# Integrating data quality into GLM for insurance pricing

Pierre Chatelain

# Integrating data quality into GLM for insurance pricing

Chatelain Pierre[1]

[1]Univ Lyon ISFA LSAF EA2429, F-69007, Lyon, France
*Corresponding author: pierre.chatelain.act@gmail.com

**Abstract**
Pricing or regulatory works done by actuaries incorporate more and more external data provided by data providers. The reliability of these external data needs to be examined, since all aspects of regression are impacted by data quality. Therefore, actuaries as others modellers need to deal with this notion of quality. This paper studies the impact of data credibility on GLMs. This latter is measured by an exogenous and individualized quality index. Under a simple hypothesis that inconsistent data have the same distribution as consistent data, this paper proposes a method to find the true impact of a variable on the predictor. Under several assumptions, this method adapts the prediction depending on each data quality. Operational remarks and actuarial applications illustrate the creation and the use of quality indexes.

**Keywords:** Credibility, Quality index, OLS Regression, Generalized Linear Model

## 1. Introduction

Actuarial pricing was traditionally limited by the number of variables used and their complexity. Indeed, variables available to actuaries stem from the underwriting process. The potential client has a limited knowledge and amount of time to answer the questionnaire. To offset this problematic and improve the risks' knowledge, insurance companies use external data, for which its reliability is debatable. External data come from a third party for which the users do not have the ability to infer in the data creation process. Its reliability depends on each observation within a same variable. Indeed, often gathering processes aggregate data sets from various sources with heterogeneous quality. Legally, insurers' entities are responsible for the data quality (articles 219, 237, 244, 245, 247 from Solvency II Commission Delegated Regulation (EU) 2015/35). Their works must assess and justify the data quality even if they are coming from a third party : *'Data used in the internal model obtained from a third party shall not be considered to be appropriate unless the insurance or reinsurance undertaking is able to demonstrate a detailed understanding of those data, including their limitations'*, article 237. In France, the French Prudential Supervision and Resolution Authority, ACPR, [1, ACPR 2011] states that 10% of the data are coming from external parties. The data quality does not have a negligible impact as illustrate Campbell [3, Campbell et al., 2006] relating several actuarial examples. Therefore, actions must be triggered to assess and to take into account the data quality problematic. These different notions of quality have already been discussed for actuarial purposes in exploratory cases on the North American side ([7, Francis, 2005]) or on the UK side [3, Campbell et al., 2006] for instance. To the best of our knowledge, advices to take into account data quality are

still very qualitative ([8, GCASB, 2014]) such as basic recommendations : deleting, imputing or correcting the problem. These solutions will be discussed but are not sufficient. Models depend on observations' quality and the latter can be represented by quality indexes given by the data provider. How can an individualized and exogenous quality index be used for predicting ?

Literature suggests a multiple dimension analysis to evaluate data quality ([21, todoran, 2014]). For instance, the completeness dimension of data is a research field where numerous methods were developed to deal with missing values ([23, Van Buuren, 2018], [14, Little and Rubin, 2019]). These methods are globally based on assumptions such as MCAR (Missing Completely At Random), MAR (Missing At Random) or MNAR (Missing Not At Random). In the present paper, the credibility dimension will be studied further. On the mismeasurement side of uncertainty dimension, some works exist using trees algorithm ([22, Trabelsi et al., 2016], [20, Tami et al. 2018]) or the EIV-mismeasurment ([24, Van Huffel and Lemmerling, 2013]) framework. On the credibility side of uncertainty dimension, robust estimation theory as RANSAC (RANdom SAmple Consensus, [6, Fischler and Bolles, 1981]) algorithm and its different extensions such as KALMANSAC [25, Vedaldi et al., 2005] deal with outliers and inliers mostly used for computer vision. The downside of these methods is the left-aside observations for which no prediction can be made. It would be operationally inconceivable that some contract may not be priced.

In our framework, the credibility of observations is quantified and called quality index. It is assumed to be perfectly measured. Observations' uncertainty is modelled by a latent variable model. In this work, quality indexes are exogenous, individualized and equal to the probability that the observation is the true one. Indeed, this framework derives from works with a data provider. In different works, the data provider delivers data and quality indexes associated to it. The goal was to price household insurance contracts using building geolocation and external data. During this work, it was clear that the given quality indexes were evaluating the credibility of each observation more than its precision. These quality indexes are exogenous (given by the data provider) and the framework and assumptions developed in this paper arise from this case.

The main assumption is that wrong observations have the same distribution as the empirical one. Under this assumption,[4, Chatelain and Milhaud , 2021] considers the case of a basic linear regression and the correlation matrices. Because GLM are preferred in insurance industry, this paper will study the GLM cases through the likelihood. The goal is to give a precise answer to the following question. Given an individualized quality index (here based on credibility dimension), how can this quality index be used in a multivariate **GLM** ? How could actuaries set up a pricing model with a variable having quality problems?

**Contributions :**This paper presents two main contributions. First, it shows how to take into account quality indexes in a GLM regression. Next, several operational and practical remarks are given to help the creation and the use of quality indexes.

**Outline of the paper :** The paper is built as follows: in the section 2, the general framework and the notation are introduced. This work specifies how uncertainty is integrated in the covariate generating process. Section 3 gives the main algorithm and theoretical results. Hereafter, a simulation study illustrates the results in the section 4. Next, section 5 brings close the different assumptions to actuarial uses. In detail, subsection 5.4 and 5.3 discuss the use of quality indexes and the case of imperfect data quality indexes. All these remarks are illustrated by a practical case on household insurance.

## 2. Data problems and imputation

### 2.1 Notations

The set of all $n - m$ matrices where all its element are in the interval $I$ is denoted $\mathcal{M}_{n\times m}(I)$. $p$ represents the number of variable without the intercept and $n$ the number of rows. Data are important :

- $\mathbf{X} = (X_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ : the data set available with data quality problems *i.e.* observed covariates ;
- $\mathbf{X}^{real} = (X_{ij}^{real}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ : the data set, in practice not available , corresponding to the "real" observations.

We want to take advantage of the exogenous information provided by an *individualized* quality index related to the confidence we can have about the $i - th$ observation of the $j - th$ covariate, further denoted $Q_{ij}$.

In this view, we introduce the following latent variable model :

$$\mathbf{X} = \mathbf{X}^{real} \circ \mathbf{\Omega} + \mathbf{Z} \circ (J_{n,(p+1)} - \mathbf{\Omega}), \tag{1}$$

where :

- $\circ$ corresponds to the Hadamard product,
- $J_{n,(p+1)}$ is the $n \times (p+1)$-identity matrix under Hadamard multiplication,
- $\mathbf{Z} = (Z_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ are considered as the "wrong" covariate values having the same distribution as $\mathbf{X}^{real}$,
- $\mathbf{\Omega} = (\omega_{ij}) \in \mathcal{M}_{n\times(p+1)}(\{0, 1\})$ is a binary mask indicating whether the $i - th$ observation of the $j - th$ covariate $X_{ij}$ is perfectly observed or not. In other words, $\mathbf{\Omega}$ tells us if one observes the "real" observation or not. Assume that covariates distribution have second moment finite.

In practice, the data at disposal are made of individualized quality indexes through some matrix $Q = (Q_{ij}) \in \mathcal{M}_{n\times(p+1)}([0, 1])$, together with $n$ *i.i.d* replications $(Y_i, X_i)_{i=1,..,n}$ of $(Y, \mathbf{X})$, where $Y_i \in \mathbb{R}$ and $\mathbf{X}_i = (1, X_{i1}, ..., X_{ip}) \in \mathbb{R}^{p+1}$. The vector of quality indexes of the i-th row is written $\mathbf{Q}_i = (1, Q_{i1}, ..., Q_{in})$. A vector of specific values' quality indexes is called a quality pattern. Each element $Q_{ij}$ of the matrix $Q$ informs us on the quality related to the observed covariate value $X_{ij}$. Let use that $Q$ is the expectation of $\mathbf{\Omega}$, leading to define the quality index as a credibility index. This means that for all $i = 1, ..., n$, $j = 1, ..., p$ the quality index $Q_{ij}$ is equal to :

$$Q_{ij} = \mathbb{E}(\omega_{ij}) = \mathbb{P}(\omega_{ij} = 1) = \begin{cases} \mathbb{P}(X_{ij} = X_{ij}^{real}) & \text{if } \mathbf{X}_j \text{ is continuous variable,} \\ \mathbb{P}(X_{ij} = X_{ij}^{real}) - \mathbb{P}(X_j^{real} = X_{ij}^{real}) & \text{if } \mathbf{X}_j \text{ is discrete variable.} \end{cases} \tag{2}$$

The quality index corresponds to the probability to have taken not the "right" but the "real" observation. For a discrete variable, the part "$-\mathbb{P}(X_j^{real} = X_{ij}^{real})$" corresponds to the probability to get the true value randomly. In other words, $Q_{ij} = 0$ means that the value $X_{ij}$ is not informative on the risk of $i$. Denote for the rest of the paper $(j = 1, ..., p)$, $\bar{Q}_j = \frac{1}{n}\sum_{i=1}^{n} Q_{ij}$ and assume $\bar{Q}_j \neq 0$. This assumption is not restrictive, especially for real-life applications where such covariates would simply be removed from the data. However, it does not mean that an individual having all quality indexes null does not exist.

In this framework, the singularity is that $\mathbf{X}^{real}$ is not fully observed, which has consequences on the estimation of the regression coefficients.

## 2.2 Inapplicability of basic recommendations

The basic recommendations proposed by different actuarial works on deleting and imputing new values on "wrong observations" are not viable solutions for this framework.

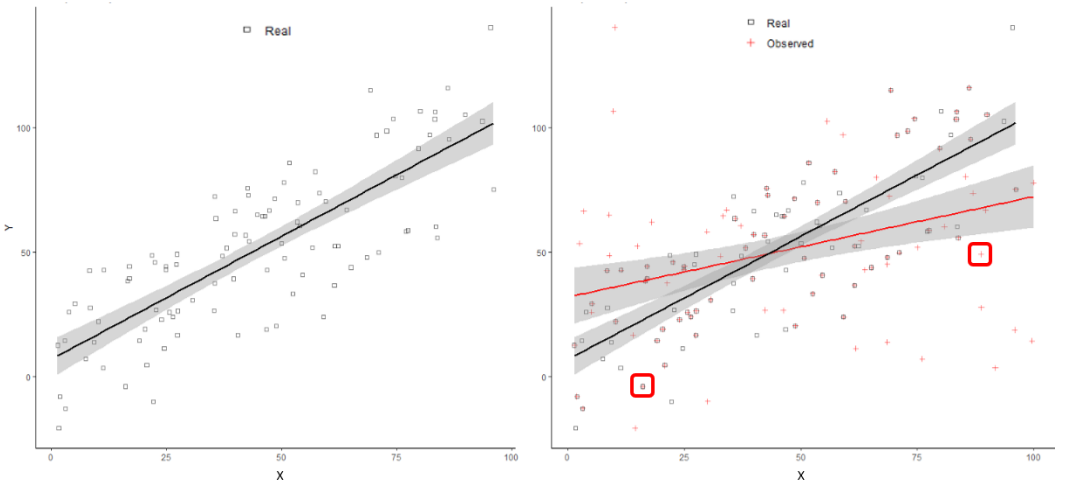



**Figure 1**: An univariate example where black squares are the real observations and crosses are the observed values. This graph is based on simulated data with $X \sim \Gamma(1,2)$ and $Y = 10 + 1X$, $Q$ follows an uniform distribution between 0.2 and 1. The data set $\mathbf{X}$ is created with the previous framework defined in subsection 2.1. Two linear regressions are fitted; one on the real data set $\mathbf{X}^{\mathbf{real}}$ and the other one on the observed dataset $\mathbf{X}$. Two points are highlighted : a real point which could be considered as an outlier and a wrong value which could be considered as an inlier.

**Imputating:** Let consider the strategy to impute new values on outliers or low quality observations[1]. Defining outliers in the multivariate case when the others covariates are not good quality is difficult. This is even more true in actuarial pricing where the outcome to model - claim cost, claim frequency, retention rate ..., has an intrinsic variability. Without taking into account exogenous information, robust estimation theory as RANSAC (RANdom SAmple Consensus, [6, Fischler and Bolles, 1981]) algorithm and its different extension have been developed using only a subsample of "real observation" (inliers) in modelling. Straightforwardly, in our framework, the data quality influences the definition of outliers for a regression, as shown in Figure 1. Indeed, the outliers' detection is bias due to the data set's quality. In that situation, some perfectly observed observations may be defined as outliers and the goal is also to predict values for individual with wrong observation(s). In the multivariate case and with variance outcome increasing, the definition of an outlier is operationally even more complex to deal with. For instance, in our framework, if $(X_{i,1}, X_{i,2})$ is defined as an outlier, is $X_{i,1}$ or $X_{i,2}$ or both wrong ?

**Deleting:** Given a data set and its joint quality index, a naive workaround of deleting low quality observations could be done. An easy one is to choose a threshold on the quality indexes and delete individuals having one of their quality indexes below. This solution can hardly be done with some low quality data or for highly dimensional datasets.

[1]Outliers detection and influential values have been studied for instance by [9, Hadi, 1991]) or [5, Cook, 1977].

Indeed, this latter issue was exemplified by [26, Zhu and al., 2019]. With an independent probability of a value missing equals to 0.05 and 300 covariates, this deleting approach would suppress 95% of the data set.

For our framework, let assume assumptions similar to Zhu et al. 2019 [26], *i.e.* in the case of complete independence of quality and observations[2]. Assume all the quality indexes independently distributed as an $Uniform(0.4, 0.8)$. Not only the low quality of the data implies a small threshold, but the different observations would highly range around the mean value. For a threshold of 0.5 and 10 variables defined as before, only 6 % rows would have all its covariates above the threshold in average. Besides, errors can be correlated spatially and this filtering process may bias the portfolio risks. For open data used in household insurance, this is in particularly true for urban area zones : covariates have often lower quality in rural areas. In short, filtering strategies are not optimal. Finally, neither imputing nor deleting are correcting the impact of quality on models.

### 2.3 An illustrative example

| Exposure | $X_1$ | $X_2$ | $X_3$ | $Q_3$ | $Y$ | Premium |
|---|---|---|---|---|---|---|
| 0.6 | 45 | 2 | 454 | 0.8 | 350 | ? |
| 1 | 30 | 3 | 1000 | 0.6 | 0 | ? |
| 1 | 43 | 2 | 2500 | 0 | 2450 | ? |
| 0.2 | 61 | 6 | 245 | 0.7 | 0 | ? |
| 1 | 53 | 3 | 723 | 1 | - | ? |
| 1 | 53 | 3 | 723 | 0.5 | - | ? |

Table 1. : The four first lines exemplifies a training dataset and the two last lines a testing data set.

Let consider a simple example : the explanatory variables, ($X_1$, $X_2$, $X_3$) and $\mathbf{Y}$. Here, only the last variable $X_3$ has an associated individualized quality index $Q_3$ where $Q_3 \in (0, 1)$. Each $X_{i;3}$ observation has a quality index $Q_{i;3}$ associated which is between 0 and 1 - 1 being an observation of perfect quality and 0 the worst one. The table 1 represents a dummy example. Here, $X_1$ refers to occupant age in year, $X_2$ to the number of rooms and $X_3$ house value in £ per $m^2$. Arbitrary, $Y$ could be the annual claims amount. From a training data set with an imperfect variable, how can actuaries predict the future mean claim cost knowing perfectly a value or in a more general knowing imperfectly a value ? Here, both last individual have the same observed characteristics but not the same quality index. How should the premium differ ?

First, the index can not be used as a weight in a multivariate regression. Indeed, the use of weights may bias the regression and does not correct the impact of quality. Secondly, table 1 displays another problematic : if an actuary fits a model with medium quality observations, how should he adapt its prediction for observations for which the covariate value is perfectly known or unknown ?

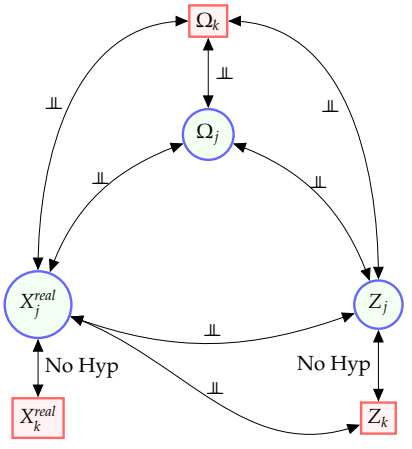
In our framework, quality indexes are associated to values between 0 and 1 as shown in example 1. In real-life application, quality indexes are exogenous information given by the data provider and take qualitative values such as "very high", "high", "medium",

[2]From the notations used in this paper: (C1) with the assumptions (X-A1) and (Z-A1).

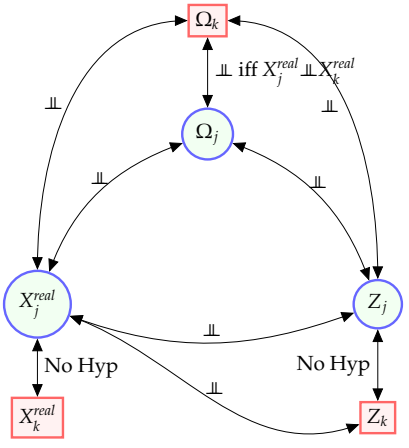
"low", "very low". To overcome this issue, the last section shows how to associate a value to a quality index modality using our theoretical framework.

### 2.4 Frameworks under study

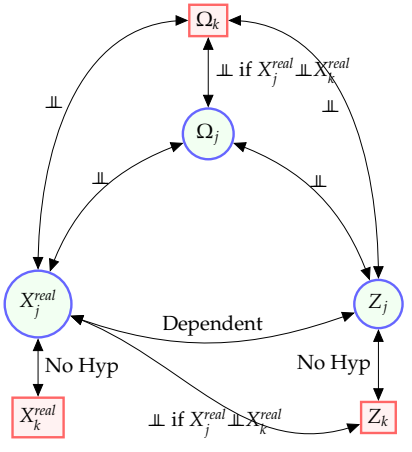
Several assumptions are looked through. They are linked with Rubin's nomenclature [17, Rubin, 1976], yet contested by [19, Seaman, 2013]. From the equation (1), different cases can be investigated depending on the correlation structure of ($\mathbf{X}^{real}$, $\mathbf{Z}$, $\mathbf{\Omega}$). Let consider the four following situations resumed in the Figures 2a, 2b, 2c, 2d. These cases depend only on the type of collection of each variable. Suppose that the information brought to the predictor from $\mathbf{Z}$ is not distinct from $\mathbf{X}^{real}$; $\mathbf{Z}$ is informative only through it correlation with $\mathbf{X}^{real}$ on $\mathbf{Y}$.



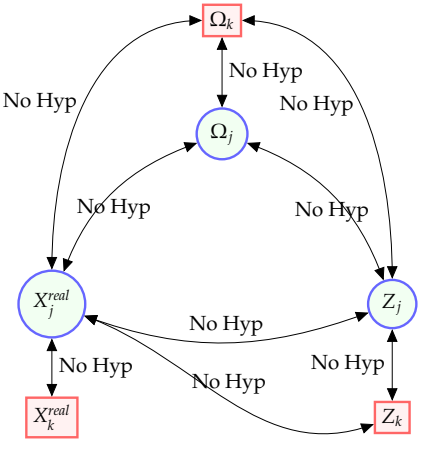

(a) Case (C1) - Total uncertainty ($j \neq k$). (No Hyp) means No hypothesis.




(b) (C2) - Local imprecision with unrelated errors.




(c) Case (C3) - Imprecision ($j \neq k$).




(d) Case (C4) ($j \neq k$).

**Figure 2:** Cases studied.

These assumptions can be linked with the missing value theory. For instance, the MCAR assumption ([17, Rubins 1976], [10, Heitjan and Basu, 1996]) can be seen as a particular case of the statement (C1) when the quality indexes are equal either to 0 or to 1. The multivariate independency between variables suggests that the errors are independent. In other words, each observation of each variable is gathered from different and unrelated sources or with unrelated errors. In the same way, MAR assumption is a particular case of (C2) and (C3). Indeed, it corresponds to some dependence between quality indexes/missing observations. In the case (C3), the wrong values $Z_j$ are correlated to the real values $X_j^{real}$. A particular case is when $(Z_j - X_j^{real})$ follows a centred distribution and is related to mismeasurement theory. The last case (C4) is closely linked to MNAR setting, where some dependence exists between each variable. In most of the cases encountered, $\Omega$ depends on specific values of $\mathbf{X}^{real}$. Therefore, the wrong values $\mathbf{Z}$ can depend on the real values $\mathbf{X}^{real}$; the errors are informative, which make the analysis more complex. The different cases are discussed in section 5.

**Remark 2.1.** *A discrete variable is considered a sum of boolean variables in regression. In between these boolean variables, the quality variables are equal. Hence, the case (C2) with fully correlated quality variables is a necessary assumption.*

## 3. Estimation Process

### 3.1 Reducing the error by mitigating on quality pattern

In this work, $\mathbf{X}$ is governed by the underlying process generating the covariates, as in the equation (1). In linear regression, the solution $\hat{\beta}$ minimizes the Residual Squared Error (RSE) calculated on the dataset $\mathbf{X}$. In GLM regression ([15, Nelder and Wedderburn, 1972]), it is the mean deviance $(1/n)Dev(\hat{\beta}|\mathbf{X}, Y)$ calculated on the dataset $\mathbf{X}$ which is minimized. Our particular framework enables to group two individuals $i$ and $i'$ having the same quality indexes (*i.e.* $\mathbf{Q}_i = \mathbf{Q}_{i'}$), which defines a quality pattern. Denote $P(\mathbf{Q})$ the set of all quality patterns present in the data. By taking it into account, the cost metric can be improved since

$$(1/n)Dev(\hat{\beta}|\mathbf{X}, Y) \geq (1/n) \sum_{K \in P(\mathbf{Q})} \sum_{i \backslash \mathbf{Q}_i = K} Dev(\hat{\beta}^K|\mathbf{X}_i, Y_i), \tag{3}$$

where $\hat{\beta}^K$ is the solution found on subset of the data with quality pattern $K$. The strategy to calculate these different coefficient is introduced in Section 3.2.

### 3.2 Prediction using quality index

This section studies in the sequel GLM given by

$$E[Y \mid \mathbf{X}^{real}] = g^{-1}(\mathbf{X}^{real}\beta),$$

and the likelihood associated $\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})$ using the real data set $\mathbf{X}^{real}$ [3].

In most cases, the previous model is unknown in our framework. Hereafter, this model is called **"Real" model**.

Denote the following naming :

[3] See the subsection 3.5 and the appendices.

- $M_2$ **("Naive" model)** : Model fitted on the observed dataset $\mathbf{X}$:

$$E[Y|\mathbf{X}] = g^{-1}(\mathbf{X}\beta^{M_2}),$$

where $\hat{\beta}^{M_2}$ the solution of $Argmax_\beta \mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})$. When $\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})$ is estimated using $\mathbf{X}^{real}$ and $\mathbf{Q}$, denote it $\mathcal{L}^{M2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q})$. Let write $\hat{\beta}^{M_2|\mathbf{X}^{real},\mathbf{Q}}$ the solution of $Argmax_\beta \mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q})$.

- $M_1$ **("Perfect quality" model)**: Model fitted on the observed dataset $\mathbf{X}$ which estimates the coefficient of the real model, $\beta$:

$$E[Y|\mathbf{X}, \mathbf{Q} = J_{n,p+1}] = g^{-1}(\mathbf{X}\beta^{M_1}).$$

In our framework, denote the solution $\hat{\beta}$ the solution of $Argmax_\beta \mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})$ and $\hat{\beta}^{M_1}$ is the solution of $Argmax_\beta \mathcal{L}^{M_1}(\beta; \mathbf{Y}|\mathbf{X}, \mathbf{Q})$ defined in the section 3.6. $\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})$ can not be determined in practice, since $\mathbf{X}^{real}$ is not fully observed;

- $M_3$ **("Pattern-adjusted" models)**: based on $\mathbf{X}$ and $\mathbf{Q}$, obtained from Algorithm 3 the models depend on each quality pattern:

$$E[Y_i \,|\, \mathbf{X}_i,\ K = (Q_{ij})_{1\leq j\leq p}] = g^{-1}(\mathbf{X}_i\beta^K),$$

where $K$ denotes the quality pattern associated to the individual $i$. In this work, notice that when $\mathbf{Q} = J_{n,1}K$, $\hat{\beta}^{M_2}$ estimates $\beta^K$.

For all the proofs, the variables are supposed centred.

### 3.3 Algorithm 3 for linear regression and GLM

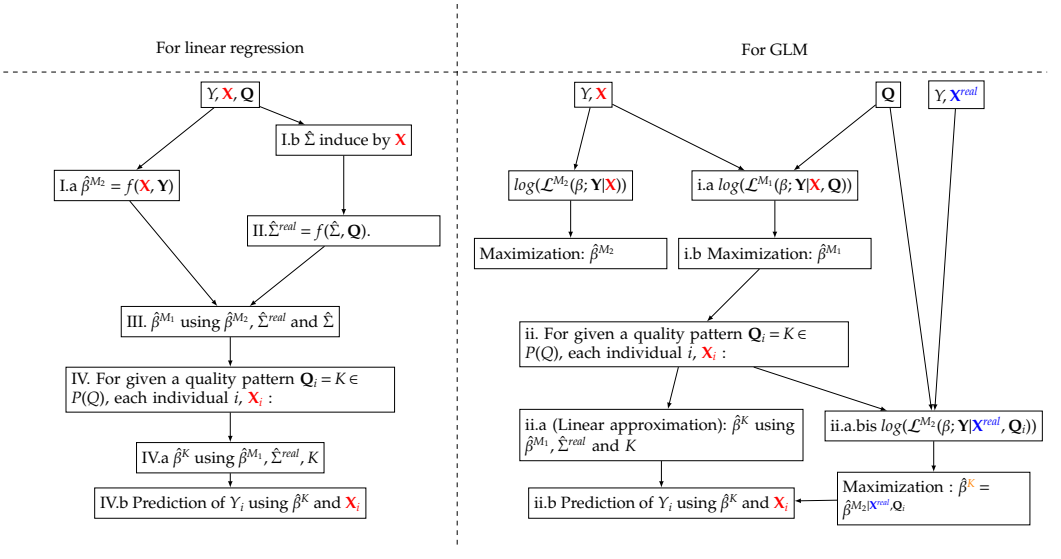



**Figure 3**: Process to take into account the quality index for linear regression and an approximation for GLM log-Poisson. In this way, this process adjusts the coefficient to each quality pattern.

For linear regression (see [4]), the algorithm associated with the Model $M_3$ is displayed in Figure 3. First, it assesses the Naive model $M_2$ from $\mathbf{X}$. Using the quality index $\mathbf{Q}$ and

the empirical correlation matrix $\hat{\Sigma}$, an estimator of the "perfect quality" correlation matrix $\Sigma^{real}$ is evaluated (see A.1). $\hat{\beta}^{M_1}$ is evaluated thanks to $\hat{\beta}^{M_2}$ and $\Sigma^{real}$. Finally, the algorithm provides $\hat{\beta}^K$ which minimizes the Residual Squared Errors for each pattern of quality $K$. Using $\hat{\beta}^K$, predictions depend on the characteristics of each individual and its quality.

For GLM regression, a similar method is suggested. To that end, the likelihood will be studied in place of the correlation matrix. However, the algorithm $M_3$ can not be applied as easily. No closed formula exists to link $\beta^{M_2}$ with $\beta$. Therefore, this work proposes to find $\beta^{M_1}$- an estimator of $\beta$ by maximizing an estimator of real model likelihood using $\mathbf{Q}$ and $\mathbf{X}$ (see section 3.6). Once $\hat{\beta}^{M_1}$ determined, I propose to use a linear correction to estimate $\hat{\beta}^K$. This approximation works well for small values of $\beta$ (see section 4.2).

In the event that $\mathbf{X}^{real}$ is known, or a large enough sample $\mathbf{X}$ is perfectly observed, $\hat{\beta}^K$ could be directly estimated from the maximization of the likelihood $\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q} = J_{n,1}K)$ (see section A.3). If the correlation structure of $\mathbf{X}^{real}$ is the same as $\mathbf{X}$ one, another solution would be to simulate a new $Y^{new}$ using $\mathbf{X}$ and $\hat{\beta}^{M1}$ to apply an estimator proposed in the subsection 3.6.

### 3.4 Assumptions under study

Assume that each covariate distribution has a finite second-order moment, and recall that $Z_j \sim X_j^{real}$ for $j = 1, ..., p$. Here, the discussion is about the assumptions underlying the correlation structure between the covariates $\mathbf{X}^{real}$, as well as for the random variables $\mathbf{Z}$. Let us thus define the five following assumptions:

*(X-A1)* All the random variables $X_j^{real}$ $(j = 1, ..., p)$ are independent.

*(X-A2)* Each variable $X_j^{real}$ is correlated with only one variable $X_k^{real}$ $(j \neq k)$.

*(X-A3)* For all $k \neq p$, the variable $X_k^{real}$ is independent of $X_p^{real}$ and $\bar{Q}_k = 1$.

*(Z-A1)* All the random variables $Z_j$ and $Z_k$ are independent.

*(Z-A2)*$(Z_j, Z_k)$ has the same correlation structure than $(X_j^{real}, X_k^{real})$, $j \neq k$.

For GLM, correlation between imperfectly observed covariates, such as (X-A2) are not considered. However, for linear regression, (X-A2) is taken into account in [4]. When the assumption (X-A3) is studied, denote $\mathbf{X}_{(*p)} = (1, X_1; \cdots; X_{p-1})$ and it's observed sample $\mathbf{X}_{i;(*p)}$. In the same way, $\beta^{(*p)}$ refers to $(\beta_0, \cdots, \beta_{p-1})$.

**Remark 3.1.** *The choice of the correlation structure of $\mathbf{Z}$ depends only on the data. Based on the same extraction and on the same key (e.g. geocoding), the correlation between two $Z_i$, $Z_j$ will be similar to $X_i^{real}$, $X_j^{real}$ ones for $i \neq j$ and $i, j \in \{1, ..., p\}$. In this case, (Z-A2) would be more appropriate. For errors completely independent, (Z-A1) would be preferred. In some other cases, the correlation structure might also differ, leading to different assumptions on $\mathbf{Z}$ dependency structure.*

### 3.5 The likelihood of the model with quality index

For actuarial pricing, most of the model used are GLMs. In the GLM case, the set of coefficient $\beta = (\beta_0, ..., \beta_p)^T$ is found by maximizing the likelihood or log-likelihood (ML-Maximum likelihood)[4];

$$\underset{\beta\in\mathbb{R}^p}{Argmax}\ \mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}) = \underset{\beta\in\mathbb{R}^p}{Argmax} \sum_{i=1}^{n} log(f_Y(Y_i|\mathbf{X}_i; \beta)), \tag{4}$$

where $\mathcal{L}$ is the likelihood function of the outcome $\mathbf{Y}$ given $\mathbf{X}$ and $\beta$ and $f_Y$ is the density function of $Y$.

Because the observations are independent and identically distributed, the previous log likelihood is the sample analogue of $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$. Assume mild regularity conditions (see A.3) for the proper convergence of our models. In our framework, these regularity conditions lead to the existence of the moment generating function for each imperfectly observed covariate. More detail on the theoretical part can be found in the appendixes. The subsection 3.6 explains how to find $\beta^{M_3|K}$ (*ii.* part of the $M_3$ algorithm) and the following part 3.7 focuses on estimating the "real" coefficient for log-Poisson GLM (*i.* part of the $M_3$ algorithm). The last part goes through different distributions and emphasizes the differences.

### 3.6 Deduce $\beta^{M_3|K}$

As already mentioned in Section 3.2, the vector $\beta^K$ exactly matches $\beta^{M_1}$ when all individualized quality indexes equal to 1, *i.e.* when $K = J_{1,p+1}$. In full generality, when $K = \mathbf{Q}_i$ is made of terms $Q_{ij} \neq 1$, the coefficients $\hat{\beta}^K$ need to be calculated. $\hat{\beta}^{K=\mathbf{Q}_i}$ is an estimator of $\beta^{M_2}$ when the model is fitted on dataset $\mathbf{X}$ but in the case $\mathbf{Q} = J_{n,1}\ \mathbf{Q}_i$. Therefore, the coefficient $\hat{\beta}^K$ is the one minimizing the mean $Dev(\hat{\beta}^K|\mathbf{X}, \mathbf{Y})$ for a given pattern of quality $K$ as wanted (see the equation 3).

For any distribution and link function, it is possible to estimate the expected $M_2$ log-likelihood for a given $\mathbf{Q}$ using $\mathbf{X}^{real}$ in the univariate case.

**Theorem 3.1.** *Let* $(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{real}, \mathbf{Q})$ *be the data sets as defined by equation 1. Suppose the assumption (C1) in the univariate case* $p = 1$. *Assume mild regularity assumptions, especially* $\int_{\mathbb{R}^2} |log(f_Y(y|z; \beta))|dF_{Z_1}(z)dF_Y(y) < \infty$ *for any value of* $\beta$. *Knowing* $(\mathbf{Y}, \mathbf{X}^{real}, \mathbf{Q})$, *a sample estimator of* $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$ *is*

$$\begin{aligned} &\bar{Q}_1 \sum_{i=1}^{n} log(f_Y(Y_i|X_{i1}^{real}; \beta)) \\ &+ (1 - \bar{Q}_1) \times \sum_{i=1}^{n} \frac{1}{n} \sum_{h=1}^{n} log(f_Y(Y_i|X_{h1}^{real}; \beta)). \end{aligned} \tag{5}$$

*This estimator converges almost surely and is denoted* $log(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|X^{real}, \mathbf{Q}))$. *The associated maximum likelihood estimator* $\hat{\beta}^{M_2|X^{real},\mathbf{Q}}$ *converges in probabilities into* $\beta^{M_2}$, *i.e.*

$$\hat{\beta}^{M_2|X^{real},\mathbf{Q}} \xrightarrow{\mathbb{P}} \beta^{M_2}.$$

The theorem can be easily extended to multivariate hypothesis (X-A3) and (Z-A1).

[4] Or equivalently minimize the deviance.

**Theorem 3.2.** *Under the assumptions (X-A3) and (Z-A1) and the same hypothesis as in the univariate case, the sample analogue of* $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta)))$

$$\begin{aligned}\bar{Q}_p \sum_{i=1}^{n} log(f_Y(Y_i|\mathbf{X}^{real}_{i;(*p)}, X_{i;p} = X^{real}_{i;p};\beta) \\ + (1-\bar{Q}_p) \sum_{i=1}^{n} \frac{1}{n} \sum_{h=1}^{n} log(f_Y(Y_i|\mathbf{X}^{real}_{i;(*p)}, X_{i;p} = X^{real}_{h;p};\beta)),\end{aligned} \tag{6}$$

*is consistent. The associated maximum likelihood estimator* $\hat{\beta}^{M_2|X^{real},\mathbf{Q}}$ *converges in probabilities into* $\beta^{M_2}$*, i.e.*

$$\hat{\beta}^{M_2|X^{real},\mathbf{Q}} \xrightarrow{\mathbb{P}.} \beta^{M_2}.$$

**Remark 3.2.** *In fact, for any correlation structure in between* $\mathbf{X}^{real}$*,* $\mathbf{\Omega}$*,* $\mathbf{Z}^{real}$*, an estimate of the expected likelihood of* $M_2$ *can be found easily through simulations. The only constraints needed are that mild regularity conditions must be verified under the chosen correlation structure.*

A downside of these methods is that $\mathbf{X}^{real}$ must be known, which is not always the case. Nonetheless, if $\mathbf{X}$ has the same the correlation structure than $\mathbf{X}^{real}$[5], a solution would be to simulate $Y^{new}$ from $\mathbf{X}$ using $\hat{\beta}$ and therefore calculate the previous estimator.

Both theorems permit to estimate $\beta^{M_2}$ through $\mathbf{X}^{real}$ for any $\mathbf{Q}$. $\hat{\beta}^{K=\mathbf{Q}_i}$ is an estimator of $\beta^{M_2}$ when the model is fitted on dataset $\mathbf{X}$ but in the case $\mathbf{Q} = J_{n,1}\ \mathbf{Q}_i$.

### 3.7 Deduce $\beta^{M_1}$ for log-Poisson GLM

This part will focus on Log-Poisson GLM under (X-A3) and (Z-A1). Estimators for other distributions or assumptions would be created exactly in the same way. Denote $V = (v_i)_{i=1,\dots,n}$ the exposure to have more traditional notations for count distributions. The exposure is supposed to be perfectly observed.

Remind that only $\mathbf{X}_p$ has a heterogeneous quality. Using the equation 34, an estimator of $log(\mathcal{L}(\hat{\beta};\mathbf{Y}|\mathbf{X}^{real},\mathbf{Q}))$ can be found as follows :

$$\begin{aligned}log(\mathcal{L}^{M_1}(\hat{\beta};\mathbf{Y}|\mathbf{X},\mathbf{Q})) = \frac{1}{\bar{Q}_p}\Big[log(\mathcal{L}^{M2}(\hat{\beta};\mathbf{Y}|\mathbf{X}) \\ - (1-\bar{Q}_p) \times log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p)})) \\ - (1-\bar{Q}_p) \times \sum_{i=1}^{n} v_i e^{\hat{\beta}^{*p}\mathbf{X}^{real}_{i,(*p)}}(1 - M_{X_p}(\hat{\beta}_p))\Big].\end{aligned} \tag{7}$$

All the right terms are known and can be evaluated. Indeed,

- $log(\mathcal{L}^{M_2}(\hat{\beta};\mathbf{Y}|\mathbf{X}))$ is the $M_2$ model log-likelihood using all the covariates;
- $M_{X_p}(\hat{\beta}_p)$ can be estimated or for particularly distributions, given the distribution parameters, the moment generating function is explicitly known ;
- $log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p)}))$ is the $M_2$ model log-likelihood using all the covariables except for $\mathbf{X}_p$; under the assumption (X-A3), $log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p)}))$ is equal to $log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}_{(*p)}))$.

[5] i.e in the (C1) case under (X-A1) and (Z-A1) or in the (C2) case with fully correlated quality variables and (Z-A2).

In the same spirit, another estimator can be put forward as a sum of the previous estimator conditioned by pattern of quality $K_p$ :

$$\begin{aligned} log(\mathcal{L}^{M_1}(\hat{\beta};\mathbf{Y}|\mathbf{X},\mathbf{Q})) = \sum_{K_p \in P(Q_p), K_p \neq 0} \frac{1}{K_p} \Big[ & log(\mathcal{L}^{M2}(\hat{\beta};\mathbf{Y}|\mathbf{X}_{Q_p=K_p}) \\ & - (1-K_p) \times log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p);Q_p=K_p})) \\ & - (1-K_p) \times \sum_{i=1}^{n} v_i e^{\hat{\beta}^{*p}\mathbf{X}^{real}_{i;(*p);Q_p=K_p}} (1 - M_{X_p}(\hat{\beta}_p)) \Big]. \end{aligned} \tag{8}$$

where $\mathbf{X}_{Q=K_p}$ represents the dataset where only individual $i$ such as $Q_{i;p} = K_p$ are kept. The second estimator $log(\mathcal{L}^{M_1}(\hat{\beta};\mathbf{Y}|\mathbf{X},\mathbf{Q}))$ from the equation 8 is often more precise by construction than the equation 7. Individual having null quality index are not taken into account. Therefore, in the following part, the second estimator will be used. These two estimators converge in probabilities to $log(\mathcal{L}(\hat{\beta};\mathbf{Y}|\mathbf{X}^{real},\mathbf{Q}))$. In the same way, the solution of the maximum likelihood converges in probability.

*Optimization program:* On the contrary of the classical optimization method: the iterative weighted least square algorithm used to fit GLM parameters can not be used. Empirically, the Nelder-Mean optimization from the *optim* function from *stats* package (R software) seems to have a more stable convergence than Newton-Raphson algorithm.

Indeed, for some distributions, the moment generating function may not exist or has extremely high value for some values of $\hat{\beta}_p$. In this case, the estimated derivative may be important. For these reasons, Newton-Raphson method can lead to important staring oscillations depending on $\hat{\beta}_p$ and $\mathbf{X}_p$ distribution. This is why Nelder-Mean optimization is here preferred and starting at $\hat{\beta}_p = 0$.

### 3.8 Estimators proprieties for other distributions

| Model GLM | Case | | Hyp $\mathbf{X}^{real}$ | Hyp $\mathbf{Z}$ | Estimator convergence | Log-likelihood ($M_1$) |
|---|---|---|---|---|---|---|
| Log-Gaussian | (C1)-(C2) | $p=1$ | No Hyp | No Hyp | $\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$ and $\frac{\hat{\beta}_1^{M2}}{Q_1} \xrightarrow{\mathbb{P}.} \beta_1$ | - |
| | (C1) | $p>1$ | X-A1 | Z-A1 | $\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$,<br>$\frac{\hat{\beta}_j^{M2}}{Q_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j=1,...,p.$ | - |
| | (C1) | $p>1$ | X-A3 | Z-A1 | $\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$,<br>$\frac{\hat{\beta}_j^{M2}}{Q_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j=1,...,p.$ | - |
| | (C2)<br>$\Omega_j = \Omega_k$ | $p>1$ | No Hyp | Z-A2 | $\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$,<br>$\frac{\hat{\beta}_j^{M2}}{Q_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j=1,...,p.$ | - |
| Log-Poisson | (C1)-(C2) | $p=1$ | No Hyp | No Hyp | $\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$ and $\hat{\beta}_1^{M_2} \xrightarrow{\mathbb{P}.} [0;\beta_1]$ | Yes |
| | (C1) | $p>1$ | X-A3 | Z-A1 | $\hat{\beta}_j^{M_2} \xrightarrow{\mathbb{P}.} \beta_j, \quad j=0,...,p-1$<br>$\hat{\beta}_p^{M_2} \xrightarrow{\mathbb{P}.} [0;\beta_p]$. | Yes |
| | (C2)<br>$\Omega_j = \Omega_k$ | $p>1$ | No Hyp | Z-A2 | - | Yes |
| Log-Gamma | (C1)-(C2) | $p=1$ | No Hyp | No Hyp | $\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$ and $\hat{\beta}_1^{M_2} \xrightarrow{\mathbb{P}.} [0;\beta_1]$ | Yes |
| | (C1) | $p>1$ | X-A3 | Z-A1 | - | Yes |
| | (C2)<br>$\Omega_j = \Omega_k$ | $p>1$ | No Hyp | Z-A2 | - | Yes |
| Inv-Gamma | (C1) | $p>1$ | X-A3 | Z-A1 | - | No |
| Probit | (C1) | $p>1$ | X-A3 | Z-A1 | - | No |

Previous table shows different results for different GLM and assumptions. For the most common GLM used in non-life pricing (Log-Gaussian, Log-Poisson and Log-Gamma GLM), some interesting results can be found thanks to the additive or multiplicative structure. However, for Probit or Inv-Gamma GLMs, no explicit formulas can be found without approximation.

**Log-Gaussian GLM:** Log-Gaussian GLM's structure leads to explicit relation between $\beta$ and $\beta^{M2}$. Therefore, the $M_1$ log-likelihood is not needed to be calculated. Because Log-Gaussian GLM and linear regression are equivalent, the same results can be state. It is important to notice that $\beta_j^{M2}$ only depends on $Q_j$ and $\beta_k$ and $Q_j$ for all $X_k$ correlated to $X_j$ in Log-Gaussian case. In other words, the coefficient of a variable is not skewed by the quality of variables not correlated to it. In the case (C2) with fully correlated quality variable under (Z-A2), *i.e.* $\Omega_j = \Omega_k$ for all $j$ and $k$, Log-Gaussian coefficients $\beta_j^{M_2}$ have a simple affine linearship with $\beta_j^{M_1}$ for $j = 1, ..., p$.

**Multiplicative structures:** In the case (C1), Log-Poisson and Log-Gamma GLM's multiplicative structure provides the calculus of $log(\mathcal{L}^{M1}(\beta; \mathbf{Y}|\mathbf{X}, \mathbf{Q}))$. However, in multivariate case, under (X-A3) and (Z-A1), $\beta_p^{M_2}$ depends on $Q_p$ , the moment generating function $M_{X_p}(t)$ and $\beta_j$ for $j = 1, ..., p-1$. The main difference is that $\beta_j^{M_2}$ depends on the distribution of $X_p$. For Log-Poisson model, $\beta_j^{M_2} = \beta_j$ and $\beta_p^{M_2}$ convergences in probability in an interval $[0, \beta_p]$. However, for Log-Gamma this property is not true and the other coefficients, $\beta_j^{M2}$, are also impacted by $Q_p$. In the case (C2), regrettably, no proprieties on the estimator can be state for Log-Gamma and Log-Poisson GLM.

## 4. Simulation study - M1 estimator

This section aims to check our theoretical results on the estimator properties for Log-Poisson GLM. In this view, all the simulated examples are created using the following steps involving all the aforementioned quantities required to generate the right data :
**Step 1:** $Q$ is in practice given. For the simulation, it is randomly generated;
**Step 2:** $\mathbf{X}^{real}$ is simulated given the marginals and the correlation structure;
**Step 3:** $\mathbf{Z} = (Z_1, ..., Z_p)$ is simulated given $\mathbf{X}^{real}$ and the assumptions;
**Step 4:** $Y$ is simulated from its relationship with $\mathbf{X}^{real}$;
**Step 5:** $\Omega$ is simulated from $Q$ through Bernoulli trials;
**Step 6:** $\mathbf{X}$ is deduced thanks to the equation (1).

The study is performed using R ([16]) statistical software.

### 4.1 Find $\hat{\beta}^{M_1}$ coefficients

Let $\mathbb{E}(Y|\mathbf{X}^{real}) = exp(1 + 0.4X_1^{real} + 0.5X_2^{real} + 0.6X_3^{real} + 0.07X_4^{real})$ with $X_1 \sim \Gamma(2, 1), X_2^{real} \sim \mathcal{N}(0, 1), X_3 \sim \mathcal{P}ois(2)$, $X_4 \sim \mathcal{N}(0, 10)$ and $Y$ following a Poisson distribution. The quality index follows an independent discrete distribution on the values (0.5; 0.75; 1) with the probability (0.25; 0.25; 0.5) for $Q_4$. Let all the other covariates be perfectly observed, *e.g.* $Q_{i,j} = 1$ for all $i \in 1, ..., n$ and $j \in \{1, 2, 3\}$.

Using the precedent result, $M_1$ likelihood can be estimated as shown in figure 4. The use of imperfectly observed dataset implies a wider variance of the estimator $M_1$ than the real model one. Here, the first estimator has wider variance than the second estimator. As shown by equation 37, the coefficients $\beta_1$, $\beta_2$ and $\beta_3$ did not change due to the independence in between the variables - figure 5 - and the coefficient associated to $X_4$ is corrected - figure 6.

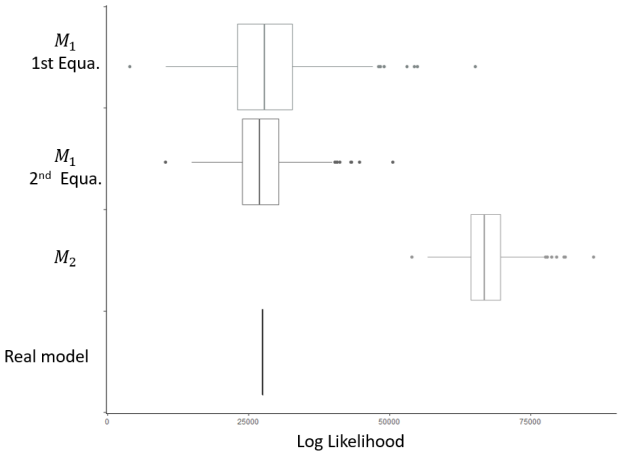



**Figure 4**: Estimation of the $M_1$ log-likelihood for log-Poisson GLM using equation 7 for a given $\mathbf{X}$ and $\mathbf{Q}$. The moment function is estimated using its empirical estimator. The true function leads to the same graph but with a smaller variance. 2000 simulations are done for a given $\mathbf{X}^{real}$ and $\mathbf{Q}$.

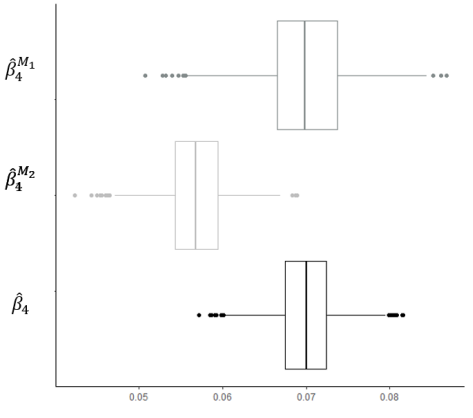



**Figure 5**: $\hat{\beta}_4^{M_2}$ is smaller than $\hat{\beta}$ because of the quality of the variable. $\hat{\beta}_4^{M_1}$ is unbias, but has a wider variance than the real coefficient.

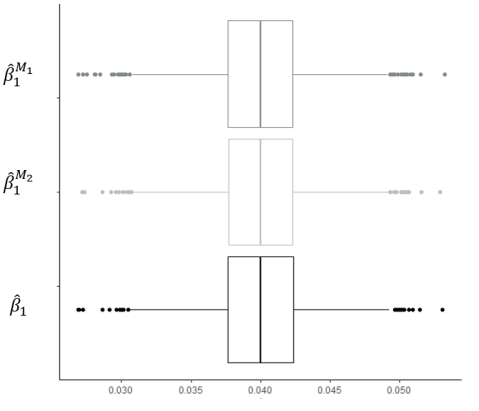



**Figure 6**: The quality of the variable $X_4$ does not impact the estimation of $\beta_1$, $\beta_2$, $\beta_3$; here, highlighted by $\beta_1$ with $X_1^{real}$ standard normal distribution and $Y$ following a Poisson distribution. Other distributions of $X_1^{real}$ have also been tested and lead to the same results. 2000 simulations are done for a given $\mathbf{Q}$.

### *4.2 Adapt the coefficient to the quality*

Unlike linear regression, no explicit relation exists between the $\beta$ and $\beta^{M_2}$ or $\beta^K$ in function of the quality. It has been shown that the coefficient is a barycenter of the $\hat{\beta}^{M_1}$ and 0. Moreover, $\hat{\beta}_p^{M_2}$ converges to 0 when $Q_p$ tends to 0. I suggest using the linear approximation, *i.e.* $\hat{\beta}_p^{K=Q_p} = Q_p \times \hat{\beta}_p^{M_1}$. Indeed, as shown on the figure 7, for small values of $\beta_4$ ($\approx 0.07$), the

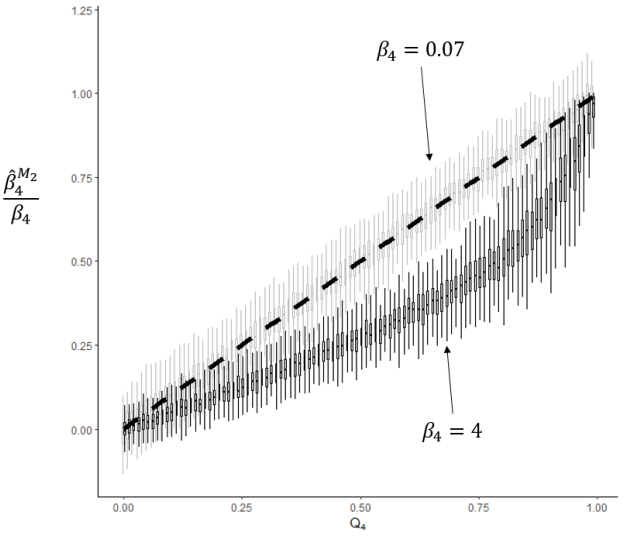



**Figure 7**: 1000 simulation for each quality. As for linear regressions, a linear evolution through the quality can be seen for low coefficient, however for higher values the relationship is not proportional to the quality.

impact of the moment on the likelihood is lower than for higher value of $\beta_4 = 4$. Therefore, the coefficient could be estimated linearly only for $\mathbb{E}(Y)$ small, but would overestimate the coefficient for higher values. In household insurance for individual, this assumption is adapted to the low annual frequency of claims.

## 5. Discussion

In this discussion, the adequacy of different assumptions and hypothesis will be evaluated. The following example comes straight up from a project on household pricing using geolocated addresses to add external data. First, the subsection 5.1 gives proper examples encountered and justifies the different hypothesis needed in our framework in the subsection 5.2. In our example, issues remain, such as imperfect quality index, its' evaluation and correlations in between variables. Sections 5.3 and 5.4 emphasize the limits and propose some solutions using interactions.

### *5.1 Examples*

In this part, the different cases are discussed under the scope of a pricing case using house geolocation. Here, the goal is to model the frequency or the claim cost of a household insurance using only the address and external data. As explained in the introduction, our particular framework is adapted to this problematic. To find the different covariates associated to characteristics of the individual, the first step is to link the address with its geocoding, then to link the geocoding to the right parcel or/then with the building. Then by geolocating external data and calculating characteristics from picture analysis or other predicting method, a database is created. The variable to model is given by insurers departments. It corresponds to the frequency or claims cost and is supposed to be perfectly observed.

Here, the collected data's quality is mainly looked through the credibility dimension. If the geocoding is wrong, all the observations would be taken on another building. The consistency of the variable and the way it is collected change also the data quality. Moreover, the reliability of predicted characteristics depends also on the reliability of covariables used in the predictive model.

Let discuss the different assumptions on the example of pricing of home insurance using geocoding.

**Example of case C1:** The collection of the variables, the presence of pool and presence of solar panels can fit the description. Suppose that the pool variable collection uses a governmental data set based on inhabitants' declaration and the solar panels variable uses the geocoding to determine pictures to analyse. The collection of the two variables are not correlated. The case (C1) and the assumption (Z-A1) would be appropriate. Indeed, if one is wrongly observed, it does not induce the other one to be and the errors are not linked with the variable value, *i.e.* $\mathbf{Q}$, $\mathbf{X}^{real}$ and $\mathbf{Z}$ are independent.

**Example of case C2:** The living surface, the number of rooms and the footprint are globally one of the most segmenting features in household pricing. Different data sets and methods are available in France to collect them, such as DVF [6]. This database geolocates parcels and contains different features such as the value of property values, the number of rooms, the surface of the parcel or the living surface among others. The database is created from all properties transfer since 2015. e. On the uncertainty dimension, errors are coming from the link between geocoding and the address or between the address and the building, each of these steps impacts the data's quality depending on the feature. A wrong geocoding would imply that the observations are taken from another building. For all these variables, the case (C2) and the assumption (Z-A2) would be appropriate since they are collected from the same building.

**Example of case C3:** The previous example acts also on the mismeasurement dimension, where $\mathbf{Z}$ and $\mathbf{X}^{real}$ are correlated. Data quality, impacted by the consistency of the collection of the database, infers on it due to the timeless dimension; houses might have changed since the last property transfer. Indeed, precision of the house's size may be bias after expansion of a house if the database is not updated in the meantime. Moreover, correlations between $\mathbf{X}^{real}$ and $\mathbf{Z}$ come also from the way that variables are collected; the best example is spatial correlation. For instance, let look into a variable informing on the floors' number being collected from pictures analysis. The impact of geocoding uncertainty is not globally the same as before. Indeed, neighbour's houses have often the same height or number of floor. Then, even if the collection of the data is done on the wrong building, $\mathbf{Z}$ will be correlated with $\mathbf{X}^{real}$.

**Example of a case C4:** All variables mentioned earlier can fit in this category due to spatial correlation. In fact, rarely in our study, the quality variable does not depend on if the building is from rural areas or urban areas. Moreover, if in megalopolis the detection of the house size may be difficult due to the building's density, a systematic uncertainty could appear on this variable for urban houses - globally smaller. Then $\mathbf{Z}$ would be correlated with the $\mathbf{X}^{real}$ automatically, but also with $\mathbf{Q}$. The same analysis could be done on high buildings, *e.g.* for the number of floors.

One of the most difficult cases is when the quality depends on others variables; for instance the material of the roof and the analysis of a roof to detect a window - see figure 8 and 9. In this case, the modality of dark slate informs on the risk, not because dark slate changes it but due to the low quality of the variable roof-windows associated to it.

[6] This database comes from a certified public service documenting the property values declared during property transfers available in open data

**Figure 8**: The detection of a window on a roof is immediate from the IGN cartography (47.183722, -1.812768) - © IGN 2018.

**Figure 9**: The detection of a window on a roof is harder because of the dark slate roof from the IGN cartography (47.179068, -1.814216) - © IGN 2018

### 5.2 Actuarial justification of this framework assumptions

**Integrity of dataset and assumption on $Z$:** In all examples seen above, the wrong observations $\mathbf{Z}$ are coming from real individual; it justifies the main assumption that "wrong" values $\mathbf{Z}_j$ follow the same distribution as $\mathbf{X}_j^{real}$ (equation 1). However, the assumption is true only if the integrity of the data set is valid. Indeed, for instance, if some wrong observations are taken from commercial buildings or flats when pricing residential household insurance, this assumption would not be verified.

**Assumption (X-A3):** The assumption (X-A3) is a very restrictive assumption. Nonetheless, it can be appropriate for underwriting used. First, the use of several imperfectly observed covariates is not recommended and not adapted when aiming to a stable model. Moreover, traditional covariates used are well-known covariates of good quality, so one or two variables with heterogeneous quality would in practice be integrated at the most. Adding some imperfect variables correlated to others also bias the coefficients of these perfectly observed variables.

**Use of the linear approximation to find adapted model:** As shown in section 4, linear approximation can be a good approximation for small values of the coefficients. In other words, the approximation can be valid when the claim count modelling is done at the individual case. Indeed, in household insurance, the mean damage frequency is around 1 % (for instance, water damage or fire damage coverage.) The other benefit is that only one model is fitted.

**Add a new variable in a pricing:** Lastly, our framework can be used to estimate $\beta$ for a new covariate. Without a data set and claims associated to it, the observations of this new variable have to be determined using external data or models. Indeed, it is impossible to request a completely new information once the contract signed. However, a question can be added in underwriting questionnaire during a quotation and therefore the covariate can be used in the new pricing model. Logically, information from underwriting questionnaire are much better quality and are often suppose perfectly observed (for most of the variables). So pricing models muss use $\beta$, adapted to perfectly observed variables, and not $\beta^{M2}$.

### 5.3 Use interactions with quality indexes

The different results also help to understand how to deal with a finite number of quality groups within a variable. one could propose to use interaction instead of this paper's framework. Indeed, the quality effect could be taken into account by adding an interaction

between the $Q_j$ and the $X_k$, $k \neq j$. Denote the following log-Gaussian GLM :

$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{9}$$

and $\rho$ the correlation between the two covariates. Suppose that the data set has another variable $Q_1$ with two modalities (High and Low) informing on the quality of $X_1$. From the results earlier, adding some interactions between $X_1$ and $Q_1$ only, i.e,

$$\mathbb{E}[Y|\mathbf{X}, Q_1] = \mathbf{1}_{Q_1=L}(\beta_0^{Q_1=L} + \beta_1^{Q_1=L} X_1) + \mathbf{1}_{Q_1=H}(\beta_0^{Q_1=H} + \beta_1^{Q_1=H} X_1) + \beta_2 X_2 \tag{10}$$

would be the best option only if they is no correlation. The interaction should be on both variables :

$$\mathbb{E}[Y|\mathbf{X}, Q_1] = \mathbf{1}_{Q_1=H}(\beta_0^{Q_1=H} + \beta_1^{Q_1=H} X_1 + \beta_2^{Q_1=H} X_2) + \mathbf{1}_{Q_1=L}(\beta_0^{Q_1=L} + \beta_1^{Q_1=L} X_1 + \beta_2^{Q_1=L} X_2). \tag{11}$$

Obviously, with more covariates and quality indexes, it adds a lot more parameters to fit exactly $n \times 2^{h-p}$ where $h$ is the sum of modalities' number of each quality index. Moreover, the coefficients $\hat{\beta}_2^{Q_1=H}$ and $\hat{\beta}_2^{Q_1=L}$ could have different signs (see [4] or the appendix). For other distributions, the whole issue is much more complex. Therefore, in such case, limiting the correlation in between variables should be the priority.

### 5.4 Determine quality indexes and the impact of imperfect quality indexes

In a pricing data set studied, the quality index was given as an ordered variable with the following modality ("very low", "low", "medium", "high","very high"). Would it be possible to determine the equivalent quality index by modality ?

By evaluating a model by modality, quality indexes can be easily found given baseline coefficients - by example $\beta$ (known or evaluated thanks to the best quality points). The difficulty resides in the way that the quality is given. By fitting an univariate linear model with variables centred and an interaction between $X_1$ and the variable representing the quality $K(X_1)$ with $M \in \mathbb{N}$ modalities,

$$E[Y|\mathbf{X}, K(X_1)] = \beta_0^{M2} + \sum_{m=1,\dots,M} \beta_1^{M2,K(X_1)=m} X_1 \mathbb{1}_{K(X_1)=m}, \tag{12}$$

each quality index modality can be evaluated. Indeed, let assume that the modality $K(X_1) = 1$ corresponds to perfect quality observations, the quality index value of the modality $m$ is equal to $Q_m = \beta_1^{K(X_1)=m} / \beta_1^{K(X_1)=1}$.

Figure 10 shows a real example of a quality index assessment. The model used is an univariate log-Poisson GLM using only the living surface to predict a water damage frequency. The values of living surface is at first coming from labels using DVF by associating a building to property sale. To complete the missing information, predicting methods are done using the house characteristics. If the confidence into the database geocoding is perfect, the confidence associated is "very high". Otherwise, the confidence is degraded depending on the reliability of the geocoding of the property sales database. On the other hand, predicted values are associated with a maximum of "high" (in majority "medium"). The credibility is degraded depending on the quality of the covariates and the score associated to each result. Two filters are considered on the addresses' geolocation to link the claims and these characteristics : a filter keeping all the building considered as the main one on the parcel and a second keeping only the building if it is link only to one address. Figure 10 helps to evaluate the quality indexes values. Supposing $\beta^{High-One\ adresse} = \beta$ perfect. The value of each $Q$ can be approach by $\frac{\hat{\beta}^Q}{\beta}$ using a linear approximation. Remind that the annual frequency of water damage is low: around 3 per 100.

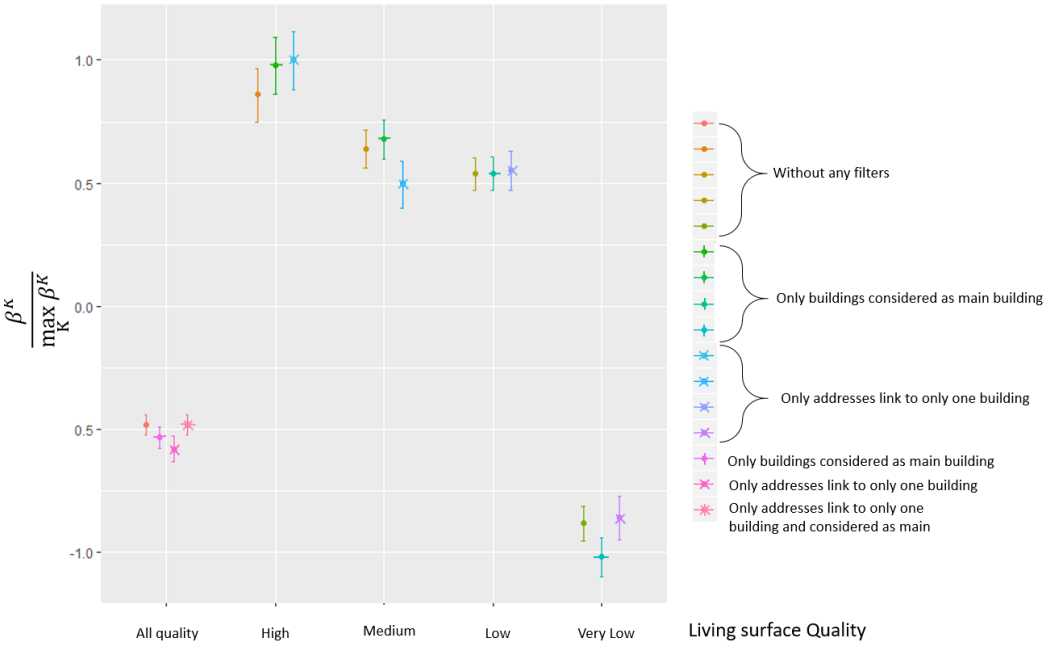



**Figure 10**: Ratio of the living surface coefficient of Log-Poisson. Because they are too few "very high" quality observations, they were regrouped with "high" quality observations. Different filters based the building geolocation information are done on the data set to challenge the quality index.

Then, "medium" quality value would be estimated by 0.6, "low" quality value by 0.5. However, the coefficient of very low quality values has an opposite sign. In fact, very low quality values are link to rural zone. Therefore, the case (C4) is the most appropriate and our evaluation method can not be used. In the same way, "low" quality values are also more link to rural density than "medium" one or "high"[7]. In consequence, the associated value to medium quality 0.5 can be debated. Indeed, the "low" and "very low" quality are correlated with others characteristics impacting the risks. The coefficients calculated on the database are therefore impacted. This is an important limit of our framework. In such case, using a threshold to set aside "very low" quality observations is recommended so that the data set verifies our assumptions. The different filters on geocoding show that leaner detail could be added within a value of the quality index. In this case, a modality may regroup different levels of quality. In other word, quality index is not perfectly determined.

In practice, a modality might regroup observation of different quality. This part considers the case of a modality regrouping two types of observations with distinct quality. Denote $m$ a modality of $n$ observations which regroups $n_\alpha$ and $n_\kappa$ number observations with the quality $\bar{Q}_\alpha$ and $\bar{Q}_\kappa$ respectively ($n_\alpha + n_\kappa = n$). The difference between model's coefficients and the real model ones can be expressed as a barycenter of the sum of the group's quality under (X-A1):

$$\beta_1^{M2,\bar{Q}_m} - \beta_1 = \frac{n_\alpha}{n}(\bar{Q}_\alpha - 1)\beta_1 + \frac{n_\kappa}{n}(\bar{Q}_\kappa - 1)\beta_1. \tag{13}$$

[7]Here, the mismeasurement side of the data quality is set aside.

Equation 13 can be easily extended to higher dimension. If groups of different quality are mixed together and are given the same quality index value, the best one should be the pondered mean of each quality in a context of linear regression with the assumption (X-A1). However, under (X-A2) (with correlation), the aggregation of the quality influences the coefficients value of other correlated covariates.

**Proposition 1** For log-Gaussian GLM, under assumptions (X-A2) and (Z-A1), given $k$ and $j$ such as $\rho_{kj}^{real} = \rho$, $Q_k \neq 0$ and $Q_j \neq 0$, if $\rho\beta_k^{M_1} \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j^{M_1}$,

$$\begin{aligned} \beta_k^{M_2} :]0,1] &\to \mathbb{R} \\ Q_k &\mapsto \beta_k^{M_2}(Q_k|Q_j). \end{aligned} \tag{14}$$

is an increasing convex function. Otherwise, it is decreasing concave.

Therefore, the weighted mean of the quality is a biased approximation. Indeed, accordingly to the Proposition 1, if $\rho\beta_k \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j$, for $i \neq j$ :

$$\forall Q_\alpha, Q_\kappa \in [0,1], \quad \beta_k^{M_2}(\frac{n_\alpha}{n}\bar{Q}_\alpha + \frac{n_\kappa}{n}\bar{Q}_\kappa) \leq \frac{n_\alpha}{n}\beta_k^{M_2}(\bar{Q}_\alpha) + \frac{n_\kappa}{n}\beta_k^{M_2}(\bar{Q}_\kappa). \tag{15}$$

In consequence, regrouping two groups of different quality bias the coefficient accordingly to the correlation. The equivalent quality index in linear regression under this assumption should be lower than the pondered mean of the quality. Because the convexity depends on the correlation, the pondered mean of the quality may be a fine approximation with low correlation between covariates.

## 6. Conclusion

This paper extends a method to take into account index quality on the credibility dimension for GLM regression. In pricing, it could correspond to an external score when open/external data are added to a traditional dataset. Moreover, as for Rubin's nomenclature, different cases exist depending on the relation structure between qualities indexes, real observations and wrongs one. Relaxing the different assumptions, especially some hypothesis between quality variable and the variable, will be the next step. These results are very useful for actuaries which are in charge of the data quality they use and models following. The different cases have been discussed under a real pricing using the geolocated addresses to find external information. Finally, actuaries should keep in mind that they are answerable of the data quality they use. Therefore, this work suggests a method to evaluate data quality and put forwards recommendation with data quality indexes in use.

To use data's quality index with correlated covariates, further research is ongoing to adapted decision trees to this use and to release assumptions between quality variable and the true values. Several issues remain generalizing for penalized likelihood optimization and quality index evaluation.

**Acknowledgements.** The author thanks the firm and the data provider which have allowed him to use the portfolio, to geolocate it and to create this data set for this paper.

## Conflict of interest

The author declares no conflict of interest.

## References


[1] Autorité de contrôle prudentiel, ACPR. Synthèse de l'enquête déclarative de 2019 sur la gestion des données alimentant les calculs prudentiels des organismes d'assurance. Technical Report 119, ACPR, Jan 2021.

[2] Ole Barndorff-Nielsen. Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics*, pages 151–157, 1978.

[3] Robert Campbell, Louise Francis, V Prevosto, Mark Rothwell, and Simon Sheaf. Report of the data quality working party. Technical report, 2006.

[4] Pierre Chatelain and Xavier Milhaud. Linear regression and data quality through individualized credibility index. preprint, May 2021.

[5] R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[6] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981.

[7] Louise A Francis. Dancing with dirty data methods for exploring and cleaning data. 2005.

[8] General Committee of the Actuarial Standards Board and Applies to All Practice Areas, GCASB. Data quality - revised edition. Technical Report 23, Actuarial Standard of Practice, May 2014.

[9] Ali S Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):761–771, 1992.

[10] Daniel F Heitjan and Srabashi Basu. Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3):207–213, 1996.

[11] Sham Kakade, Ohad Shamir, Karthik Sindharan, and Ambuj Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 381–388. JMLR Workshop and Conference Proceedings, 2010.

[12] Kananart Kuwarannancharoen and Shreyas Sundaram. On the location of the minimizer of the sum of two strongly convex functions. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1769–1774. IEEE, 2018.

[13] EL Lehmann and George Casella. Unbiasedness. *Theory of Point Estimation*, pages 83–146, 1998.

[14] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[15] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[17] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[18] Shaun Seaman and Ian White. Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, 43(16):3499–3515, 2014.

[19] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

[20] Myriam Tami, Marianne Clausel, Emilie Devijver, Adrien Dulac, Eric Gaussier, Stefan Janaqi, and Meriam Chebre. Uncertain trees: Dealing with uncertain inputs in regression trees. *arXiv preprint arXiv:1810.11698*, 2018.

[21] Ion-George Todoran, Laurent Lecornu, Ali Khenchaf, and Jean-Marc Le Caillec. Toward the quality evaluation of complex information systems. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, volume 9091, page 90910N. International Society for Optics and Photonics, 2014.

[22] Asma Trabelsi, Zied Elouedi, and Eric Lefevre. Handling uncertain attribute values in decision tree classifier using the belief function theory. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 26–35. Springer, 2016.

[23] Stef Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.

[24] Sabine Van Huffel and Philippe Lemmerling. *Total least squares and errors-in-variables modeling: analysis, algorithms and applications*. Springer Science & Business Media, 2013.

[25] Andrea Vedaldi, Hailin Jin, Paolo Favaro, and Stefano Soatto. Kalmansac: Robust filtering by consensus. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 633–640. IEEE, 2005.

[26] Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.

# Appendix A. Theoretical framework

## A.1 The covariance impacted by quality index

Given a data set with two covariates and their joint quality $(X_j, Q_j)$, $(X_k, Q_k)$, $j \neq k$ as in the equation (1), the following lemma states the relation between real covariance $Cov_{jk}^{real}$ and the observed covariance $Cov_{jk}^{real}$ under various assumptions :

**Lemma 1.** *In the case (C1), the relation yields*

$$\textbf{\textit{Under (Z-A1)}}\ Cov_{jk} = Q_j Q_k \times Cov_{jk}^{real}. \tag{16}$$

$$\textbf{\textit{Under(Z-A2)}}\ Cov_{jk} = (1 + 2Q_j Q_k - Q_j - Q_k) \times Cov_{jk}^{real}. \tag{17}$$

*In case (C2), if $X_j^{real}$ and $X_k^{real}$ are independent, both (16) and (17) hold true. Otherwise, if the joint quality are completely and positively dependent, the following results yield :*

$$\textbf{\textit{Under (Z-A1)}}\ Cov_{jk} = Q_j \times Cov_{jk}^{real}, \tag{18}$$

$$\textbf{\textit{Under (Z-A2)}}\ Cov_{jk} = Cov_{jk}^{real}. \tag{19}$$

The proof of the case (C1) is available in [4, Chatelain and Milhaud, 2021]. The proof of the case (C2) is a trivial extension of the precedent. In case (C3), an additional term corresponding to the correlation between "wrong" value and the "right" one would appear. The results could therefore be extended to such cases both under (Z-A1) or (Z-A2), but one would need to specify the correlation structure between $\mathbf{X}^{real}$ and $\mathbf{Z}$. Because each covariate $\mathbf{X}^{real}$ and $\mathbf{Z}$ have the same distribution, $Var(X_j) = Var(X_j^{real}) = Var(Z_j)$. Therefore, the same relation between Pearson's correlation is also true. Thanks to Lemma 1, $\Sigma^{real}$ can be evaluated from $\mathbf{Q}$ and $\Sigma$.

## A.2 Regression model under consideration

Given the independent variables $(Y_1, ..., Y_n)$, the corresponding explanatory variables $(\mathbf{X}_1, ..., \mathbf{X}_n)$, and individualized quality indexes $(\mathbf{Q}_1, ..., \mathbf{Q}_n)$ where $\mathbf{Q}_i = (Q_{i1}, ..., Q_{ip})$, this part will study Generalize Linear Model (GLM). GLM is defined by three components : the distribution's response variable $Y$ which is from the exponential family, a linear predictor $\mathbf{X}\beta$ and a link function $g$ defined such as

$$\mu = E[Y|\mathbf{X} = \mathbf{x}] = g^{-1}(\mathbf{x}\beta). \tag{20}$$

where $\mathbf{X}$ is the vector of covariates including a constant (see Section 2.1), and $\beta = (\beta_0, \beta_1, ..., \beta_p) \in \mathbb{R}^{p+1}$ is the vector of regression coefficients. $\beta$ is found through maximum likelihood optimization. The classical linear regression model is a particular case of GLM where $Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2)$ and the link $g$ is the identity-function. The following sections aim to link $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$ and $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta)))$.

## A.3 Univariate analysis in GLM

This section focuses on the univariate case (p = 1). In the case (C1), (C2) and (C3), the quality variable $\mathbf{\Omega}$ is independent from the others variables. For $\beta$ in $\mathbb{R}^2$, the model $M_2$ maximizes the following log-likelihood

$$\begin{aligned}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta))) = \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))|\Omega = 1) \times \mathbb{E}(\Omega_1 = 1) \\ + \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z}; \beta))|\Omega = 0) \times \mathbb{E}(\Omega_1 = 0).\end{aligned} \tag{21}$$

In the equation 21,the expected value of $\Omega = 0$, equals to $Q_1$, is known and when $\Omega = 0$, $X_1^{real}$ is not observed and $Z_1$ is given.

In the case (C1), (C2) and (C3), the quality variable $\mathbf{\Omega}$ is independent from the others variables, which leads to :

$$\begin{aligned}\mathbb{E}(log(f_Y(Y|\mathbf{X}^{real}; \beta)))|\Omega = 1) = \mathbb{E}(log(f_Y(Y|\mathbf{X}^{real}; \beta))), \\ \mathbb{E}(log(f_Y(Y|\mathbf{Z}; \beta))|\Omega = 0) = \mathbb{E}(log(f_Y(Y|\mathbf{Z}; \beta))).\end{aligned} \tag{22}$$

Recall one main regularity condition which is mandatory for the MLE convergence of the exponential family based $\mathbf{X}^{real}$ (see section 6.2 of [13, Lehmann and CAsella, 1998] for all the conditions needed):

**Regularity condition A.1.** *Assume that for every $\beta$, $log(f_Y(Y|\mathbf{X}^{real}; \beta))$ is integrable i.e, $\mathbb{E}(|log(f_Y(Y|\mathbf{X}^{real}; \beta))|) < +\infty$.*

Assumption A.1 and the other ones for the exponential family lead to the *Bartlett identities* and both derivatives can be passed under the integral sign ([2, Barndorff-Nielsen, 1978]). The Bartlett identities are :

$$\begin{aligned} &\mathbb{E}\Big(\frac{\delta}{\delta\beta} log(f_Y(Y|\mathbf{X}^{real};\beta))\Big) = 0, \\ &Var\Big(\frac{\delta}{\delta\beta} log(f_Y(Y|\mathbf{X}^{real};\beta))\Big) = -\mathbb{E}\Big(\frac{\delta^2}{\delta\beta^2} log(f_Y(Y|\mathbf{X}^{real};\beta))\Big). \end{aligned} \tag{23}$$

Moreover, same assumptions on $\mathbf{Z}$ are also needed for the MLE convergence of the exponential family based on $\mathbf{X}$ in particular :

**Regularity condition A.2.** *Assume that for every $\beta$, $log(f_Y(y|z;\beta))$ is integrable i.e $\mathbb{E}(|log(f_Y(Y|Z;\beta))|) < +\infty$.*

Because $Z_j$ has the same marginal distribution than $X_j^{real}$, the regularity conditions A.1 and A.2 highly overlap. The main difference is the independence of $\mathbf{Z}$ from $Y$. Therefore, the *Bartlett identities* are still verified.

**Remark A.1.** *In the univariate case, the condition $\int_{\mathbb{R}^2} |log(f_Y(y|z;\beta))| dF_{Z_1}(z) dF_Y(y) < \infty$ implies $\int_{\mathbb{R}} |log(f_Y(y|\ z;\beta))| dF_{Z_1}(z) < \infty$ for any value of $y$ and $\beta$. Remind that $\mathbf{Z}_1$ distribution has the same distribution than $\mathbf{X}_1^{real}$. Hereafter, the canonical link function is used. In this case, the log-likelihood maximized can be written as follows :*

$$Y_i(\beta_0 + \beta_1 X_1) - b(\beta_0 + \beta_1 X_1) + C^{st}, \tag{24}$$

*where $C^{st}$ is a constant independent of $\beta$ and $X$ and $b(.)$ a real function. In the Bernoulli case supposing $\beta_1 \geq 0$ without loss of generality, the case $\beta_1 = 0$ being trivial, the condition*

$$\begin{aligned} \int_{\mathbb{R}} |log(f_Y(y|z;\beta))| dF_{Z_1}(z) &\underset{Triangle\ ineq.}{\leq} \int_{\mathbb{R}} y_i|\beta_0 + \beta_1 z_1| dF_{Z_1}(z) + \int_{\mathbb{R}} |log(1 + exp(\beta_0 + \beta_1 z_1))| dF_{Z_1}(z), \\ &\underset{\frac{x}{1+x} \leq log(x) \leq x}{\leq} \int_{\mathbb{R}} y|\beta_0 + \beta_1 z_1| dF_{Z_1}(z) + \int_{\mathbb{R}} exp(\beta_0 + \beta_1 z_1) dF_{Z_1}(z), \end{aligned} \tag{25}$$

*can be fulfilled with a condition on the $\mathbf{Z}_1$ generating moment function existence for all $\beta$ and with $\mathbb{E}(|Z_1|) < +\infty$. In the Poisson case, the same sufficient condition can easily be shown :*

$$\int_{\mathbb{R}} |y(\beta_0 + \beta_1 x_1) - exp(\beta_0 + \beta_1 x_1)| dF_{Z_1}(z) \underset{Triangle\ ineq.}{\leq} \int_{\mathbb{R}} y|\beta_0 + \beta_1 z_1| dF_{Z_1}(z) + \int_{\mathbb{R}} exp(\beta_0 + \beta_1 z_1) dF_{Z_1}(z). \tag{26}$$

*As the second moment existence was needed for linear regression convergence, the moment function existence is also needed for a proper maximum likelihood convergence.*

How could we estimate $\mathbb{E}(log(f_Y(\mathbf{Y}|Z_1;\beta)))$ with $X_1^{real}$ ? In the multivariate case, the value $Z_{i,1}$ could be estimated using the other covariables depending on the case. If $X_1^{real}$ and $Z_1$ are correlated or dependent, a function $g$ could exist such as $g(X_1^{real})$ is a good estimator of $Z_1$. Under the case (C1), none of these solutions can be applied. Indeed, the quality index $\mathbf{Q}_1$, the real data-set $X_1^{real}$ and the wrong values $Z_1$ are completely independent.

The following estimator,

$$\bar{Q}_1 \sum_{i=1}^{n} log(f_Y(y_i|X_{i,1}^{real};\hat{\beta})) + (1 - \bar{Q}_1) \times \sum_{i=1}^{n} \frac{1}{n} \sum_{h=1}^{n} log(f_Y(y_i|X_{h,1}^{real};\hat{\beta})). \tag{27}$$

converges almost surely to $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\hat{\beta})))$ (see the proof C.1).

In the multivariate case, the previous assumptions can easily be extended depending on the correlation structure. Under assumptions (X-A3) and (Z-A1), remind the following notation $\mathbf{X}_{(*p)} = (1, X_1; \cdots; X_{p-1})$ and its observed sample $X_{i;(*p)}$. In the same way, $\beta^{*p}$ refers to $(\beta_0, \cdots, \beta_{p-1})$. The expected likelihood

$$\begin{aligned} \mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) = Q_p\ &\mathbb{E}(log(f_Y(Y|\mathbf{X}_{(*p)}^{real}, \mathbf{X}_p = \mathbf{X}_p^{real};\beta))) \\ &+ (1 - Q_p) \int_{\mathbb{R}} \mathbb{E}(log(f_Y(Y|\mathbf{X}_{(*p)}^{real}, \mathbf{X}_p = z;\beta))) f_{Z_p}(z) dz). \end{aligned} \tag{28}$$

can be written in a similar way than in the univariate case under the mild regulatory conditions.

Under these assumptions, each estimator $\hat{\beta}$ and $\hat{\beta}^{M2}$ converges in probabilities respectively to $\beta$ and $\beta^{M2}$ .

## A.4 Example 1: Log-Gaussian GLM

In this section, let focus on Log-Gaussian GLM case. Without loss of generality, remind that the covariate are centred (See the paper [4, Chatelain and Milhaud, 2021] for uncentred case).

Let $\beta \in \mathbb{R}^{p+1}$. The likelihood to optimize is :

$$log(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})) \propto \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\beta)^2, \tag{29}$$

where for the purposes of notation $\mathbf{X}_i = {}^t(1, \mathbf{X}_{i;1})$ and $\beta = {}^t(\beta_0, \beta_1)$.

In the univariate case, the Bartlett identities give the same results as the OLS' one :

$$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0, \ \frac{\hat{\beta}_1^{M2}}{\bar{Q}_1} \xrightarrow{\mathbb{P}.} \beta_1. \tag{30}$$

Indeed, linear regressions and Log-Gaussian GLM models are equivalent. This proof can be easily extended in the multivariate case under the assumption (X-A1) and (Z-A1).

Under the assumption (X-A3) and (Z-A1), only the $\beta_p$ is impacted by

$$\hat{\beta}_j^{M_2} \xrightarrow{\mathbb{P}.} \beta_j, \ \frac{\hat{\beta}_p^{M2}}{\bar{Q}_p} \xrightarrow{\mathbb{P}.} \beta_p, \quad j = 0, ..., p-1. \tag{31}$$

In the particular case (C2) when the quality variables are fully correlated i.e, $\Omega_j = \Omega_k \to Q_j = Q_k$ $(j \neq k)$, under the assumption (Z-A2) without any assumption on correlation structure of $\mathbf{X}^{real}$, it can be shown that :

$$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0, \ \frac{\hat{\beta}_j^{M2}}{\bar{Q}_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j = 0, ..., p. \tag{32}$$

The proof of these results are in C.2. The paper [4, Chatelain and Milhaud, 2021] proved the following theorem, where $\rho$ represents the Pearson correlation of two variable.

**Theorem A.1.** *For all $j \neq k$, if $|\rho| = |\rho_{jk}^{real}| \neq 1$, under (X-A2) and as $n \to +\infty$ :*

$$\begin{aligned}
&\textit{Under (Z-A1):} \ \frac{1}{1-\rho^2}\left(\frac{\hat{\beta}_j^{M_2}}{\bar{Q}_j}(1-\rho^2\,\bar{Q}_j\bar{Q}_k) + \sqrt{\frac{Var(X_k)}{Var(X_j)}}\frac{\hat{\beta}_k^{M_2}}{\bar{Q}_k}\rho\,(\bar{Q}_j\bar{Q}_k - 1)\right) \to \beta_j, \\
&\textit{Under (Z-A2):} \ \frac{1}{1-\rho^2}\left(\frac{\hat{\beta}_j^{M_2}}{\bar{Q}_j}(1-\rho^2\,(1 + 2\bar{Q}_j\bar{Q}_k - \bar{Q}_j - \bar{Q}_k)) + \sqrt{\frac{Var(X_k)}{Var(X_j)}}\frac{\hat{\beta}_k^{M_2}}{\bar{Q}_k}\rho\,(2\bar{Q}_j\bar{Q}_k - \bar{Q}_j - \bar{Q}_k)\right) \\
&\qquad\qquad \to \beta_j.
\end{aligned} \tag{33}$$

## A.5 Example 2: Log-Poisson GLM

In the Poisson case, the additive structure simplifies some calculus. Under assumptions (X-A3) and (Z-A1), the existence of the moment generating function $M_{X_p^{real}}(t) = M_{X_p}(t) = M_{Z_p}(t)$ for all $t \in \mathbb{R}$ and its derivatives' existence are ensured by the mild regularity condition A.2. Denote $V$ the exposure. To keep the notation simple, let omit the exposure $V$ in the expected likelihood $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$. Let $\beta \in \mathbb{R}^{p+1}$. The sample estimator of the expected likelihood is equal to

$$\begin{aligned}
log(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})) &= \bar{Q}_p\, log(\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{\mathbf{real}})) + (1-\bar{Q}_p)\, log(\mathcal{L}(\beta^{*p}; \mathbf{Y}, \mathbf{X}_{(*\mathbf{p})}^{\mathbf{real}})) \\
&\quad + (1-\bar{Q}_p)\sum_{i=1}^{n} V_i e^{\beta^{*p}\mathbf{X}_{i;(*p)}^{real}}(1 - M_{X_p^{real}}(\beta_p)).
\end{aligned} \tag{34}$$

Under (X-A3) and (Z-A1), $X_{(*p)}^{real} = X_{(*p)}$ allows us to evaluate the $M_1$ likelihood using only $X_{(*p)}$ and $\mathbf{Q}$,

$$\begin{aligned}
log(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{Y}|\mathbf{X}, \mathbf{Q})) &= \frac{1}{\bar{Q}_p}(log(\mathcal{L}^{M2}(\hat{\beta}; \mathbf{Y}|\mathbf{X}) - (1-\bar{Q}_p) \times log(\mathcal{L}(\hat{\beta}^{*p}; \mathbf{Y}|X_{(*p)})) \\
&\quad - (1-\bar{Q}_p) \times \sum_{i=1}^{n} V_i e^{\hat{\beta}^{*p}\mathbf{X}_{i;(*p)}^{real}}(1 - M_{X_p}(\hat{\beta}_p))).
\end{aligned} \tag{35}$$

The expected likelihood can be bounded :

$$
\begin{aligned}
Q_p\ log(\mathcal{L}(\beta;\mathbf{Y}|\mathbf{X}^{real})) + (1-Q_p)\ log(\mathcal{L}(\beta^{*p};\mathbf{Y},\mathbf{X}^{\mathbf{real}}_{(*\mathbf{p})})) \\
+ (1-Q_p)\sum_{i=1}^{n} V_i e^{\beta^{*p} X^{real}_{(i);(*p)}} \left(1-\exp\left(\frac{\beta_p^2 \mathbb{V}(\mathbf{X}^{\mathbf{real}}_{\mathbf{p}})}{2}\right)\right) \\
\leq log(\mathcal{L}^{M_2}(\beta;\mathbf{Y}|\mathbf{X})) \leq Q_p\ log(\mathcal{L}(\beta;\mathbf{Y}|X^{real})) + (1-Q_p)\ log(\mathcal{L}(\beta^{*p};\mathbf{Y},\mathbf{X}^{\mathbf{real}}_{(*\mathbf{p})})).
\end{aligned}
\tag{36}
$$

By introducing the normalized coefficient $b_p = \frac{\beta_p}{\sqrt{\mathbb{V}(X_p^{real})}}$, one can see that a small normalize coefficient implies a narrower the interval. In other words, the impact of variable quality on the likelihood logically depends on the normalize coefficient.

*Proof.* See C.3.1. □

**Lemma 2.** *Let $\beta \in \mathbb{R}^{p+1}$. Under the assumption (X-A3), the derivatives of the $M_2$ log-likelihood for $j$ in $\{0,...,p-1\}$ are equal to :*

$$
\begin{aligned}
\frac{\delta}{\delta\beta_p}\mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) &= Q_p\ d_p(\beta) \\
&\quad - (1-Q_p)\int_{\mathbb{R}^{p-1}} v e^{\beta^{*p} x^{real}_{(*p)}} dF_{\mathbf{X}^{real}_{(*p)}}(x^{real}_{(*p)})\ M'_{X_p}(\hat{\beta}_p); \\
\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) &= d_j(\beta),
\end{aligned}
\tag{37}
$$

*where $d_i$ is the derivative according to $\beta_i$ of $\mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}^{real})))$ .*

**Remark A.2.** *Unlike the Log-gaussian case, the difference $\beta_p^{M_2}$ and $\beta_p$ depends on the distribution of $X_p$.*

When $\hat{\beta}_p \to 0$, $M'_{X_p}(\hat{\beta}_p) \to 0$. It can be easily shown that the derivatives - equation 37 are a constant function of the mean quality. Therefore, the following proposition A.5 can be deduced.

**Proposition 1** *Suppose the framework of this paper with log-Poisson distribution. Under the Assumptions (X-A3), i.e. $Q_j = 1$ for $j \in \{1,\cdots,p-1\}$ and $Q_p \in (0,1)$,*

$$
\begin{aligned}
\beta_p^{M_2} :[0,1] &\to \mathbb{R} \\
Q_p &\mapsto \beta_p^{M_2}(Q_p)
\end{aligned}
\tag{38}
$$

*is a monotonic function of the quality.*

Using the Lemma 2 it is straightforward to show the following theorem.

**Theorem A.2.** *Under the assumption (X-A3) and (Z-A1),*

$$
\begin{aligned}
\hat{\beta}_j^{M_2} &\xrightarrow{\mathrm{P}} \beta_j, \quad j \in 0,...,p-1, \\
\hat{\beta}_p^{M_2} &\xrightarrow{\mathrm{P}} [0;\beta_p].
\end{aligned}
\tag{39}
$$

A particular application of this theorem would be under the univariate case $p = 1$. Remark that in the univariate case (C1) and (C2) are equal. In multivariate case, under (Z-A2) and (C2) with fully correlated quality variable and without any assumption on the structure of $\mathbf{X}^{real}$, the expected log likelihood can be written only using $\mathbf{X}$ and the quality index $\mathbf{Q}$, (see C.3.4) :

$$
\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) = \frac{1}{Q_1}\Big(\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X},\beta)) - (1-Q_1)\Big(\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0)) + Vexp(\beta_0)\,(1-M_{\mathbf{X}^{real}}(\beta_*))\Big)\Big),
\tag{40}
$$

where $M_{\mathbf{X}^{real}}(\beta_*)$ is the multivariate generating function of $X_1^{real},...,X_p^{real}$ and $\beta_* = {}^t(\beta_1,....,\beta_p)$. Unfortunately, no bounds explicit can be state. The paper [12, kuwaranancharoen and Sundaram, 2018] provides an upper bound on the location of local minimum of the sum of two strongly convex function under the assumption of bounded gradient. The difficulty is that the log-likelihood exponential family is almost strongly convex, *i.e.* strongly convex in the neighbourhood of $\beta$ as proved in [11, Kakade et al. 2010].

## A.6 Example 3: Log-Gamma GLM

The expected log-likelihood of Log-Gamma $Y \sim \Gamma(\mu, \nu)$ can be written :

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) = \int_{\mathbb{R}^{p+1}} \nu(-y\ exp(-\mathbf{x}\beta) - \mathbf{x}\beta + (\nu-1)log(y) - log(\Gamma(\nu))dF_{X_1,...,X_p,Y}(\mathbf{x}, y). \tag{41}$$

Here, the only interest is to maximize the log likelihood according to $\beta$ for a known $\nu$. Therefore, the expected log-likelihood will be studied,

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) \propto \int_{\mathbb{R}^{p+1}} -y\ exp(-\mathbf{x}\beta) - \mathbf{x}\beta dF_{X_1,...,X_p,Y}(\mathbf{x}, y). \tag{42}$$

Under (X-A3) and (Z-A1), the expected log likelihood $\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X})))$ is equal to

$$Q_p\ \mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1-Q_p)\mathbb{E}(log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))M_{X_p}(-\hat{\beta}_p). \tag{43}$$

The $M_1$ estimator can be calculated

$$\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) = \frac{1}{Q_p}\Big(\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}))) - (1-Q_p)\mathbb{E}(log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))M_{X_p}(-\hat{\beta}_p)\Big). \tag{44}$$

**Lemma 3.** *Let $\beta \in \mathbb{R}^{p+1}$. Under the assumption (X-A3), the derivative of the $M_2$ log-likelihood for $j$ in $\{0, ..., p-1\}$ are equal to :*

$$\begin{aligned}
\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_j(\beta) \\
&\quad + (1-Q_p)\ \frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta^{*\mathbf{p}}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))M_{X_p}(\beta_p), \\
\frac{\delta}{\delta\beta_p}\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_p(\beta) \\
&\quad + (1-Q_p)\ \mathbb{E}(log(f_Y(\beta^{*p}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))M'_{X_p}(\beta_p),
\end{aligned} \tag{45}$$

*where $d_j$ is the derivative according to $\beta_j$ of $\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}^{real})))$ .*

The lemma 3 can not lead to a theorem like in the Log-Poisson case. The minimization of a sum of concave function in $\mathbb{R}^{p+1}$ does not necessary lead to $\beta_j^{M2} \in [min(\beta_j^{-p}, \beta_j), max(\beta_j^{-p}, \beta_j)]$ where $\beta_j^{-p}$ is the maximum likelihood estimator $\mathbb{E}(log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))$ and with $\beta_p^{-p} = 0$. Therefore, the Log-Gamma coefficients' evolution are not easily bounded in the general case. Nevertheless, the $\beta_p^{M2}$ are still continuous according to the quality.

Under (C2) and (Z-A2), $\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X})))$ is equal :

$$Q_p\ \mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1-Q_p)M_{\mathbf{X}^{real}}(-\beta)\mathbb{E}(log(f_Y(\hat{\beta}_0; \mathbf{Y}))). \tag{46}$$

The proof is no different from the precedents. Still no bound can be state when $p > 2$.

## A.7 Without multiplicative properties: Inv-Gamma GLM and Probit GLM

The expected log-likelihood of Inv-Gamma $Y \sim \Gamma(\mu, \nu)$ will be maximized for a known $\nu$. The maximum likelihood estimator will maximize the sample analogue of

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) \propto \int_{\mathbb{R}^{p+1}} -y\ \mathbf{x}\beta + log(\mathbf{x}\beta)dF_{X_1,...,X_p,Y}(\mathbf{x}, y). \tag{47}$$

The expected log likelihood $\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X})))$ is equal to

$$Q_p\ \mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1-Q_p)\int_{\mathbb{R}^{p+1}} log(\mathbf{x}^{real}\beta^{*p} + z_p\beta_p)dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}^{real}, y)dF_{Z_p}(z_p). \tag{48}$$

Because of $log(\mathbf{X}^{real}\beta^{*p} + Z_p\beta_p)$, the sample analogue can not be estimated using only $\mathbf{X}^{real}$ and $\mathbf{X}$ which will not allow us to find a relation between the likelihood using only these two data sets.

For the Bernoulli distribution using its canonical link function, the expected log-likelihood :

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) \propto \int_{\mathbb{R}^{p+1}} -y\ \mathbf{x}\beta + log(1 + exp(\mathbf{x}\beta))dF_{X_1,...,X_p,y}(\mathbf{x}, y), \tag{49}$$

can not be calculated using only $\mathbf{X}^{real}$. Looking for an approximation, $log(1+exp(x)) = log(2+exp(x)-1) = log(2) + log(1+(exp(x)-1)/2) \sim log(2) + (exp(x)-1)/2 + o((exp(x)-1))$ when $exp(x)$ is close to 1. Therefore, when $\mathbf{x}\beta$ is close to 0, a fine approximation with a multiplicative structure can be state.

# Appendix B. Various operational remarks

In this section, let focus on the simplest case of GLM : log-Gaussian.

## B.1 Quality impact and attenuation

The different results show that the "attenuation"[8] on $\hat{\beta}$ due to data quality can be explained. However, the quality impacts might not always decrease the coefficients, as shown in [4, Chatelain and Milhaud, 2021]. The quality of a variable impacts all coefficients related to other correlated variables in the uncentred case. Figure 11 under (X-A2) and (Z-A1) in linear regression shows that even in the simple case the "attenuation" is not always true. With some correlation, the coefficient can be higher than the usual one (the true coefficient equal to 1 and is represented the line on Figures 11 and 12) and even might change sign.

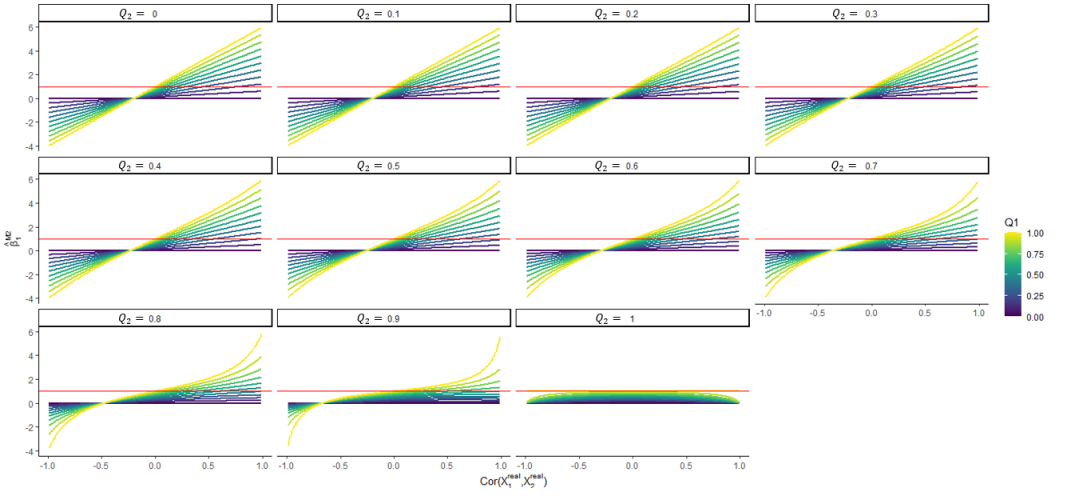



**Figure 11**: Log-Gaussian GLM: Value of $\beta_1^{M_2}$ wrapped by $Q_2$ and grouped by $Q_1$. The coefficient $\beta$ are all equal to 1 and the ratio of the standard deviation $\sqrt{Var(X_1^{real})/Var(X_2^{real})}$ equals to 6. The red straight line represents, $\beta_1$ which is equal to 1.

This is especially harmful to insurance pricing, where covariates' choice must be justified by their impacts. Indeed, some coefficients may seem counter-intuitive due to quality impacts. Figures 11 and 12 provide an illustration of the impact of $Q_1$ depending on $Q_2$ ($\beta_1, \beta_2$ always equal to 1). The coefficients' evolution is not linear with the correlation. Figure 11 shows that if $\rho < 0$, $\beta_1^{M_2}$ could be negative, even if $\beta_1 > 0$. Another point is that the coefficients could be considered as null even if the variable's quality is not low. For instance, for $Q_2 = 0.7$ and $\rho \approx -0.4$, $\beta_1^{M_2} \approx 0$. In this case, dropping the variable $X_1$ would not have any impact on $\beta_2^{M_2}$ even if the true coefficient is different from 0. Moreover, by finding the $\beta^{M_1}$ - thanks to $\mathbf{X}$ and $\mathbf{Q}$, the modeler can find the "real" impact of a variable in models, thus justifying it.

**Remark B.1.** *Here, the discussion was done with the simplest hypothesis under the case (C1) and for Gaussian distribution where the variable quality does not impact others independent variables coefficients. For other distributions, the quality impacts would complicate the whole issue further.*

[8] As called in the econometric literature.

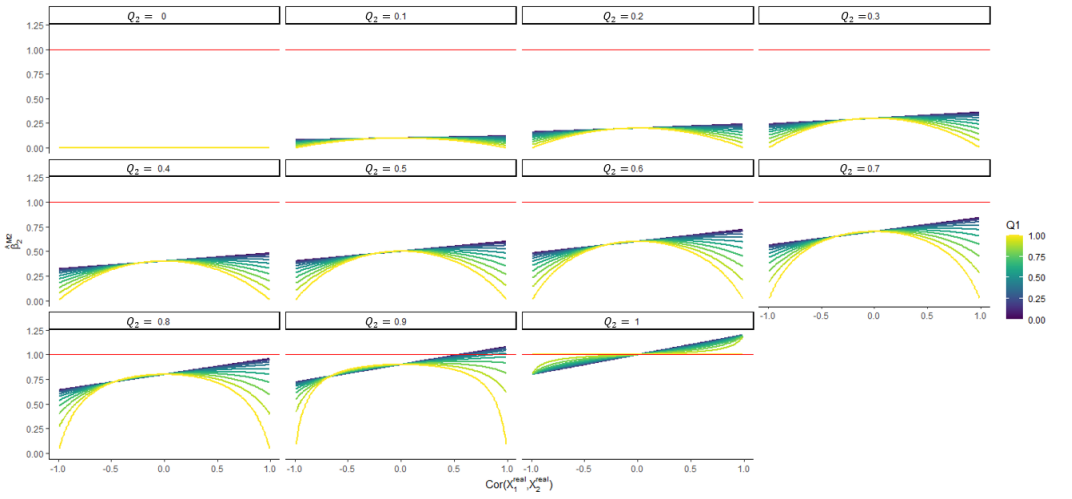



**Figure 12**: Log-Gaussian GLM: Value of $\beta_2^{M_2}$ wrapped by $Q_2$ and grouped by $Q_1$. The coefficient $\beta$ are all equal to 1 and the ratio of the standard deviation $\sqrt{Var(X_1^{real})/Var(X_2^{real})}$ equals to 6.

## B.2 Missing data

The case of missing values could be seen as a particular case of this framework, where missing values are observations with a null quality. In the case of linear regression under (C1), (X-A1) and (Z-A1), the mean imputation is equivalent to the process explained in this paper. Denote the following model $E[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and $\rho$ the correlation between the two covariates. First, suppose $\rho = 0$ and the individual $i$ having its $X_{i,1}$ missing ; using a simple mean imputation, the predicted value of $Y_i$ would be

$$\hat{Y}_i = \beta_0 + \beta_1 \mathbb{E}(X_1^{real}) + \beta_2 X_{i;2}^{real}, \tag{50}$$

using the process 3. Under (X-A1) and (Z-A1), the predicted value of $y_i$ would be

$$\hat{Y}_i = \beta_0^K + \beta_1^K Z_{i;1} + \beta_2^K X_{i;2}^{real}, \tag{51}$$

where $K$ is the pattern of the quality - here $K = (0, 1)$, $\hat{\beta}_j^K$ is the estimator found thanks to the process 3, $j \in \{0; 1; 2; 3\}$ and $Z_{i,1}$ is a value drawn randomly from the empiric distribution of $X_1^{real}$. Due to the different assumptions and $K = (Q_1 = 0, Q_2 = 1)$, the coefficients can be written as

$$(\beta_0^K, \beta_1^K, \beta_2^K) = (\beta_0 + \beta_1 \mathbb{E}(X_1^{real}), 0, \beta_2), \tag{52}$$

showing the equivalence between the two methods. However, in correlated cases for instance under (X-A2) and (Z-A1), the coefficients would equal to :

$$(\beta_0^K, \beta_1^K, \beta_2^K) = (\beta_0 + \beta_1 \mathbb{E}(X_1^{real}) - \sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}} \beta_1 \rho \mathbb{E}(X_2^{real}), 0, \beta_2 + \sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}} \beta_1 \rho). \tag{53}$$

Thus, the equivalent imputation here for $X_{i;1}^{real}$ should be $\mathbb{E}(X_1^{real}) - \sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}} \rho(\mathbb{E}(X_2^{real}) - X_{i;2})$. This imputation corresponds to the result of a linear regression to predict $X_1^{real}$ using only $X_2^{real}$ for the individual which is the best one according to the linear regression. In fact, $\sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}} \beta_1^{M_1} \rho \mathbb{E}(X_2^{real})$ corresponds to the linear part of the information $X_1^{real}$ already taken into account by $X_2^{real}$.

**Multivariate case** By extrapolating these results, it seems that for one missing observation the equivalent imputation should be the prediction of the linear regression of the other covariates. However, this remark does not take into account the issue with other covariates' quality. To go further than the case (C1), the credibility of other covariates may be also correlated with the fact that a value is missing. This remark is close to the analysis of Seaman 2014 [18, Seaman and White, 2014] about how to impute with fully conditional specifications.

For Log-Poisson GLM, it has been shown that under (X-A3) and (Z-A1) the mean imputation is also equivalent, which is not always true for other assumptions. For instance, in Log-Gamma model the other coefficients are also impacted. To find the best solution into account, one should find $\beta^K$ and modify all the other coefficients.

# Appendix C. Proofs

## C.1 Proof of the Theorem 3.1

*Proof.* Let $(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{real}, \mathbf{Q})$ be the data sets as defined by the equation 1. In the univariate case $p = 1$, the expected log-likelihood of the model $M_2$ depends on the quality index,

$$\begin{aligned}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) = \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))|\Omega = 1) \times \mathbb{E}(\Omega = 1) \\ + \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z};\beta))|\Omega = 0) \times \mathbb{E}(\Omega = 0),\end{aligned} \tag{54}$$

thanks to the independance between $\Omega$ and respectively $\mathbf{X}^{real}$ and $\mathbf{Z}$. The first term is known and because $Z$ is independent of $Y$, the second can rewritten, using Fubini's theorem :

$$\mathbb{E}(log(f_Y(Y|\mathbf{Z};\hat{\beta}))) = \mathbb{E}_Y \int_{\mathbb{R}} log(f_Y(Y|z;\beta))dF_{Z_1}(z). \tag{55}$$

Because $Z_1$ have the same distribution as the $X_1^{real}$, $X_1^{real}$ can be used to estimate the density $f_{Z_1}$ so $dF_{Z_1}(s) = dF_{X_1^{real}}(s)$. Finally, the previous equation can be estimated by the mean sample. Because $\{X_{1;1}^{real}, ..., X_{n;1}^{real}\}$ are *i.i.d* observations, the sample estimator would be

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{h=1}^{n} log(f(y_i|X_{h,1}^{real};\hat{\beta})). \tag{56}$$

Having $\mathbb{E}(|log(f_Y(Y|\ \mathbf{Z},\beta)|) < \infty$ using the strong law of large numbers, this sample estimator converges almost surely. The sample estimator $\sum log(f_Y(y_i|X_{i;1}^{real};\hat{\beta}))$ converges almost surely to $\mathbb{E}(log(f_Y(Y|\mathbf{X}^{real};\hat{\beta}))$. Using the strong law of large number, $\bar{Q}_1$ converges in probability towards $Q_1$. Thus,

$$\bar{Q}_1\sum_{i=1}^{n} log(f_Y(y_i|X_{i,1}^{real};\hat{\beta})) + (1-\bar{Q}_1) \times \sum_{i=1}^{n}\frac{1}{n}\sum_{h=1}^{n} log(f_Y(y_i|X_{h,1}^{real};\hat{\beta})). \tag{57}$$

converges almost surely to $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\hat{\beta})))$. Denote this estimator $log(\mathcal{L}^{M_2}(\beta|\mathbf{Y}, \mathbf{X}^{real}, \mathbf{Q}))$.

Finally, the following points are true :

- observations are i.i.d and the density is Lebesgue measurable;
- the parameter space of $\beta$ is compact and open;
- the previous estimator is concave as sum of concave function and is differentiable according to $\beta$;
- Identifiability : the estimator function is a smooth function of $\beta$ and converges in probability for all $\beta$ towards $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\hat{\beta})))$ which has the unique solution.

Therefore, using the Cramer-Rao conditions - Collorary 3.8 of [13, Lehmann and Casella, 1998], the global maximum exists, is unique and converges in probability to $\beta^{M_2}$, *i.e.*

$$\hat{\beta}^{M_2|X^{real},\mathbf{Q}} \xrightarrow{\mathbb{P}} \beta^{M_2},$$

meaning that the estimator is consistent. □

## C.2 Log-Gaussian proofs

### C.2.1 Proof of equation 30

*Proof.* Remind that $X_1^{real}$ is supposed centred. The likelihood maximization solution can be found as the solution of the derivative equal to 0. Deriving by $\beta$, the derived sample estimator can be written as follows :

$$\begin{aligned}\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\hat{\beta}^{M2}))}{\delta\beta} =& \frac{\delta \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta)))}{\delta\beta} \times Q_1 + \frac{\delta \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z};\beta)))}{\delta\beta} \times (1-Q_1)\\ =& Q_1 \frac{\delta}{\delta\beta}\int_{\mathbb{R}^2}(y - x\beta_1 + \beta_0)^2 dF_{X_1^{real},Y}(x,y)\\ &+ (1-Q_1)\frac{\delta}{\delta\beta}\int_{\mathbb{R}}\int_{\mathbb{R}}(y - z\beta_1 - \beta_0)^2 dF_{Z_1}(z)dF_Y(y) = 0.\end{aligned} \tag{58}$$

Remind that the MLE solution exists and is unique. It is a well-known fact that, if the identity $\int f_Y(y;\beta)d(y) = 1$ is twice differentiable with respect to $\beta$ and, both derivatives can be passed under the integral sign. Therefore,the theorem of differential under the integral can be applied and leads to the first *Bartlett identity*,

$$\begin{aligned}\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\hat{\beta}^{M2}))}{\delta\beta_0} &= -2\,Q_1\int_{\mathbb{R}^2}(y - x\beta_1 - \beta_0)dF_{X_1^{real},Y}(x,y)\\ &\quad -2\,(1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}}(y - z\beta_1 - \beta_0)dF_{Z_1}(z)dF_Y(y) = 0,\\ \frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\hat{\beta}^{M2}))}{\delta\beta_0} &= -2\,Q_1\int_{\mathbb{R}^2}x(y - x\beta_1 - \beta_0)dF_{X_1^{real},Y}(x,y)\\ &\quad -2\,(1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}}z(y - z\beta_1 - \beta_0)dF_{Z_1}(z)dF_Y(y) = 0.\end{aligned} \tag{59}$$

Remind that $\int_{\mathbb{R}} z dF_{Z_1}(z) = 0 = \int_{\mathbb{R}} x dF_{X_1^{real}}(x)$. Therefore, the solutions of the precedent equations are :

$$\begin{aligned}\beta_0^{M2} &= Q_1\int_{\mathbb{R}^2} y dF_{X_1^{real},Y}(x,y) + (1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}} y dF_{Z_1}(z)dF_Y(y),\\ &= \int_{\mathbb{R}} y dF_Y(y) = \beta_0,\\ \beta_1^{M2} &= \frac{Q_1\int_{\mathbb{R}^2} xy dF_{X_1^{real},Y}(x,y)}{Q_1\int_{\mathbb{R}^2} x^2 dF_{X_1^{real},Y}(x,y) + (1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}} z^2 dF_{Z_1}(z)dF_Y(y)} = Q_1\beta_1.\end{aligned} \tag{60}$$

Let end the proof by replacing the $\beta_0$, $Q_1$ and $\beta_1$ by their estimators. Each of them converges in probabilities; $\hat{\beta}_0$ and $\hat{\beta}_1$ thanks to the asymptotics MLE proprieties and $\bar{Q}_1$ using the strong law of large number. The proof can be generalized exactly in the same way under (X-A1) and (Z-A1) for $p > 1$. □

### C.2.2 Proof of equation 31

*Proof.* The first *Bartlett identities* under the assumption (X-A3) are equal, for $j = 1, ..., p-1$:

$$\begin{aligned}\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta^{M_2}))}{\delta\beta_0} &= -2\,Q_1\int_{\mathbb{R}^{p+1}}(y - \mathbf{x}_{*p}^{real}\beta^{*p;M_2} - x_p^{real}\beta_p^{M_2})dF_{X_1^{real},...,X_p^{real},Y}(\mathbf{x}_{*p}^{real}, x_p^{real}, y)\\ &\quad -2\,(1-Q_1)\int_{\mathbb{R}^p}\int_{\mathbb{R}}(y - \mathbf{x}_{*p}^{real}\beta^{*p;M2} - z_p\beta_p^{M2})dF_{Z_p}(z_p)dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}_{*p}^{real}, y) = 0,\\ \frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta^{M2}))}{\delta\beta_j} &= -2\,Q_1\int_{\mathbb{R}^{p+1}}x_j^{real}(y - \mathbf{x}_{*p}^{real}\beta^{*p;M_2} - x_p^{real}\beta_p^{M_2})dF_{X_1^{real},...,X_p^{real},Y}(\mathbf{x}_{*p}^{real}, x_p^{real}, y)\\ &\quad -2\,(1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}^p}x_j^{real}(y - \mathbf{x}_{*p}^{real}\beta^{*p;M_2} - z_p\beta_p^{M_2})dF_{Z_p}(z_p)dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}_{*p}^{real}, y) = 0,\\ \frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta^{M2}))}{\delta\beta_p} &= -2\,Q_1\int_{\mathbb{R}^{p+1}}x_p^{real}(y - \mathbf{x}_{*p}^{real}\beta^{*p;M_2} - x_p^{real}\beta_p^{M_2})dF_{X_1^{real},...,X_p^{real},Y}(\mathbf{x}_{*p}^{real}, x_p^{real}, y)\\ &\quad -2\,(1-Q_1)\int_{\mathbb{R}^p}\int_{\mathbb{R}}z_p(y - \mathbf{x}_{*p}^{real}\beta^{*p;M_2} - z_p\beta_p^{M_2})dF_{Z_p}(z_p)dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}_{*p}^{real}, y) = 0.\end{aligned} \tag{61}$$

Remind that $\int_{\mathbb{R}} z_j dF_{Z_j}(z_j) = 0 = \int_{\mathbb{R}} x_j^{real} dF_{X_j^{real}}(x_j^{real})$ for $j = 1, ..., p$. Therefore, the solutions of the precedent equations are:

$$\beta_0^{M2} = \beta_0, \beta_j^{M2} = \beta_j, \beta_p^{M2} = Q_p \beta_p, \quad j = 1, ..., n. \tag{62}$$

Let end the proof by replacing the $\beta^{M_2}$ and $Q_1$ by their estimators. Each of them converges in probabilities; $\hat{\beta}^{M_2}$ thanks to asymptotics MLE proprieties and $\bar{Q}_1$ using the strong law of large number. □

### C.2.3 Proof of equation 32

*Proof.* For this proof, denote $\beta_* = (\beta_1, ..., \beta_p)$. In the case (C2) with perfectly correlated quality variable , *i.e.*, $\Omega_j = \Omega_k \to Q_j = Q_k$ $(j \neq k)$, it leads to the following equation,

$$\begin{aligned}
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M_2}))}{\delta \beta} &= 0 \\
&= \frac{\delta \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta^{M_2})))}{\delta \beta} \times Q_1 + \frac{\delta \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z}; \beta^{M_2})))}{\delta \beta} \times (1 - Q_1) \\
&= Q_1 \frac{\delta}{\delta \beta} \int_{\mathbb{R}^{p+1}} (y - x^{real} \beta_*^{M_2} - \beta_0^{M_2})^2 dF_{X_1^{real}, ..., X_p^{real}, Y}(x, y) \\
&\quad + (1 - Q_1) \frac{\delta}{\delta \beta} \int_{\mathbb{R}} \int_{\mathbb{R}^p} (y - \mathbf{z} \beta_*^{M_2} - \beta_0^{M_2})^2 dF_{Z_1, ..., Z_p}(\mathbf{z}) dF_Y(y).
\end{aligned} \tag{63}$$

The first *Bartlett identity* under the assumption (Z-A2) are equal:

$$\begin{aligned}
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M_2}))}{\delta \beta_0} &= -2\, Q_1 \int_{\mathbb{R}^{p+1}} (y - \mathbf{x}^{real} \beta_*^{M_2} - \beta_0^{M_2}) dF_{X_1^{real}, ..., X_p^{real}, Y}(\mathbf{x}^{real}, y) \\
&\quad - 2\,(1 - Q_1) \int_{\mathbb{R}} \int_{\mathbb{R}^p} (y - \mathbf{z} \beta_*^{M_2} - \beta_0^{M_2}) dF_{Z_1, ..., Z_p}(\mathbf{z}) dF_Y(y) = 0, \\
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M_2}))}{\delta \beta_j} &= -2\, Q_1 \int_{\mathbb{R}^{p+1}} \mathbf{x}_j^{real} (y - \mathbf{x}^{real} \beta_*^{M_2} - \beta_0^{M_2}) dF_{X_1^{real}, ..., X_p^{real}, Y}(\mathbf{x}^{real}, y) \\
&\quad - 2\,(1 - Q_1) \int_{\mathbb{R}} \int_{\mathbb{R}^p} z_j (y - \mathbf{z} \beta_*^{M_2} - \beta_0^{M_2}) dF_{Z_1, ..., Z_p}(\mathbf{z}) dF_Y(y) = 0.
\end{aligned} \tag{64}$$

Remind that $\int_{\mathbb{R}} z_j dF_{Z_j}(z_j) = 0 = \int_{\mathbb{R}} x_j^{real} dF_{X_j^{real}}(x_j^{real})$ for $j = 1, ..., p$. Under the assumption (Z-A2), $\int_{\mathbb{R}^p} z_j \mathbf{z} dF_{Z_1, ..., Z_p}(\mathbf{z}) = \int_{\mathbb{R}^p} \int_{\mathbb{R}} x_j^{real} \mathbf{x}^{real} dF_{X_1^{real}, ..., X_p^{real}}(\mathbf{x}^{real})$. Therefore, the solutions of the precedent equations are:

$$\beta_0^{M2} = \beta_0, \beta_j^{M2} = Q_j \beta_j, \quad j = 1, ..., n. \tag{65}$$

Let end the proof by replacing the $\beta^{M_2}$ and $Q_1$ by their estimators. Each of them converges in probabilities; $\hat{\beta}^{M_2}$ thanks to asymptotics MLE proprieties and $\bar{Q}_1$ using the strong law of large number. □

## C.3 Proof for the GLM Log-Poisson

### C.3.1 Proof of equations 34 and 36

*Proof.* To keep the notation simple, I omit the exposure $V$ in $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta)))$. Under the assumption (X-A3), using the Fubini's theorem, the expected likelihood (without the constant part) is equal to

$$\begin{aligned}
\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) \propto & Q_p \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^{p+1}} -ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z} + n(\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z) dF_{X^{real}_1,\dots X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)dF_{Z_p}(z) \\
\propto & Q_p \mathbb{E}(log(f_Y(Y|\mathbf{X}^{real};\beta))) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^{p+1}} +ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} - ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z} dF_{X^{real}_1,\dots X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)dF_{Z_p}(z) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^{p+1}} -ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} + y(\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z) dF_{X^{real}_1,\dots,X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)dF_{Z_p}(z) \\
\propto & Q_p \mathbb{E}(log(f_Y(Y|\mathbf{X}^{real};\beta))) + (1-Q_p)\mathbb{E}(log(f_Y(Y|\mathbf{X}^{real}_{(*p)};\beta^{*p}))) \\
& + (1-Q_p)\int_{\mathbb{R}^{p+1}} ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} dF_{X^{real}_1,\dots X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)(1-M_{X_p}(\beta_p)).
\end{aligned} \tag{66}$$

Because all the input centred, the last term of the integral is null. Moreover, the moment generating function $M_{\mathbf{X^{real}_p}}(t)$ exists for all $t \in \mathbb{R}$ and the expected likelihood has at sample analogue using only $\mathbf{X}^{real}$

$$\begin{aligned}
& \bar{Q}_p\ log(\mathcal{L}(\beta|\mathbf{Y},\mathbf{X}^{real})) + (1-\bar{Q}_p)\ log(\mathcal{L}(\beta^{*p}|\mathbf{Y},\mathbf{X}^{real}_{*p})) \\
& + (1-\bar{Q}_p)\sum_{i=1}^{n} v_i e^{\beta^{*p}\mathbf{X}^{real}_{i,(*p)}}(1-M_{X_p}(\beta_p)).
\end{aligned} \tag{67}$$

If $X_p$ is bounded, the Hoeffding Lemma gives us a proper upper bound and Jensen inequality gives us the inferior one. Indeed,

$$exp\left(\beta\mathbb{E}(X)\right) \leq \mathbb{E}(e^{\beta X}) \leq exp\left(\beta\mathbb{E}(X) + \frac{\beta^2(max(X)-min(X))^2}{8}\right).$$

With Hoeffding inequality, another bound can be deduced without needing a bounded variable[9] :

$$exp\left(\beta\mathbb{E}(X)\right) \leq \mathbb{E}(e^{\beta X}) \leq exp\left(\beta\mathbb{E}(X) + \frac{\beta^2\mathbb{V}(X)}{2}\right).$$

These inequalities lead to the equation 36. □

### C.3.2 Proof of the Lemma 2

*Proof.* For $j$ in $0,\dots,p$, the gradient

$$\begin{aligned}
\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) = & Q_p \frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta))) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^{p}} -v\mathbf{x}^{real}_j e^{\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z} + n\mathbf{x}^{real}_j dF_{X^{real}_1,\dots X^{real}_{p-1},N}(\mathbf{x}^{real}_{(*p)},n)dF_{Z_p}(z)
\end{aligned} \tag{68}$$

can be separated in two part. Remind that $dF_{Z_p}(z) = dF_{X^{real}_p}(x)$ because $Z_p$ and $X^{real}_p$ have the same distribution. Under the assumption (X-A3), $dF_{X^{real}_1,\dots X^{real}_p}(\mathbf{x}^{real}) = dF_{X^{real}_1,\dots X^{real}_{p-1}}(\mathbf{x}^{real}_{(*p)})dF_{X^{real}_p}(x^{real}_p)$ for $j = 1,\dots,p-1$. By replacing these

[9] Recall that $X_p$ is assumed to possess a second moment through the mild regularity conditions A.1 -A.2.

values in the previous equation, we have :

$$\begin{aligned}\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) &= Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta)))\\ &\quad+(1-Q_p)\int_{\mathbb{R}^{p+1}} -vx_j^{real}e^{\beta\mathbf{x}^{real}}+n\mathbf{x}_j^{real}dF_{X_1^{real},...X_p^{real},N}(\mathbf{x}^{real},n)\\ &= Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta)))\\ &\quad+(1-Q_p)\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta)))=d_j(\beta),\end{aligned} \tag{69}$$

The derivative according to $\beta_p$ is calculated thanks to equation 67 without difficulty.
This end the proof for equation 37. □

### C.3.3 Proof of the theorem A.2

*Proof.* The solution (MLE) $\beta^{M_2}$ exists and is unique. Moreover, the solution $\beta^{M_2}$ is a global maxima. Therefore, the solution $\beta^{M_2}$ nullifies the partial derivatives, $d_j^{M_2}$ for $j=0,...,p$, i.e $d_j^{M_2}(\beta^{M_2})=0$. In the same way, $d_j(\beta)=0$. One can remark that

$$\begin{aligned}d_j(\beta) &= \int_{\mathbb{R}}\int_{\mathbb{R}^p} -vx_j^{real}e^{\beta^{*p}\mathbf{x}_{*p}^{real}+\beta_p x^{real}}dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)dF_{X_p^{real}}(x^{real})\\ &= \int_{\mathbb{R}^p} -vx_j^{real}e^{\beta^{*p}\mathbf{x}_{*p}^{real}+\beta_p x^{real}}dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)\underbrace{\int_{\mathbb{R}}e^{\beta_p x^{real}}dF_{X_p^{real}}(x^{real})}_{>0}=0,\end{aligned} \tag{70}$$

which leads to $\int_{\mathbb{R}^p} -vx_j^{real}e^{\beta^{*p}\mathbf{x}_{*p}^{real}+\beta_p x^{real}}dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)=0$.

Denote $b$ a set of coefficient such as $b_{*p}=\beta_{*p}$ and $b_p\in\mathbb{R}^*$, $j=1,...,p-1$. The derivative $d_j^{M_2}(\beta^{M_2})$,

$$\begin{aligned}d_j^{M_2}(b)=d_j(b) &= \int_{\mathbb{R}}\int_{\mathbb{R}^p} -vx_j^{real}e^{\beta_{*p}\mathbf{x}_{(*p)}^{real}+b_p x^{real}}dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)dF_{X_p^{real}}(x^{real})\\ &= \underbrace{\int_{\mathbb{R}^p} -vx_j^{real}e^{\beta_{*p}\mathbf{x}_{*p}^{real}}dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)}_{=0}\int_{\mathbb{R}}e^{b_p x^{real}}dF_{X_p^{real}}(x^{real}),\end{aligned} \tag{71}$$

is null for $j=1,...,p-1$. Deriving by $\beta_p$, the derivatives equals to

$$d_p^{M_2}(b)=Q_p\,d_p(b)-(1-Q_p)\int_{\mathbb{R}^p} ve^{\beta^{*p}X_{i,*p}^{real}}dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)M_{X_p}^{'}(b_p). \tag{72}$$

If $b_p>\beta_p$, $d_p(b)>0$ and if $b_p<0$, $d_p(b)<0$. In the same way, if $b_p>\beta_p$, $-M_{X_p}^{'}(b_p)>0$ and if $b_p<0$, $-M_{X_p}^{'}(b_p)<0$. These inequalities lead to if $b_p>\beta_p$, $d_p^{M_2}(b)<0$ and if $b_p<0$, $d_p^{M_2}(b)>0$.

Because $b\mapsto d_p^{M_2}(b)$ is a continuous function, a $b_p\in[0,\beta_p^{M_1}]$ exists, such as $d_p^{M_2}(b)=0$.

We have proven that it exists $b$ with the following characteristic's $b_{*p}=\beta_{*p}$ and $b_p\in[0,\beta_p^{M_1}]$ such as :

$$d_j^{M_2}(b)=0,\ d_p^{M_2}(b)=0. \tag{73}$$

Because the solution of $M_2$ log-likelihood maximization is unique, the previous solution is the global maximum $\beta^{M_2}$.

For $j=1,...,p$, we end the proof by replacing the $\beta_j$, $Q_1$ and $\beta_j$ by their estimators. Each of them converges in probabilities; $\hat{\beta}_0$ and $\hat{\beta}_p$ the asymptotics MLE proprieties and $\bar{Q}_1$ using the strong law of large number,

$$\hat{\beta}_j^{M_2}\xrightarrow{\mathbb{P}}\beta_j,\ \hat{\beta}_p^{M_2}\xrightarrow{\mathbb{P}}[0;\beta_p],\quad j\in 0,...,p-1. \tag{74}$$

□

### C.3.4 Proof of the Log-poisson results for (C2) assumption

*Proof.* Denote $\beta_* = (\beta_1, ..., \beta_p)$. In the case (C2) with perfectly correlated quality variables, *i.e.* $\Omega_j = \Omega_k \rightarrow Q_j = Q_k$ $(j \neq k)$, the equation under (Z-A2) can be written:

$$\begin{aligned}
\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta)) =& Q_1 \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) + (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z};\beta))) \\
=& Q_1 \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) + (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0))) \\
&+ (1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}^p} -ve^{\beta_0+\beta_*\mathbf{z}} + n(\beta_0+\beta_*\mathbf{z}) dF_{Z_1,...Z_p}(\mathbf{z}) dF_N(n) \\
=& Q_1 \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) + (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0))) \\
&+ (1-Q_1)\, v exp(\beta_0)\, M_{\mathbf{Z}}(\beta_*),
\end{aligned} \tag{75}$$

where $M_{\mathbf{Z}}(\beta_*)$ is the multivariate generating function of $Z_1, ..., Z_p$ and under (Z-A2) is equals to $M_{\mathbf{X}^{real}}(\beta_*)$. The first of *Bartlett identities*,

$$\begin{aligned}
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta^{M2}))}{\delta\beta} =& Q_1 \frac{\delta}{\delta\beta}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta^{M_2}))) + (1-Q_1)\frac{\delta}{\delta\beta}\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0^{M_2}))) \\
&+ (1-Q_1)V\frac{\delta}{\delta\beta}\Big(exp(\beta_0^{M_2})(1-M_{\mathbf{X}^{real}}(\beta_*^{M_2}))\Big) = 0,
\end{aligned} \tag{76}$$

does not permit to find a bound on $\beta^{M_2}$ (see the remark for log-Gamma GLM). However, $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta)))$ can be calculated using only $\mathbf{X}$,

$$\begin{aligned}
\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) = \frac{1}{Q_1}\Big(&\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta))) - (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0))) \\
&- (1-Q_1)\, V\, exp(\beta_0)\, M_{\mathbf{X}^{real}}(\beta_*)\Big).
\end{aligned} \tag{77}$$

□

## C.4 Proof for GLM log gamma

### C.4.1 Proof of the lemma 3

*Proof.* The expected log-likelihood to maximize is equivalent to:

$$\int_{\mathbb{R}^{p+1}} -y\, exp(-\mathbf{x}\beta) - \mathbf{x}\beta dF_{X_1,...,X_p,Y}(\mathbf{x},y). \tag{78}$$

The expected log likelihood $\mathbb{E}(log(f_Y(\hat{\beta};\mathbf{Y}|\mathbf{X})))$ is equal to

$$Q_p\ \mathbb{E}(log(f_Y(\hat{\beta};\mathbf{Y}|\mathbf{X}^{real}))) + (1-Q_p)\mathbb{E}(log(f_Y(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real})))M_{X_p}(-\hat{\beta}_p). \tag{79}$$

Under the assumption (X-A3), the derivative of the $M_2$ log-likelihood for $j$ in $\{0, ..., p-1\}$ are equals to

$$\begin{aligned}
\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) =& Q_p\ d_j(\beta) + (1-Q_p) \\
&\frac{\delta}{\delta\beta_j}\int_{\mathbb{R}}\int_{\mathbb{R}^p} -y\, exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p) - \mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real},y)dF_{Z_p}(\mathbf{z}_p) \\
=& Q_p\ d_j(\hat{\beta}) + (1-Q_p) \\
&\int_{\mathbb{R}}\int_{\mathbb{R}^p} -yx_j^{real} exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p)dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real},y)dF_{Z_p}(\mathbf{z}_p) \\
=& Q_p\ d_j(\beta) + (1-Q_p)\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta^{*p};\mathbf{Y}|\mathbf{X}^{real})))M_{X_p}(-\beta_p),
\end{aligned} \tag{80}$$

$$\begin{aligned}
\frac{\delta}{\delta\beta_p}\mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) =& Q_p\ d_p(\beta) + (1-Q_p) \\
&\int_{\mathbb{R}}\int_{\mathbb{R}^p} -yz_p exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p)dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real},y)dF_{Z_p}(\mathbf{z}_p) \\
=& Q_p\ d_j(\beta) - (1-Q_p)\mathbb{E}(log(f_Y(\beta^{*p};\mathbf{Y}|\mathbf{X}^{real})))M'_{X_p}(-\beta_p).
\end{aligned}$$

□

## C.5 Convexity: Propositions 1

*Proof.* Denote the covariates $X_j$, $X_k$ $(i \neq j)$ with a Pearson correlation $\rho$ for which $|\rho| \neq 1$ and suppose $\beta_k$ and $\beta_j$ non-null. Using the corollary of [4, Chatelain and Xavier, 2021], the following derivatives are found :

$$\begin{aligned}\frac{\delta \beta_k^{M_2}(Q_k|Q_j)}{\delta Q_k} &= A \times \frac{1 + Q_j^2 Q_k^2 \rho^2}{(1 - Q_j^2 Q_k^2 \rho^2)^2}, \\ \frac{\delta^2 \beta_k^{M_2}(Q_k|Q_j)}{\delta Q_k^2} &= A \times \frac{2 Q_j^2 Q_k \rho^2}{(1 - Q_j^2 Q_k^2 \rho^2)^3}(3 + Q_k^2 Q_j^2 \rho^2),\end{aligned} \tag{81}$$

with $A = \beta_k(1 - Q_j^2\rho^2) + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j \rho(1 - Q_j^2)$.

$A$ is positive only if $\rho\beta_k > -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j$. Indeed,

$$\begin{aligned}0 \leq & \beta_k(1 - Q_j^2\rho^2) + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j^{M_1}\rho(1 - Q_j^2) \\ 0 \leq & \beta_k + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j\rho - Q_j^2(\rho^2\beta_k - \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j\rho),\end{aligned} \tag{82}$$

$$\text{if } \rho \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k} \text{ and } \beta_k \geq 0 \text{ or } \rho \leq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k} \text{ and } \beta_k \leq 0.$$

Then $\beta_k^{M_2}(Q_k|Q_j))$ is convex if $\rho \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k}$ and $\beta_k \geq 0$ or $\rho \leq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k}$ and $\beta_k \leq 0$ and concave in the two other cases. If $Q_1$, $Q_2$ or $\rho$ are null, $\frac{\delta^2 \beta_k^{M_2}(Q_k)}{\delta Q_k^2} = 0$ which ends the proof. □