statistic for global test and coefficient test.

In the figure 24 we can see two cases, where no differences have been introduced and where only β_0 from segment B has been modified. In first case no one statistic detects any difference as it is expected, and in second case they do so, even though the detection seems to have more difficulties than in MLR case.

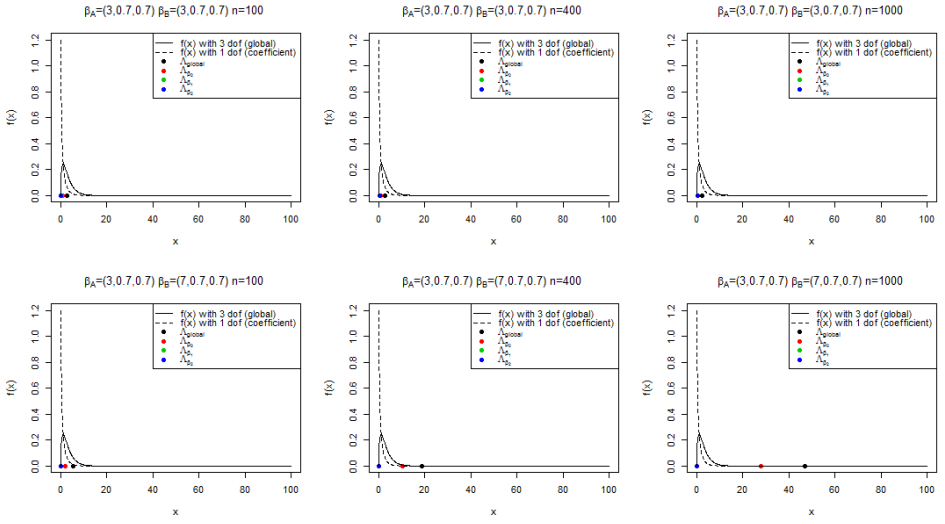

Figure 24: Comparison of Λ-statistics for global and coefficient hypothesis tests, for increasing sample size for each coefficient design configuration.

From figure 25 we can see that for small differences in β_1 coefficient from segment B the detection is hard (three top panels). It starts to find such a differences when the alteration in that coefficient is higher and for greater sample sizes (three bottom panels).

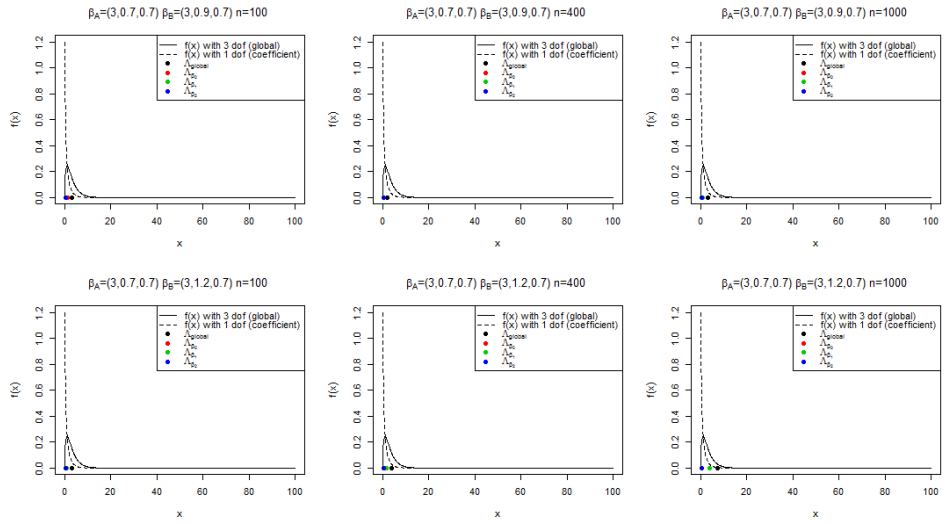

Figure 25: Comparison of Λ-statistics for global and coefficient hypothesis tests, for increasing sample size for each coefficient design configuration.

In the figure 26 the detection is quite good for $n > 100$ (in three top panels). But in three bottom panels we have both coefficients β_1 and β_2 from segment B slightly modified. This adds additional difficulty to the detection mechanism.

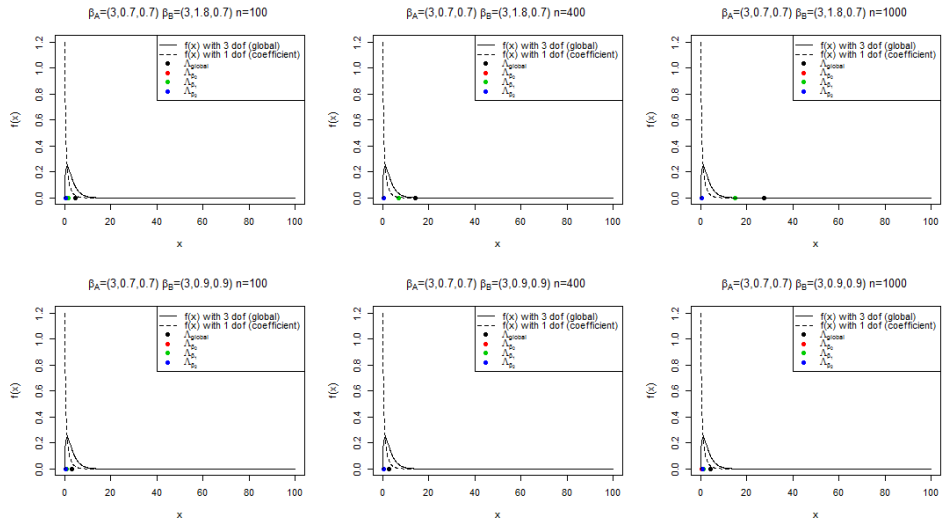

Figure 26: Comparison of Λ-statistics for global and coefficient hypothesis tests, for increasing sample size for each coefficient design configuration.

Finally, in the figure 27 we find a better detection as we increase the difference in both coefficients β_1 and β_2 from segment B, specially in the three bottom panels.

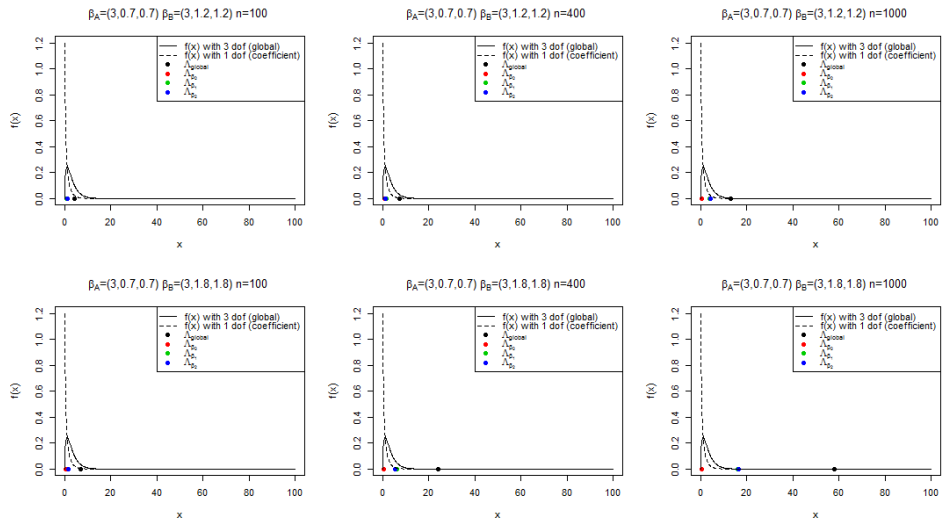

Figure 27: Comparison of Λ-statistics for global and coefficient hypothesis tests, for increasing sample size for each coefficient design configuration.

C
Software development

The Software development consisted in several tasks

- Review of functions from "genpathmox" **R** package (Lamberti, 2014) available on CRAN.
- Changes and improvements related to the F/Λ-statistics and d.o.f. calculations.
- Debug and double-check of the code step-by-step.
- Design and implementation of simulations for
  - Multiple Linear Regression models
  - Logistic Regression models (GLM)
  - Multiple Linear Regression models (alternative F-coefficient statistic)

We have paid a special attention to the improvement of the test statistics computations by differentiating the type of link function as two cases (*identity* and others) and calculating F-statistic and Λ-statistic, respectively.

We have also simplified the d.o.f. calculation by taking into account only the number of coefficient estimates for linear predictor, η. For instance, we do not count the σ parameter in MLR.

References

[1] Consultations with Dr. Tomàs Aluja Banet.

[2] Lebart L., Morineau A., and Féenelon J.P. Traitement des donneés statistiques. Paris. Dunod, 1979.

[3] Robert A. Beezer. A First Course in Linear Algebra. 2004.

[4] Peña D. Análisis de Datos Multivariantes. McGraw - Hill. 2002.

[5] P. McCullagh and J.A. Nelder. Generalized Linear Models. Chapman and Hall; Second edition. 1989.

[6] B. Francis, M. Green, C. Payne. GLIM 4: the statistical system for generalized linear interactive modelling. New York: Oxford University Press Inc. ISBN 0 19 (852231), 2. 1993.

[7] Peña D. Estadística Modelos y métodos 1. Fundamentos. Alianza ed., 1989.

[8] Lamberti G. Modeling with Heterogeneity. Doctoral Dissertation. Retrieved on 15/08/2016 from *http://www.tesisenred.net/handle/10803/309295.*

[9] Michael Baron. Probability and Statistics for Computer Scientists, Second Edition. Chapman and Hall/CRC, 2013.

[10] Peter E. Rossi. Bayesian Non- and Semi-parametric Methods and Applications. Princeton University Press, 2014.

[11] Paul E. Johnson. GLM - 2; Residuals and analysis of fit. 2016.

[12] Intuitive proof of the Theorem of the three perpendiculars. Retrieved on 15/09/2016 from
http://www.math-only-math.com/theorem-of-three-perpendiculars.html.

[13] D. C. Montgomery, E. A. Peck. Introduction to Linear Regression Analysis. Second edition, John Wiley and Sons, Inc. 1992.

[14] Morineau A. Note sur la Caracterisation Statistique d'une Classe et les Valeurs-tests. Bulletin Technique du CESIA, Vol. 2 nº 1-2. Paris, 1984.

[15] Charles M. Stanton. Introduction to Probability - Notes. Spring, 2012. Retrieved on 28/11/2016 from
http://www.math-only-math.com/theorem-of-three-perpendiculars.html

[16] Kyle Siegrist. Probability, Mathematical Statistics, Stochastic Processes. Department of Mathematical Sciences, University of Alabama in Huntsville. 2015. Retrieved on 14/12/2016 from *http://www.math.uah.edu/stat/*.

[17] González J.A., Cobo E., Muñoz P., Martí M. Estadística per a enginyers informàtics. Edicions UPC, 2008.

[18] Lloyd M. 2-Sample t-Distribution Approximation. Academic Forum 31 (2013–14).

[19] Sánchez, G. PATHMOX Approach: Segmentation Trees in Partial Least Squares Path Modeling. Doctoral Dissertation (2009). Retrieved on 1/08/2016 from
http://gastonsanchez.com/.

[20] Taner O., Alpaydın E. Univariate and Multivariate Decision Trees. Dept. of Comp. Engineering, Bogazici University, Istanbul. 2000.

[21] B.H. Robbins. Non-parametric tests. Scholars Series, Dept. of Biostatistics, Vanderbilt University. 2010.

[22] Molemar, W. Simple Approximation to the Poisson, Binomial and Hypergeometric Distributions. Biometrcis, vol. 29, pp. 403-407. 1973.

[23] Chin, W., Dibbern, J. An introduction to a Permutation Based Procedure for Multi-Group PLS Analysis. Handbook of Partial Least Squares. Springer, part 1, 171-193. 2010.