



HAL
open science

Integrating data quality into GLM for insurance pricing

Pierre Chatelain

► **To cite this version:**

| Pierre Chatelain. Integrating data quality into GLM for insurance pricing. 2020. hal-03252640v2

HAL Id: hal-03252640

<https://hal.science/hal-03252640v2>

Preprint submitted on 7 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Integrating data quality into GLM for insurance pricing

Chatelain Pierre¹

¹Univ 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

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

35 Literature suggests a multiple dimension analysis to evaluate data quality ([21,
36 todoran, 2014]). For instance, the completeness dimension of data is a research field
37 where numerous methods were developed to deal with missing values ([23, Van Buuren,
38 2018], [14, Little and Rubin, 2019]). These methods are globally based on assumptions
39 such as MCAR (Missing Completely At Random), MAR (Missing At Random) or MNAR
40 (Missing Not At Random). In the present paper, the credibility dimension will be studied
41 further. On the mismeasurement side of uncertainty dimension, some works exist using
42 trees algorithm ([22, Trabelsi et al., 2016], [20, Tami et al. 2018]) or the EIV-mismeasurment
43 ([24, Van Huffel and Lemmerling, 2013]) framework. On the credibility side of uncer-
44 tainty dimension, robust estimation theory as RANSAC (RANDOM SAmple Consensus,
45 [6, Fischler and Bolles, 1981]) algorithm and its different extensions such as KALMANSAC
46 [25, Vedaldi et al., 2005] deal with outliers and inliers mostly used for computer vision.
47 The downside of these methods is the left-aside observations for which no prediction can
48 be made. It would be operationally inconceivable that some contract may not be priced.

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

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

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

70 **Outline of the paper :** The paper is built as follows: in the section 2, the general frame-
71 work and the notation are introduced. This work specifies how uncertainty is integrated
72 in the covariate generating process. Section 3 gives the main algorithm and theoretical
73 results. Hereafter, a simulation study illustrates the results in the section 4. Next, section
74 5 brings close the different assumptions to actuarial uses. In detail, subsection 5.4 and
75 5.3 discuss the use of quality indexes and the case of imperfect data quality indexes. All
76 these remarks are illustrated by a practical case on household insurance.

77 2. Data problems and imputation

78 2.1 Notations

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

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

86 We want to take advantage of the exogenous information provided by an *individualized*
 87 quality index related to the confidence we can have about the $i - th$ observation of the
 88 $j - th$ covariate, further denoted Q_{ij} .

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

$$89 \quad \mathbf{X} = \mathbf{X}^{real} \circ \mathbf{\Omega} + \mathbf{Z} \circ (J_{n,(p+1)} - \mathbf{\Omega}), \quad (1)$$

where :

- 90 • \circ corresponds to the Hadamard product,
- 91 • $J_{n,(p+1)}$ is the $n \times (p+1)$ -identity matrix under Hadamard multiplication,
- 92 • $\mathbf{Z} = (Z_{ij}) \in \mathcal{M}_{n \times (p+1)}(\mathbb{R})$ are considered as the "wrong" covariate values having the
 93 same distribution as \mathbf{X}^{real} ,
- 94 • $\mathbf{\Omega} = (\omega_{ij}) \in \mathcal{M}_{n \times (p+1)}(\{0, 1\})$ is a binary mask indicating whether the $i - th$ observation
 95 of the $j - th$ covariate X_{ij} is perfectly observed or not. In other words, $\mathbf{\Omega}$ tells us
 96 if one observes the "real" observation or not. Assume that covariates distribution
 97 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 $(Y_i, \mathbf{X}_i)_{i=1, \dots, n}$ of (Y, \mathbf{X}) ,
 where $Y_i \in \mathbb{R}$ and $\mathbf{X}_i = (1, X_{i1}, \dots, X_{ip}) \in \mathbb{R}^{p+1}$. The vector of quality indexes of the i -th row is
 written $\mathbf{Q}_i = (1, Q_{i1}, \dots, Q_{in})$. 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 $i = 1, \dots, n, j = 1, \dots, p$ the quality index
 Q_{ij} is equal to :

$$98 \quad Q_{ij} = \mathbb{E}(\omega_{ij}) = \mathbb{P}(\omega_{ij} = 1) = \begin{cases} \mathbb{P}(X_{ij} = X_{ij}^{real}) & \text{if } \mathbf{X}_j \text{ is continuous variable,} \\ \mathbb{P}(X_{ij} = X_{ij}^{real}) - \mathbb{P}(X_j^{real} = X_{ij}^{real}) & \text{if } \mathbf{X}_j \text{ is discrete variable.} \end{cases} \quad (2)$$

99 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
 100 the probability to get the true value randomly. In other words, $Q_{ij} = 0$ means that the
 101 value X_{ij} is not informative on the risk of i . Denote for the rest of the paper ($j = 1, \dots, p$),
 102 $\bar{Q}_j = \frac{1}{n} \sum_{i=1}^n Q_{ij}$ and assume $\bar{Q}_j \neq 0$. This assumption is not restrictive, especially for real-life
 103 applications where such covariates would simply be removed from the data. However,
 104 it does not mean that an individual having all quality indexes null does not exist.

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

107 **2.2 Inapplicability of basic recommendations**

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

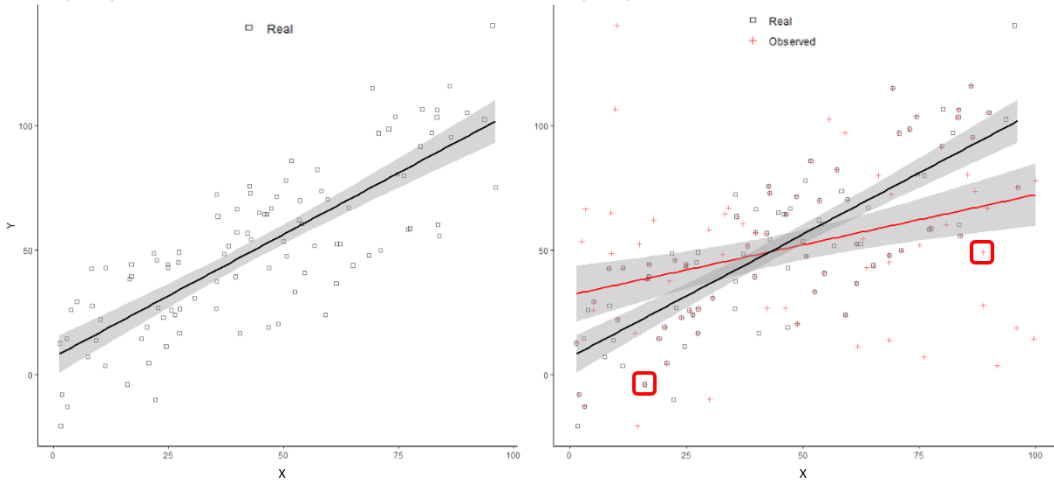


Figure 1: An univariate example where black squares are the real observations and crosses are the observed values. This graph is based on simulated data with $X \sim \Gamma(1, 2)$ and $Y = 10 + 1X$, Q follows a uniform distribution between 0.2 and 1. The data set \mathbf{X} is created with the previous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

111 **Imputating:** Let consider the strategy to impute new values on outliers or low qual-
 112 ity observations¹. Defining outliers in the multivariate case when the others covariates
 113 are not good quality is difficult. This is even more true in actuarial pricing where the
 114 outcome to model - claim cost, claim frequency, retention rate ..., has an intrinsic vari-
 115 ability. Without taking into account exogenous information, robust estimation theory as
 116 RANSAC (RANDOM SAMPLE CONSENSUS, [6, Fischler and Bolles, 1981]) algorithm and its
 117 different extension have been developed using only a subsample of "real observation"
 118 (inliers) in modelling. Straightforwardly, in our framework, the data quality influences
 119 the definition of outliers for a regression, as shown in Figure 1. Indeed, the outliers'
 120 detection is bias due to the data set's quality. In that situation, some perfectly observed
 121 observations may be defined as outliers and the goal is also to predict values for indi-
 122 vidual with wrong observation(s). In the multivariate case and with variance outcome
 123 increasing, the definition of an outlier is operationally even more complex to deal with.
 124 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
 125 wrong ?

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

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

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

133 For our framework, let assume assumptions similar to Zhu et al. 2019 [26], *i.e.* in
 134 the case of complete independence of quality and observations². Assume all the quality
 135 indexes independently distributed as an *Uniform*(0.4, 0.8). Not only the low quality of
 136 the data implies a small threshold, but the different observations would highly range
 137 around the mean value. For a threshold of 0.5 and 10 variables defined as before, only
 138 6 % rows would have all its covariates above the threshold in average. Besides, errors
 139 can be correlated spatially and this filtering process may bias the portfolio risks. For
 140 open data used in household insurance, this is in particularly true for urban area zones
 141 : covariates have often lower quality in rural areas. In short, filtering strategies are not
 142 optimal. Finally, neither imputing nor deleting are correcting the impact of quality on
 143 models.

144 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.

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

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

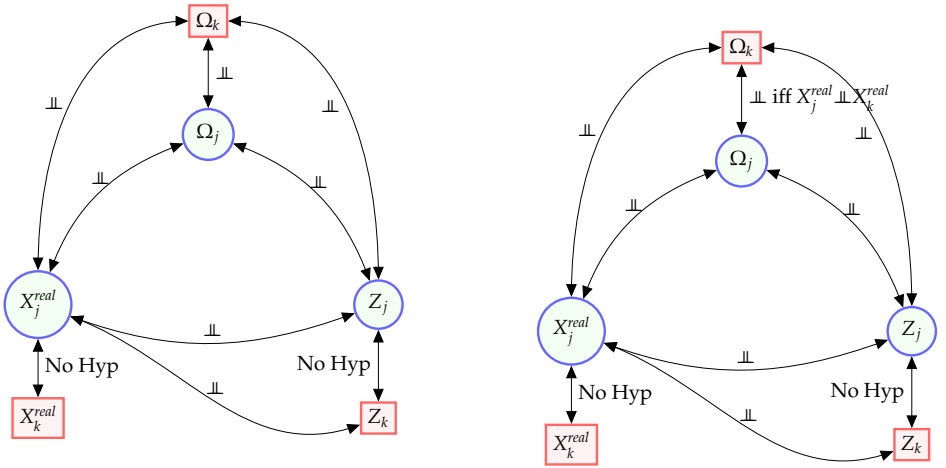
160 In our framework, quality indexes are associated to values between 0 and 1 as shown
 161 in example 1. In real-life application, quality indexes are exogenous information given
 162 by the data provider and take qualitative values such as "very high", "high", "medium",

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

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

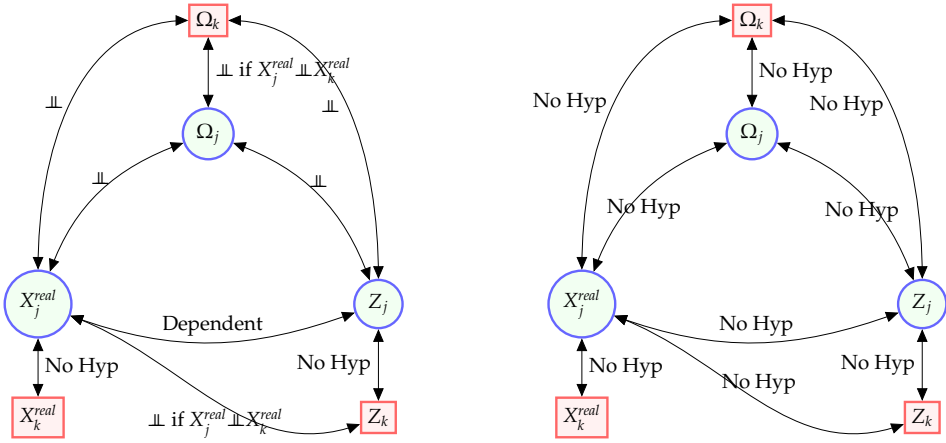
165 **2.4 Frameworks under study**

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

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

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

190 3. Estimation Process

191 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

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

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

194 3.2 Prediction using quality index

This section studies in the sequel GLM given by

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

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

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

198 Denote the following naming :

³See 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 $\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q})$. 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, β :

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

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 β^K .

For all the proofs, the variables are supposed centred.

3.3 Algorithm 3 for linear regression and GLM

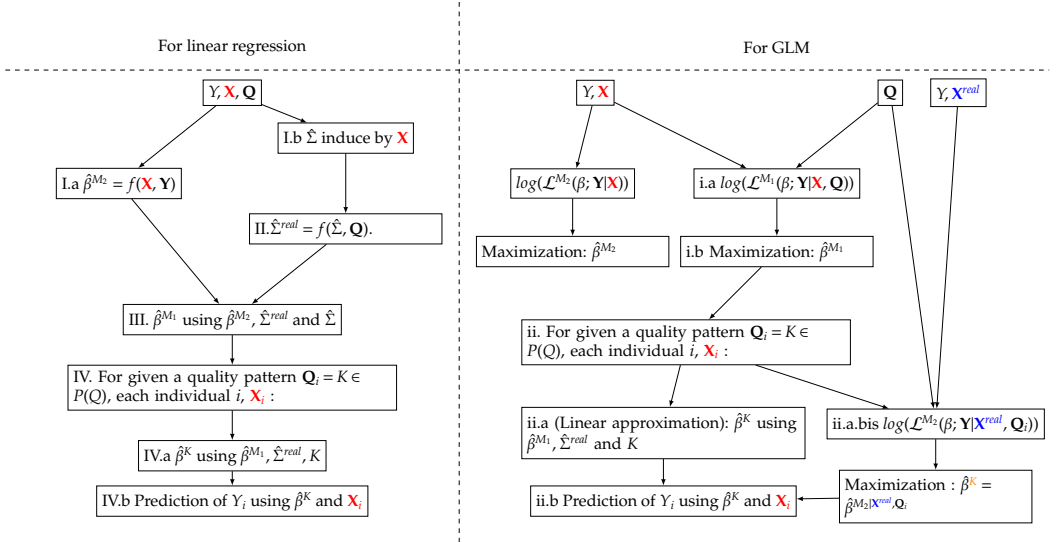


Figure 3: Process to take into account the quality index for linear regression and an approximation for GLM log-Poisson. In this way, this process adjusts the coefficient to each quality pattern.

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

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

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

221 In the event that \mathbf{X}^{real} is known, or a large enough sample \mathbf{X} is perfectly observed, $\hat{\beta}^K$
 222 could be directly estimated from the maximization of the likelihood $\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q} =$
 223 $J_{n,1}K)$ (see section A.3). If the correlation structure of \mathbf{X}^{real} is the same as \mathbf{X} one, another
 224 solution would be to simulate a new \mathbf{Y}^{new} using \mathbf{X} and $\hat{\beta}^{M_1}$ to apply an estimator proposed
 225 in the subsection 3.6.

226 3.4 Assumptions under study

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

- 231 (X-A1) All the random variables X_j^{real} ($j = 1, \dots, p$) are independent.
- 232 (X-A2) Each variable X_j^{real} is correlated with only one variable X_k^{real} ($j \neq k$).
- 233 (X-A3) For all $k \neq p$, the variable X_k^{real} is independent of X_p^{real} and $\bar{Q}_k = 1$.
- 234 (Z-A1) All the random variables Z_j and Z_k are independent.
- 235 (Z-A2) (Z_j, Z_k) has the same correlation structure than (X_j^{real}, X_k^{real}) , $j \neq k$.

236 For GLM, correlation between imperfectly observed covariates, such as (X-A2) are not
 237 considered. However, for linear regression, (X-A2) is taken into account in [4]. When the
 238 assumption (X-A3) is studied, denote $\mathbf{X}_{(sp)} = (1, X_1; \dots; X_{p-1})$ and it's observed sample
 239 $\mathbf{X}_{i;(sp)}$. In the same way, $\beta^{(sp)}$ refers to $(\beta_0, \dots, \beta_{p-1})$.

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

246 **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 $\beta = (\beta_0, \dots, \beta_p)^T$ is found by maximizing the likelihood or log-likelihood (ML- Maximum likelihood)⁴;

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

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

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

258 **3.6 Deduce $\beta^{M_3|K}$**

259 As already mentioned in Section 3.2, the vector β^K exactly matches β^{M_1} when all indi-
260 vidualized quality indexes equal to 1, i.e. when $K = J_{1,p+1}$. In full generality, when $K = \mathbf{Q}_i$
261 is made of terms $Q_{ij} \neq 1$, the coefficients $\hat{\beta}^K$ need to be calculated. $\hat{\beta}^{K=Q_i}$ is an estimator
262 of β^{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
263 coefficient $\hat{\beta}^K$ is the one minimizing the mean $Dev(\hat{\beta}^K|\mathbf{X}, \mathbf{Y})$ for a given pattern of quality
264 K as wanted (see the equation 3).

265 For any distribution and link function, it is possible to estimate the expected M_2
266 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 β . Knowing $(\mathbf{Y}, \mathbf{X}^{real}, \mathbf{Q})$, a sample estimator of $\mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$ is*

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

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

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

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

⁴Or 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 $\mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$

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

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

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

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 β^{M_2} through \mathbf{X}^{real} for any \mathbf{Q} . $\hat{\beta}^{K=\mathbf{Q}_i}$ is an estimator of β^{M_2} when the model is fitted on dataset \mathbf{X} but in the case $\mathbf{Q} = \mathbf{J}_{n,1} \mathbf{Q}_i$.

3.7 Deduce β^{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 $\log(\mathcal{L}(\hat{\beta}; \mathbf{Y} | \mathbf{X}^{\text{real}}, \mathbf{Q}))$ can be found as follows :

$$\begin{aligned} \log(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{Y} | \mathbf{X}, \mathbf{Q})) &= \frac{1}{\bar{Q}_p} \left[\log(\mathcal{L}^{M_2}(\hat{\beta}; \mathbf{Y} | \mathbf{X})) \right. \\ & \quad - (1 - \bar{Q}_p) \times \log(\mathcal{L}(\hat{\beta}^{*p}; \mathbf{Y} | \mathbf{X}_{(sp)}^{\text{real}})) \\ & \quad \left. - (1 - \bar{Q}_p) \times \sum_{i=1}^n v_i e^{\hat{\beta}^{*p} \mathbf{X}_{i^{(*)}}^{\text{real}}} (1 - M_{\mathbf{X}_p}(\hat{\beta}_p)) \right]. \end{aligned} \quad (7)$$

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

⁵i.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 :

$$\begin{aligned}
 \log(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{Y}|\mathbf{X}, \mathbf{Q})) &= \sum_{K_p \in P(\mathbf{Q}_p), K_p \neq 0} \frac{1}{K_p} \left[\log(\mathcal{L}^{M_2}(\hat{\beta}; \mathbf{Y}|\mathbf{X}_{Q_p=K_p}) \right. \\
 &\quad - (1 - K_p) \times \log(\mathcal{L}(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}_{(*)p}^{real}; Q_p=K_p)) \\
 &\quad \left. - (1 - K_p) \times \sum_{i=1}^n v_i e^{\hat{\beta}^{*p} \mathbf{X}_{i(*)p}^{real}} (1 - M_{X_p}(\hat{\beta}_p)) \right].
 \end{aligned} \tag{8}$$

287 where $\mathbf{X}_{Q=K_p}$ represents the dataset where only individual i such as $Q_{i,p} = K_p$ are kept.
 288 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
 289 construction than the equation 7. Individual having null quality index are not taken into
 290 account. Therefore, in the following part, the second estimator will be used. These two
 291 estimators converge in probabilities to $\log(\mathcal{L}(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q}))$. In the same way, the solution
 292 of the maximum likelihood converges in probability.

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

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

3.8 Estimators proprieties for other distributions

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

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

317 **Multiplicative structures:** In the case (C1), Log-Poisson and Log-Gamma GLM's multi-
318 plicative structure provides the calculus of $\log(\mathcal{L}^{M_1}(\beta; \mathbf{Y}|\mathbf{X}, \mathbf{Q}))$. However, in multivariate
319 case, under (X-A3) and (Z-A1), $\beta_p^{M_2}$ depends on Q_p , the moment generating function
320 $M_{X_p}(t)$ and β_j for $j = 1, \dots, p - 1$. The main difference is that $\beta_j^{M_2}$ depends on the distribu-
321 tion of X_p . For Log-Poisson model, $\beta_j^{M_2} = \beta_j$ and $\beta_p^{M_2}$ convergences in probability in an
322 interval $[0, \beta_p]$. However, for Log-Gamma this property is not true and the other coeffi-
323 cients, $\beta_j^{M_2}$, are also impacted by Q_p . In the case (C2), regrettably, no proprieties on the
324 estimator can be state for Log-Gamma and Log-Poisson GLM.

325 4. Simulation study - M1 estimator

326 This section aims to check our theoretical results on the estimator properties for Log-
327 Poisson GLM. In this view, all the simulated examples are created using the following
328 steps involving all the aforementioned quantities required to generate the right data :

329 **Step 1:** Q is in practice given. For the simulation, it is randomly generated;

330 **Step 2:** X^{real} is simulated given the marginals and the correlation structure;

331 **Step 3:** $Z = (Z_1, \dots, Z_p)$ is simulated given X^{real} and the assumptions;

332 **Step 4:** Y is simulated from its relationship with X^{real} ;

333 **Step 5:** Ω is simulated from Q through Bernoulli trials;

334 **Step 6:** X is deduced thanks to the equation (1).

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

336 4.1 Find $\hat{\beta}^{M_1}$ coefficients

337 Let $\mathbb{E}(Y|X^{real}) = \exp(1 + 0.4X_1^{real} + 0.5X_2^{real} + 0.6X_3^{real} + 0.07X_4^{real})$ with $X_1 \sim \Gamma(2, 1)$, $X_2^{real} \sim$
338 $\mathcal{N}(0, 1)$, $X_3 \sim \mathcal{Pois}(2)$, $X_4 \sim \mathcal{N}(0, 10)$ and Y following a Poisson distribution. The quality
339 index follows an independent discrete distribution on the values (0.5; 0.75; 1) with the
340 probability (0.25; 0.25; 0.5) for Q_4 . Let all the other covariates be perfectly observed, *e.g.*
341 $Q_{i,j} = 1$ for all $i \in 1, \dots, n$ and $j \in \{1, 2, 3\}$.

342 Using the precedent result, M_1 likelihood can be estimated as shown in figure 4.
343 The use of imperfectly observed dataset implies a wider variance of the estimator M_1
344 than the real model one. Here, the first estimator has wider variance than the second
345 estimator. As shown by equation 37, the coefficients β_1 , β_2 and β_3 did not change due to
346 the independence in between the variables - figure 5 - and the coefficient associated to X_4
347 is corrected - figure 6.

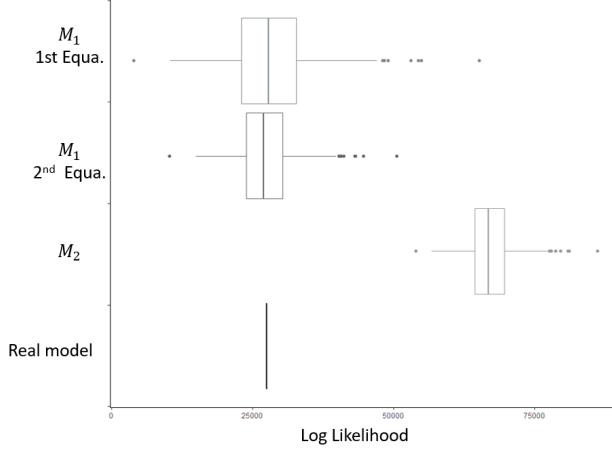


Figure 4: Estimation of the M_1 log-likelihood for log-Poisson GLM using equation 7 for a given \mathbf{X} and \mathbf{Q} . The moment function is estimated using its empirical estimator. The true function leads to the same graph but with a smaller variance. 2000 simulations are done for a given \mathbf{X}^{real}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

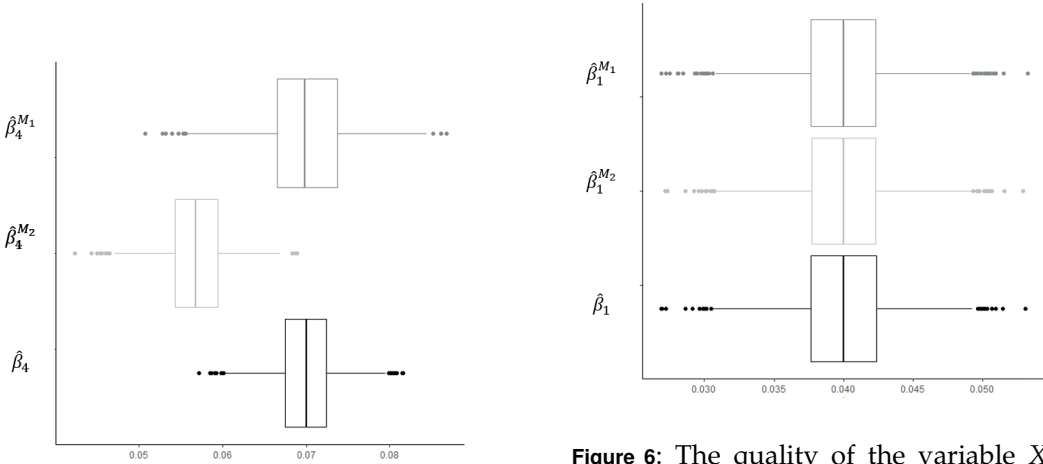


Figure 5: $\hat{\beta}_4^{M_2}$ is smaller than $\hat{\beta}_4$ because of the quality of the variable. $\hat{\beta}_4^{M_1}$ is unbiased, but has a wider variance than the real coefficient.

Figure 6: The quality of the variable X_4 does not impact the estimation of β_1 , β_2 , β_3 ; here, highlighted by β_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 β and β^{M_2} or β^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 β_4 (≈ 0.07), the

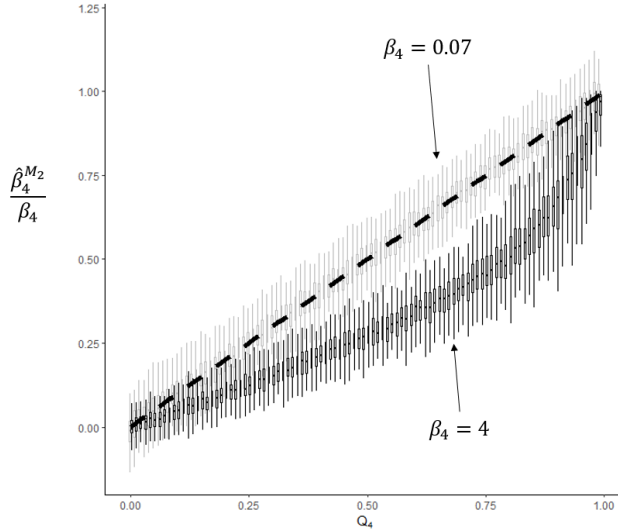


Figure 7: 1000 simulation for each quality. As for linear regressions, a linear evolution through the quality can be seen for low coefficient, however for higher values the relationship is not proportional to the quality.

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

357 5. Discussion

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

365 5.1 Examples

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

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

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

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

390 **Example of case C2:** The living surface, the number of rooms and the footprint are
391 globally one of the most segmenting features in household pricing. Different data sets and
392 methods are available in France to collect them, such as DVF⁶. This database geolocates
393 parcels and contains different features such as the value of property values, the number
394 of rooms, the surface of the parcel or the living surface among others. The database is
395 created from all properties transfer since 2015. *e.* On the uncertainty dimension, errors
396 are coming from the link between geocoding and the address or between the address and
397 the building, each of these steps impacts the data's quality depending on the feature. A
398 wrong geocoding would imply that the observations are taken from another building.
399 For all these variables, the case (C2) and the assumption (Z-A2) would be appropriate
400 since they are collected from the same building.

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

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

419 One of the most difficult cases is when the quality depends on others variables; for
420 instance the material of the roof and the analysis of a roof to detect a window - see figure
421 8 and 9. In this case, the modality of dark slate informs on the risk, not because dark slate
422 changes it but due to the low quality of the variable roof-windows associated to it.

⁶This 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.

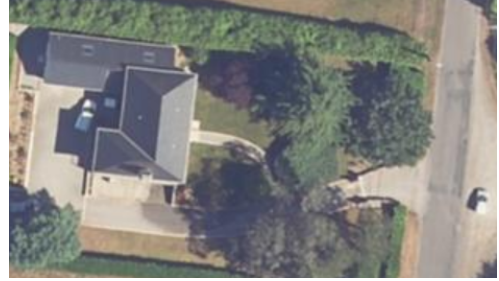


Figure 9: The detection of a window on a roof is harder because of the dark slate roof from the IGN cartography (47.179068, -1.814216) - © IGN 2018

5.2 Actuarial justification of this framework assumptions

Integrity of dataset and assumption on Z : In all examples seen above, the wrong observations Z are coming from real individual; it justifies the main assumption that “wrong” values Z_j follow the same distribution as 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 β 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 β , adapted to perfectly observed variables, and not β^{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 :

$$E[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (9)$$

and ρ 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,

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

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

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

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 β (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,

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

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.

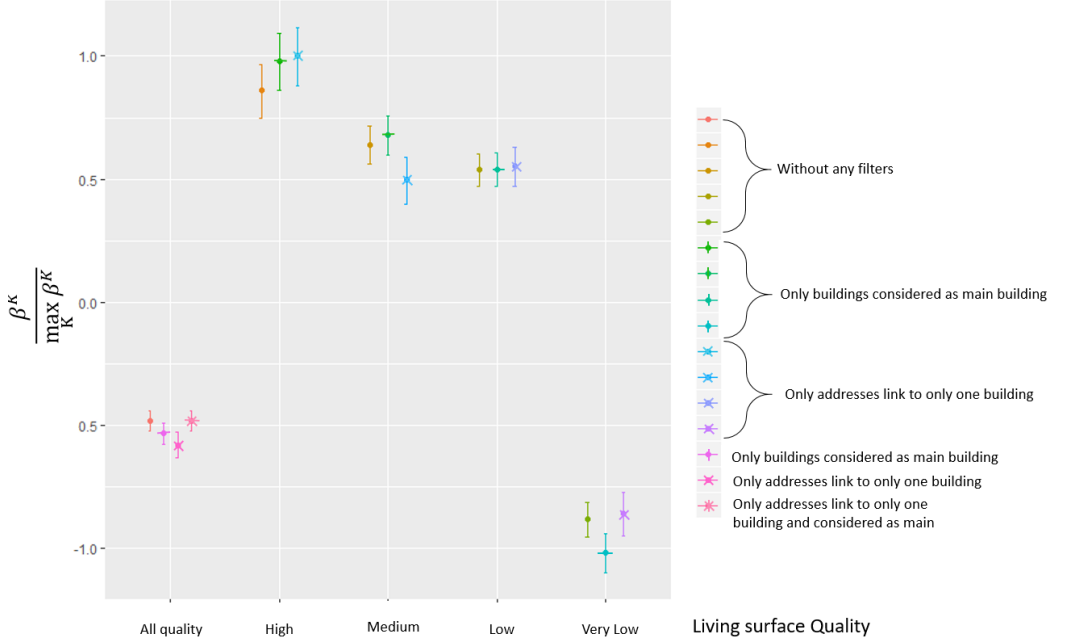


Figure 10: Ratio of the living surface coefficient of Log-Poisson. Because they are too few “very high” quality observations, they were regrouped with “high” quality observations. Different filters based the building geolocation information are done on the data set to challenge the quality index.

480 Then, “medium” quality value would be estimated by 0.6, “low” quality value by 0.5.
 481 However, the coefficient of very low quality values has an opposite sign. In fact, very low
 482 quality values are link to rural zone. Therefore, the case (C4) is the most appropriate and
 483 our evaluation method can not be used. In the same way, “low” quality values are also
 484 more link to rural density than “medium” one or “high”⁷. In consequence, the associated
 485 value to medium quality 0.5 can be debated. Indeed, the “low” and “very low” quality
 486 are correlated with others characteristics impacting the risks. The coefficients calculated
 487 on the database are therefore impacted. This is an important limit of our framework. In
 488 such case, using a threshold to set aside “very low” quality observations is recommended
 489 so that the data set verifies our assumptions. The different filters on geocoding show that
 490 leaner detail could be added within a value of the quality index. In this case, a modality
 491 may regroup different levels of quality. In other word, quality index is not perfectly
 492 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_α and n_κ number observations with the quality \bar{Q}_α and \bar{Q}_κ 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):

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

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

493 Equation 13 can be easily extended to higher dimension. If groups of different quality are
 494 mixed together and are given the same quality index value, the best one should be the
 495 pondered mean of each quality in a context of linear regression with the assumption (X-
 496 A1). However, under (X-A2) (with correlation), the aggregation of the quality influences
 497 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 $\rho\beta_k^{M_1} \geq -\sqrt{\frac{Var(X_i)}{Var(X_k)}}\beta_j^{M_1}$,

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

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

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

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

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

504 6. Conclusion

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

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

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

523 Conflict of interest

524 The author declares no conflict of interest.

525 **References**

526 [1] Autorité de contrôle prudentiel, ACPR. Synthèse de l'enquête déclarative de 2019 sur la gestion des données
527 alimentant les calculs prudentiels des organismes d'assurance. Technical Report 119, ACPR, Jan 2021.

528 [2] Ole Barndorff-Nielsen. Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of*
529 *statistics*, pages 151–157, 1978.

530 [3] Robert Campbell, Louise Francis, V Prevosto, Mark Rothwell, and Simon Sheaf. Report of the data quality
531 working party. Technical report, 2006.

532 [4] Pierre Chatelain and Xavier Milhaud. Linear regression and data quality through individualized credibility
533 index. preprint, May 2021.

534 [5] R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

535 [6] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with
536 applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981.

537 [7] Louise A Francis. Dancing with dirty data methods for exploring and cleaning data. 2005.

538 [8] General Committee of the Actuarial Standards Board and Applies to All Practice Areas, GCASB. Data quality
539 - revised edition. Technical Report 23, Actuarial Standard of Practice, May 2014.

540 [9] Ali S Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B*
541 *(Methodological)*, 54(3):761–771, 1992.

542 [10] Daniel F Heitjan and Srabashi Basu. Distinguishing “missing at random” and “missing completely at
543 random”. *The American Statistician*, 50(3):207–213, 1996.

544 [11] Sham Kakade, Ohad Shamir, Karthik Sindhara, and Ambuj Tewari. Learning exponential families in high-
545 dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial*
546 *intelligence and statistics*, pages 381–388. JMLR Workshop and Conference Proceedings, 2010.

547 [12] Kananart Kuwarananchroen and Shreyas Sundaram. On the location of the minimizer of the sum of two
548 strongly convex functions. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1769–1774. IEEE,
549 2018.

550 [13] EL Lehmann and George Casella. Unbiasedness. *Theory of Point Estimation*, pages 83–146, 1998.

551 [14] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons,
552 2019.

553 [15] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical*
554 *Society: Series A (General)*, 135(3):370–384, 1972.

555 [16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing,
556 Vienna, Austria, 2019.

557 [17] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

558 [18] Shaun Seaman and Ian White. Inverse probability weighting with missing predictors of treatment assignment
559 or missingness. *Communications in Statistics-Theory and Methods*, 43(16):3499–3515, 2014.

560 [19] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data.
561 *Statistical methods in medical research*, 22(3):278–295, 2013.

562 [20] Myriam Tami, Marianne Clausel, Emilie Devijver, Adrien Dulac, Eric Gaussier, Stefan Janaqi, and Meriam
563 Chebre. Uncertain trees: Dealing with uncertain inputs in regression trees. *arXiv preprint arXiv:1810.11698*,
564 2018.

565 [21] Ion-George Todoran, Laurent Lecornu, Ali Khenchaf, and Jean-Marc Le Caillec. Toward the quality evaluation
566 of complex information systems. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*,
567 volume 9091, page 90910N. International Society for Optics and Photonics, 2014.

568 [22] Asma Trabelsi, Zied Elouedi, and Eric Lefevre. Handling uncertain attribute values in decision tree classifier
569 using the belief function theory. In *International conference on artificial intelligence: Methodology, systems, and*
570 *applications*, pages 26–35. Springer, 2016.

571 [23] Stef Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.

572 [24] Sabine Van Huffel and Philippe Lemmerling. *Total least squares and errors-in-variables modeling: analysis,*
573 *algorithms and applications*. Springer Science & Business Media, 2013.

574 [25] Andrea Vedaldi, Hailin Jin, Paolo Favaro, and Stefano Soatto. Kalmansac: Robust filtering by consensus. In
575 *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 633–640. IEEE,
576 2005.

577 [26] Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with
578 heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.

579 Appendix A. Theoretical framework

580 A.1 The covariance impacted by quality index

581 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
 582 following lemma states the relation between real covariance Cov_{jk}^{real} and the observed covariance Cov_{jk}^{real} under
 583 various assumptions :

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

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

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

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 :

$$\text{Under (Z-A1)} \quad Cov_{jk} = Q_j \times Cov_{jk}^{real}, \quad (18)$$

$$\text{Under (Z-A2)} \quad Cov_{jk} = Cov_{jk}^{real}. \quad (19)$$

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

590 A.2 Regression model under consideration

Given the independent variables (Y_1, \dots, Y_n) , the corresponding explanatory variables (X_1, \dots, X_n) , and individualized quality indexes (Q_1, \dots, Q_n) where $Q_i = (Q_{i1}, \dots, Q_{ip})$, 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 $X\beta$ and a link function g defined such as

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

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

595 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 Ω is independent from the others variables. For β in \mathbb{R}^2 , the model M_2 maximizes the following log-likelihood

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

596 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
 597 and Z_1 is given.

In the case (C1), (C2) and (C3), the quality variable Ω is independent from the others variables, which leads to :

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

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

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

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 :

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

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 β , $\log(f_Y(y|\mathbf{z};\beta))$ is integrable i.e $\mathbb{E}(\log(f_Y(Y|\mathbf{Z};\beta))) < +\infty$.

Because \mathbf{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 $\int_{\mathbb{R}^2} |\log(f_Y(y|\mathbf{z};\beta))| dF_{Z_1}(\mathbf{z}) dF_Y(y) < \infty$ implies $\int_{\mathbb{R}} |\log(f_Y(y|\mathbf{z};\beta))| dF_{Z_1}(\mathbf{z}) < \infty$ for any value of y and β . Remind that \mathbf{Z}_1 distribution has the same distribution than X_1^{real} . Hereafter, the canonical link function is used. In this case, the log-likelihood maximized can be written as follows :

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

where C^{st} is a constant independent of β and X and $b(\cdot)$ 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

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

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

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

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 $\mathbb{E}(\log(f_Y(Y|\mathbf{Z}_1;\beta)))$ with X_1^{real} ? In the multivariate case, the value Z_{i1} could be estimated using the other covariables depending on the case. If X_1^{real} and \mathbf{Z}_1 are correlated or dependent, a function g could exist such as $g(X_1^{real})$ is a good estimator of \mathbf{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 \mathbf{Z}_1 are completely independent.

The following estimator,

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

converges almost surely to $\mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X};\beta)))$ (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 $\mathbf{X}_{(sp)} = (1, X_1; \dots; X_{p-1})$ and its observed sample $X_{i,(sp)}$. In the same way, β^{sp} refers to $(\beta_0, \dots, \beta_{p-1})$. The expected likelihood

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

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 β and β^{M2} .

615 **A.4 Example 1: Log-Gaussian GLM**

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

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

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

618 where for the purposes of notation $\mathbf{X}_i = {}^t(1, \mathbf{x}_{i1})$ and $\beta = {}^t(\beta_0, \beta_1)$.

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

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

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

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

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

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

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

621 The proof of these results are in C.2. The paper [4, Chatelain and Milhaud, 2021] proved the following theorem,
 622 where ρ 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 $n \rightarrow +\infty$:

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

623 **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 $M_{\mathbf{X}_p^{real}}(t) = M_{\mathbf{X}_p}(t) = M_{Z_p}(t)$ 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

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

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

$$\begin{aligned} \log(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{Y}|\mathbf{X}, \mathbf{Q})) &= \frac{1}{\bar{Q}_p} (\log(\mathcal{L}^{M_2}(\hat{\beta}; \mathbf{Y}|\mathbf{X})) - (1-\bar{Q}_p) \times \log(\mathcal{L}(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}_{(sp)}))) \\ &- (1-\bar{Q}_p) \times \sum_{i=1}^n V_i e^{\hat{\beta}^{*p} \mathbf{X}_{i(sp)}^{real}} (1 - M_{\mathbf{X}_p}(\hat{\beta}_p)). \end{aligned} \quad (35)$$

The expected likelihood can be bounded :

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

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

627 *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 $\{0, \dots, p-1\}$ are equal to :

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

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

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

630 When $\hat{\beta}_p \rightarrow 0$, $M'_{X_p}(\hat{\beta}_p) \rightarrow 0$. It can be easily shown that the derivatives - equation 37 are a constant function of
631 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 $j \in \{1, \dots, p-1\}$ and $Q_p \in (0, 1)$,

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

632 is a monotonic function of the quality.

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

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

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

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) :

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

634 where $M_{\mathbf{X}^{real}}(\beta_*)$ is the multivariate generating function of $\mathbf{X}_1^{real}, \dots, \mathbf{X}_p^{real}$ and $\beta_* = {}^t(\beta_1, \dots, \beta_p)$. Unfortunately, no
635 bounds explicit can be state. The paper [12, kuwarananchaoren and Sundaram, 2018] provides an upper bound
636 on the location of local minimum of the sum of two strongly convex function under the assumption of bounded
637 gradient. The difficulty is that the log-likelihood exponential family is almost strongly convex, i.e. strongly convex
638 in the neighbourhood of β as proved in [11, Kakade et al. 2010].

639 **A.6 Example 3: Log-Gamma GLM**

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

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

640 Here, the only interest is to maximize the log likelihood according to β for a known ν . Therefore, the expected
641 log-likelihood will be studied,

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

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

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

The M_1 estimator can be calculated

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

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 $\{0, \dots, p-1\}$ are equal to :

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

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

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

643 The lemma 3 can not lead to a theorem like in the Log-Poisson case. The minimization of a sum of concave
644 function in \mathbb{R}^{p+1} does not necessary lead to $\beta_j^{M_2} \in [\min(\beta_j^{-p}, \beta_j), \max(\beta_j^{-p}, \beta_j)]$ where β_j^{-p} is the maximum likelihood
645 estimator $\mathbb{E}(\log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}_{(sp)}^{real})))$ and with $\beta_p^{-p} = 0$. Therefore, the Log-Gamma coefficients' evolution are not easily
646 bounded in the general case. Nevertheless, the $\beta_j^{M_2}$ 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 :

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

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

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

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

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

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

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

651 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
652 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 :

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

653 can not be calculated using only X^{real} . Looking for an approximation, $\log(1 + \exp(x)) = \log(2 + \exp(x) - 1) = \log(2) +$
 654 $\log(1 + (\exp(x) - 1)/2) \sim \log(2) + (\exp(x) - 1)/2 + o((\exp(x) - 1))$ when $\exp(x)$ is close to 1. Therefore, when $x\beta$ is close
 655 to 0, a fine approximation with a multiplicative structure can be state.

656 Appendix B. Various operational remarks

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

658 B.1 Quality impact and attenuation

659 The different results show that the “attenuation”⁸ on $\hat{\beta}$ due to data quality can be explained. However, the quality
 660 impacts might not always decrease the coefficients, as shown in [4, Chatelain and Millhaud, 2021]. The quality of a
 661 variable impacts all coefficients related to other correlated variables in the uncentred case. Figure 11 under (X-A2)
 662 and (Z-A1) in linear regression shows that even in the simple case the “attenuation” is not always true. With some
 663 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.

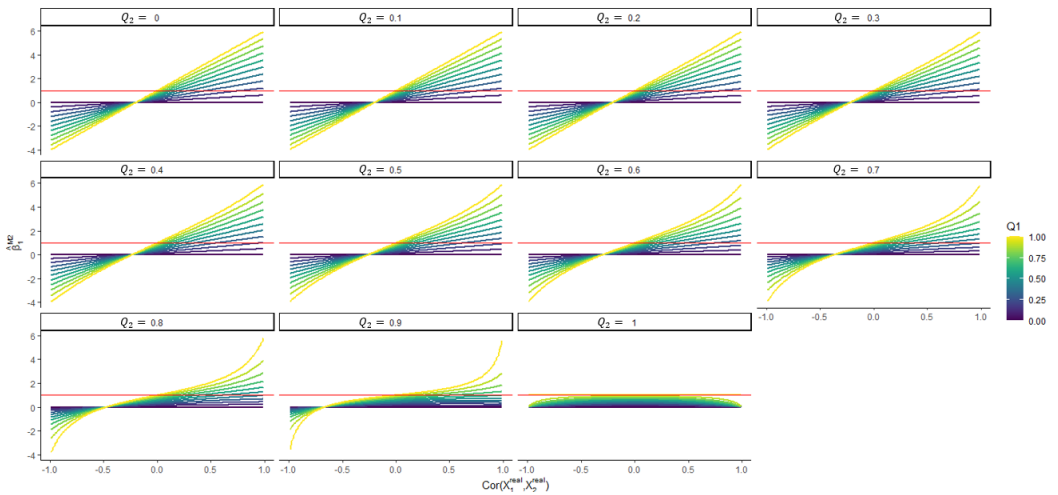


Figure 11: Log-Gaussian GLM: Value of $\beta_1^{M_2}$ wrapped by Q_2 and grouped by Q_1 . The coefficient β are all equal to 1 and the ratio of the standard deviation $\sqrt{\text{Var}(X_1^{real})/\text{Var}(X_2^{real})}$ equals to 6. The red straight line represents, β_1 which is equal to 1.

664 This is especially harmful to insurance pricing, where covariates’ choice must be justified by their impacts.
 665 Indeed, some coefficients may seem counter-intuitive due to quality impacts. Figures 11 and 12 provide an
 666 illustration of the impact of Q_1 depending on Q_2 (β_1, β_2 always equal to 1). The coefficients’ evolution is not linear
 667 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
 668 the coefficients could be considered as null even if the variable’s quality is not low. For instance, for $Q_2 = 0.7$
 669 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
 670 coefficient is different from 0. Moreover, by finding the β^{M_1} - thanks to \mathbf{X} and \mathbf{Q} , the modeler can find the “real”
 671 impact of a variable in models, thus justifying it.
 672

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

⁸As called in the econometric literature.

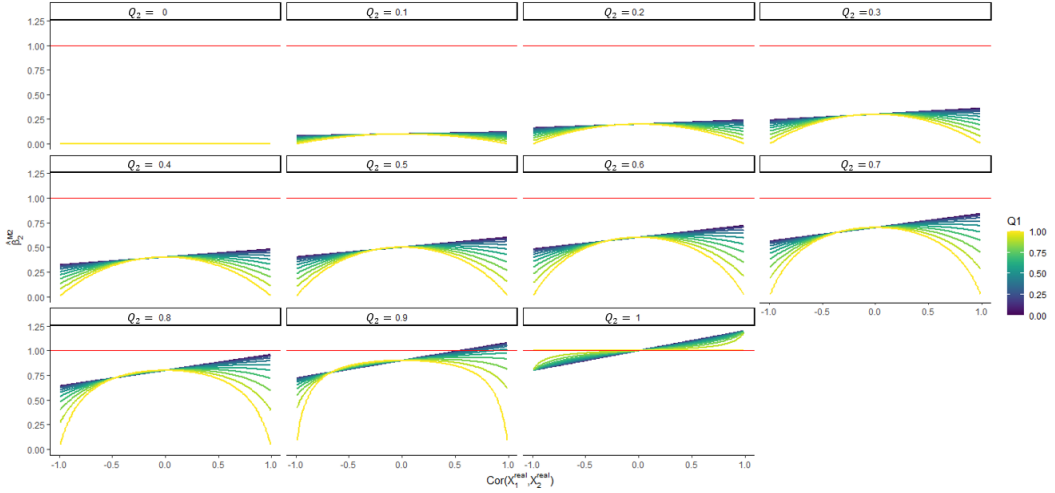


Figure 12: Log-Gaussian GLM: Value of β_2^{M2} wrapped by Q_2 and grouped by Q_1 . The coefficient β are all equal to 1 and the ratio of the standard deviation $\sqrt{\text{Var}(X_1^{\text{real}})/\text{Var}(X_2^{\text{real}})}$ equals to 6.

676 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 ρ 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

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

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

$$\hat{Y}_i = \beta_0^K + \beta_1^K Z_{i,1} + \beta_2^K X_{i,2}^{\text{real}}, \quad (51)$$

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, $j \in \{0; 1; 2\}$ 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

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

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

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

677 Thus, the equivalent imputation here for $X_{i,1}^{\text{real}}$ should be $E(X_1^{\text{real}}) - \sqrt{\frac{\text{Var}(X_2^{\text{real}})}{\text{Var}(X_1^{\text{real}})}} \rho (E(X_2^{\text{real}}) - X_{i,2})$. This imputation
 678 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
 679 according to the linear regression. In fact, $\sqrt{\frac{\text{Var}(X_2^{\text{real}})}{\text{Var}(X_1^{\text{real}})}} \beta_1^{M1} \rho E(X_2^{\text{real}})$ corresponds to the linear part of the information
 680 X_1^{real} already taken into account by X_2^{real} .

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

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

689 Appendix C. Proofs

690 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,

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

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

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

Because Z_1 have the same distribution as the $X_{11}^{real}, X_{11}^{real}$ can be used to estimate the density f_{Z_1} so $dF_{Z_1}(s) = dF_{X_{11}^{real}}(s)$. Finally, the previous equation can be estimated by the mean sample. Because $\{X_{11}^{real}, \dots, X_{n1}^{real}\}$ are *i.i.d* observations, the sample estimator would be

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

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,

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

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

692 Finally, the following points are true :

- 693 • observations are i.i.d and the density is Lebesgue measurable;
- 694 • the parameter space of β is compact and open;
- 695 • the previous estimator is concave as sum of concave function and is differentiable according to β ;
- 696 • Identifiability : the estimator function is a smooth function of β and converges in probability for all β
 697 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 β^{M_2} , *i.e.*

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

698 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 β , the derived sample estimator can be written as follows :

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

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 β 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*,

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

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 :

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

701 Let end the proof by replacing the β_0 , Q_1 and β_1 by their estimators. Each of them converges in probabilities;
702 $\hat{\beta}_0$ and $\hat{\beta}_1$ thanks to the asymptotics MLE proprieties and \hat{Q}_1 using the strong law of large number. The proof can
703 be generalized exactly in the same way under (X-A1) and (Z-A1) for $p > 1$. \square

C.2.2 Proof of equation 31

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

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

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 $j = 1, \dots, p$. Therefore, the solutions of the precedent equations are:

$$\beta_0^{M_2} = \beta_0, \beta_j^{M_2} = \beta_j, \beta_p^{M_2} = Q_p \beta_p, \quad j = 1, \dots, n. \quad (62)$$

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

707 C.2.3 Proof of equation 32

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

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

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

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

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

$$\beta_0^{M_2} = \beta_0, \beta_j^{M_2} = Q_j \beta_j, \quad j = 1, \dots, n. \quad (65)$$

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

710 **C.3 Proof for the GLM Log-Poisson**

711 **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

$$\begin{aligned}
 \mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}; \beta))) &\propto Q_p \mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))) \\
 &+ (1 - Q_p) \int_{\mathbb{R}} \int_{\mathbb{R}^{p+1}} -v e^{\beta^{*p} \mathbf{x}_{(sp)}^{real} + \beta_p z} + n(\beta^{*p} \mathbf{x}_{(sp)}^{real} + \beta_p z) dF_{X_1^{real}, \dots, X_{p-1}^{real}, V, Y}(\mathbf{x}_{(sp)}^{real}, v, y) dF_{Z_p}(z) \\
 &\propto Q_p \mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))) \\
 &+ (1 - Q_p) \int_{\mathbb{R}} \int_{\mathbb{R}^{p+1}} +v e^{\beta^{*p} \mathbf{x}_{(sp)}^{real}} - v e^{\beta^{*p} \mathbf{x}_{(sp)}^{real} + \beta_p z} dF_{X_1^{real}, \dots, X_{p-1}^{real}, V, Y}(\mathbf{x}_{(sp)}^{real}, v, y) dF_{Z_p}(z) \\
 &+ (1 - Q_p) \int_{\mathbb{R}} \int_{\mathbb{R}^{p+1}} -v e^{\beta^{*p} \mathbf{x}_{(sp)}^{real}} + y(\beta^{*p} \mathbf{x}_{(sp)}^{real} + \beta_p z) dF_{X_1^{real}, \dots, X_{p-1}^{real}, V, Y}(\mathbf{x}_{(sp)}^{real}, v, y) dF_{Z_p}(z) \\
 &\propto Q_p \mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))) + (1 - Q_p) \mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}_{(sp)}^{real}; \beta^{*p}))) \\
 &+ (1 - Q_p) \int_{\mathbb{R}^{p+1}} v e^{\beta^{*p} \mathbf{x}_{(sp)}^{real}} dF_{X_1^{real}, \dots, X_{p-1}^{real}, V, Y}(\mathbf{x}_{(sp)}^{real}, v, y) (1 - M_{X_p}(\beta_p)).
 \end{aligned} \tag{66}$$

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

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

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

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

With Hoeffding inequality, another bound can be deduced without needing a bounded variable⁹ :

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

712 These inequalities lead to the equation 36. □

713 **C.3.2 Proof of the Lemma 2**

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

$$\begin{aligned}
 \frac{\delta}{\delta \beta_j} \mathbb{E}(\log(f_Y(\mathbf{Y}|\mathbf{X}; \beta))) &= Q_p \frac{\delta}{\delta \beta_j} \mathbb{E}(\log(f_Y(\mathbf{N}|\mathbf{X}^{real}; \beta))) \\
 &+ (1 - Q_p) \int_{\mathbb{R}} \int_{\mathbb{R}^p} -v \mathbf{x}_j^{real} e^{\beta^{*p} \mathbf{x}_{(sp)}^{real} + \beta_p z} + n \mathbf{x}_j^{real} dF_{X_1^{real}, \dots, X_{p-1}^{real}, N}(\mathbf{x}_{(sp)}^{real}, n) dF_{Z_p}(z)
 \end{aligned} \tag{68}$$

can be separated in two part. Remind that $dF_{Z_p}(z) = dF_{X_p^{real}}(x)$ because Z_p and X_p^{real} have the same distribution. Under the assumption (X-A3), $dF_{X_1^{real}, \dots, X_p^{real}}(\mathbf{x}^{real}) = dF_{X_1^{real}, \dots, X_{p-1}^{real}}(\mathbf{x}_{(sp)}^{real}) dF_{X_p^{real}}(x_p^{real})$ for $j = 1, \dots, p - 1$. By replacing these

⁹Recall 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 :

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

714 The derivative according to β_p is calculated thanks to equation 67 without difficulty.

715 This end the proof for equation 37. □

716 C.3.3 Proof of the theorem A.2

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

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

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

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

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

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

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

718 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 $b_p < 0$, $-M'_{X_p}(b_p) < 0$.

719 These inequalities lead to if $b_p > \beta_p$, $d_p^{M_2}(b) < 0$ and if $b_p < 0$, $d_p^{M_2}(b) > 0$.

720 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 $b_{sp} = \beta_{sp}$ and $b_p \in [0, \beta_p^{M_1}]$ such as :

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

721 Because the solution of M_2 log-likelihood maximization is unique, the previous solution is the global maximum β^{M_2} .

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

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

723 □

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

Proof. Denote $\beta_* = (\beta_1, \dots, \beta_p)$. 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:

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

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

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

does not permit to find a bound on β^{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} ,

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

□

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:

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

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

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

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

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

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

□

730 **C.5 Convexity: Propositions 1**

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

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

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

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

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

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

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