Integrating data quality into GLM for insurance pricing

Pierre Chatelain

Integrating data quality into GLM for insurance pricing

P. Chatelain, Univ Lyon, UCBL, ISFA LSAF EA2429, F-69007, Lyon, France,
pierre.chatelain.act@gmail.com

ARTICLE HISTORY
Compiled June 7, 2021

ABSTRACT
The different dimensions of the data quality impact the feature selection and the regression in different ways. Actuaries as others modellers need to deal with this notion of quality. Looking through GLMs, we show how to find the real impact of variables with heterogeneous quality using an individualized quality index. Under a simple assumption that inconsistent data have the same distribution as the real one, we propose a method to estimate the covariate true impact on the predictor and provide a method to predict the outcome depending on quality indexes values under several assumptions. Different operational remarks on the creation and the use of quality index for actuarial uses are done.

KEYWORDS
Credibility, Quality index, OLS Regression, General nonlinear regression

AMS CLASSIFICATION
62J05, 62J10, 62J05

1. Introduction

Actuarial pricing was traditionally limited by the number of variables used and their complexity. Indeed, the number of variables stems from the underwriter answers who has a limited amount of time to answer all questions and also a limited or imprecised knowledge. To offset this problem, more and more insurance companies use external data to improve their model. However, for a same individual, its known features gathered from external data may not be all as precise. Moreover, the reliability of information gathered from different external databases varies within a same feature and also depending on individual. Indeed, often gathering processes aggregate data sets of various quality from various sources. As is reasonably logical to expect, modelling should depend on the quality of observations. Then, how can an individualized quality index be used for predicting ?

Because the data quality subject has myriads of applications and issues, literature is stemming from research on evaluating data quality. A consensus has been reached on the need of a multiple dimension analysis for data quality evaluation ([19]). For instance, completeness is a research field where numerous methods were developed to deal with missing values ([21], [12]). These methods are globally based on assumptions such as MCAR (Missing Completely At Random), MNAR (Missing Not At Random), or MAR (Missing At Random). On mismeasurement side, the book of Efromovich [6] gives some insight for an analysis in univariate solution and with biased predictors or response or in a more multivariate case ([14]). In the present paper, the credibility dimension will be studied further. The uncertainty of observations

is quantified and called quality index and supposed to be perfectly measured. It refers to the uncertainty of the covariates, where the observed covariate values are generated by a latent variable model based on the quality. On the uncertainty of covariates, some works exists on the mismeasurement side using trees algorithm ([20], [18]).

Actuaries, responsible for the data quality within the insurance company (articles 219, 237, 244, 245, 247 from Solvency II Commission Delegated Regulation (EU) 2015/35), must assess and justify the data quality even if it is 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, ACPR [1] evaluated that 10% of datas are coming from external parties. Campbell [3] relates several actuarial examples showing that data quality does not have a negligible impact.

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 from the North Americain side ([7]) or from the UK side [3] for instance. To the best of our knowledge, advices to take into account data quality are still very qualitative ([8]) for actuaries; basic recommendations such as deleting, imputing or correcting the problem.

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

1.0.0.1. Contributions.

This paper present two main contributions. First, we show how to take into account quality indexes in a GLM regression, how to predict with a quality index and some estimators proprieties are given. Next, several operational and practical remarks are given to help the creation and the use of quality indexes.

1.0.0.2. Outline of the paper.

The paper is built as follows: in the section 2, we introduce the general framework, the notation and precise how uncertainty is integrated in the covariate generating process. In the section 3 gives the main algorithm and theoretical results. Hereafter, a simulation study illustrates the results in the section 4. Next, section 5 brings close the different assumptions to actuarial uses. Finally, section 6 discusses the creation of quality indexes, the case of imperfect data quality indexes and links this work with the missing value theory.

2. Data problems and imputation

2.1. Notations

This framework is the same as Chatelain and Milhaud [4] one. We want to take advantage of the information provided by an *individualized* quality index related to the confidence we can have about the $i-th$ observation of the $j-th$ covariate, further denoted Q_{ij}.

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

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

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

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

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

We denote for the rest of the paper $(j = 1, ..., p)$, $\bar{Q}_j = \frac{1}{n}\sum_{i=1}^{n} Q_{ij}$ and assume $\bar{Q}_j \neq 0$. This assumption is not restrictive, especially for real-life applications where such covariates would simply be removed from the data. However, it does not mean that an individual having all quality indexes null does not exist.

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

2.2. *Illustrative example*

Let consider a simple example, with a set of observations $(Y_1, ..., Y_n)$ to study and the corresponding explanatory variables $(\mathbf{X}_1, ..., \mathbf{X}_4)$ where each observation is i.i.d associated.Here, only the last variable $\mathbf{X}_4$ has an individualized quality index $(Q_1, ..., Q_n)$ where $Q_i \in (0,1)$. In our case, the quality index is evaluated as values between 0 and 1 as shown in example 1. If the index could be use as a weigh in an univariate regression, the use of quality as a weigh in the regression cannot be done in a multivariate regression. Moreover, the use as a weigh may bias the regression. Example 1 displays another issue : if an actuary fits a model with medium quality observations, how should he adapt its prediction for observations where the covariates value is perfectly known or unknown ?

Table 1.: Dummy example. Here, X_1 could refer to occupant age in Year, X_2 is the covariate informing on if the insured person is a tenant or not, X_3 to the number of rooms and X_4 house value in € per m^2. Arbitrary, Y could be the annual claims amount.

Exposure	X_1	X_2	X_3	X_4	Q_4	Y
0.6	45	True	2	454	0.8	350 €
1	30	True	3	1000	0.6	0 €
1	43	True	2	2500	0	2450 €
0.2	61	False	6	245	0.7	0 €

(a) Example of a training dataset. Each $x_{i;4}$ observation has a quality index $Q_{i;4}$ associated which is between 0 and 1 - 1 being an observation of perfect quality and 0 the worst one.

X_1	X_2	X_3	X_4	Q_4	Y
35	True	3	723	1	? €
53	True	1	613	0.5	? €

(b) Testing dataset. From a training data set with an imperfect variable, how can we predict the future claim knowing perfectly a value or in a more general knowing imperfectly a value ?

2.3. *Deleting and imputation*

Let consider the strategy to impute new values on outliers or low quality observations[1]. Defining outliers in the multivariate case when the others covariates are not good quality is difficult. Straightforwardly in our framework, most of the outliers are also incorrect observations. In the univariate case, the uncertainty can influence the definition of outliers for a regression as shown in figure 1. Indeed, the outliers' detection is bias due to the data set's quality. In that situation, some perfectly observed observations may be defined as outliers. For instance, in figure 1, the top-right point corresponds to a real value however using the dataset with a lower quality, the point would be more considered as an outlier than using the real dataset when $\mathbf{X}^{real}$. This illustrates that data quality influences outliers detection.

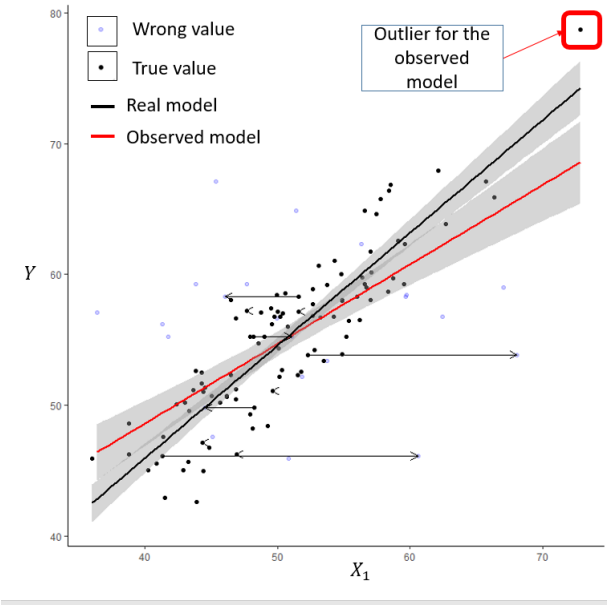


Figure 1.: An univariate example where black points are the real observations and grey points are the wrong observations. The arrows show translations of some data points between the real value and their observed one. This graph is based on simulated data with $X \sim \Gamma(1, 2)$ and $Y = 10 + 1x$, Q follows an uniform distribution between 0.5 and 1. The quality here is defined in the notation subsection 2.1. The thick line is the linear regression estimated on the true observations and the other on the data set of lower quality.

2.3.0.1. Naive solution. Given a data set and its joint quality index, a naive workaround of deleting could be done. As before, an easy one is to choose a threshold on the quality indexes and delete individuals having one of their quality indexes below. This solution can hardly be done with some low quality data or for highly dimensional datasets. Indeed, this latter issue was exemplified by Zhu and al. 2019 [22]. With an independent probability of a value missing equals to 0.05 and 300 covariates, this deleting approach would suppress 0.95% of the data set.

For our framework, suppose assumptions similar to Zhu et al. 2019 [22], i.e. in the case complete independence of quality and observations[2]. Assume all the quality indexes independently distributed as $Uniform(0.4, 0.8)$. Not only the low quality of the data implies a small threshold,

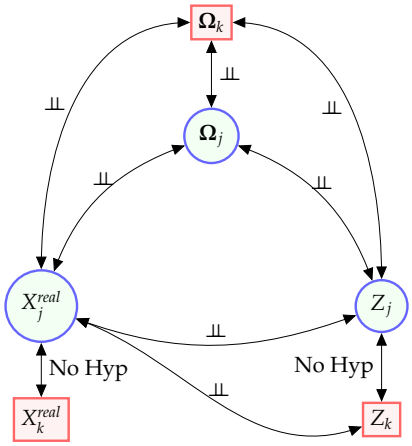
[1]Outliers detection or highly influential observations in a multivariate case was well studied in the general case for instance (Hadi, 1991 [9]) as in linear regression context (Cook 1977 [5]).

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

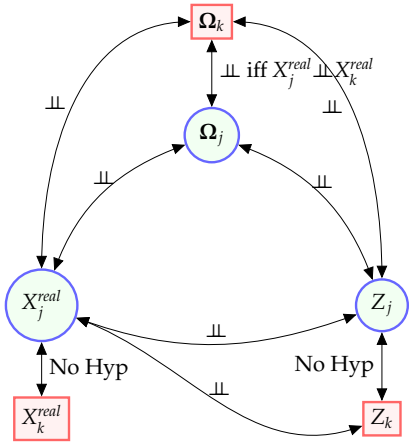
but the different observations would highly range around the mean value. For a threshold of 0.5 and 10 variables defined as before, only 6 % rows would have all its covariates above the threshold in average. Besides, errors can be correlated spatially and this filtering process may bias the portfolio risks. For open data used in household insurance, this is in particularly true for urban area zones : covariates have often lower quality in rural areas. Thus, filtering strategies are not optimal. To the best of our knowledge, integrating quality indexes in multivariate regression was never investigated for GLM.

2.4. Frameworks under study

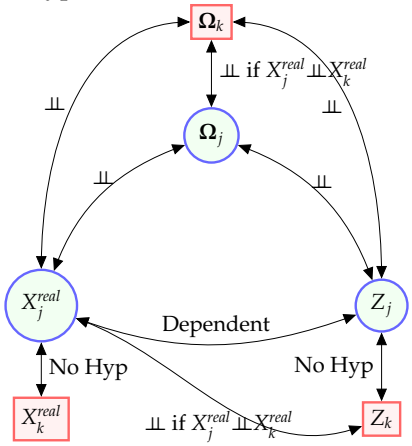
Several assumptions are looked through. They are linked with Rubin's nomenclature (Rubin, 1976 [15]), yet contested (Seaman, 2013 [17]). From the equation (1), different cases can be investigated depending on the correlation structure of ($\mathbf{X}^{real}$, $\mathbf{Z}$, $\mathbf{\Omega}$). We will consider the four following situations resume in the Figures 2a, 2b, 2c, 2d.



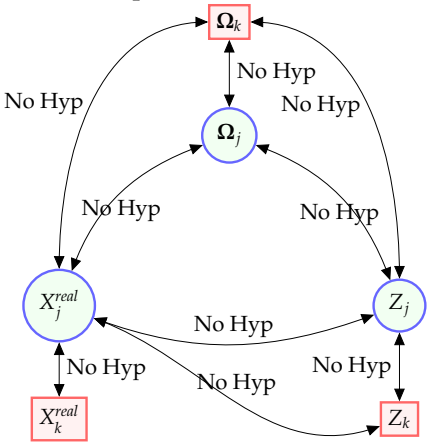
(a) Case (C1) - Total uncertainty ($j \neq k$). (No Hyp) means No hypothesis.



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



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



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

Figure 2.: Cases studied.

These assumptions can be linked with the missing value theory. For instance, the MCAR assumption (Rubins 1976 [15], Heitjan 1996 [10]) can be seen as a particular case of the statement (C1) when the quality indexes are equal either to 0 or to 1. The multivariate independency between the variables suggests that the errors are independent. In other words, each observation of each variable is gathered from different and unrelated sources or with unrelated errors.

In the same way, MAR assumption is a particular case of (C2) and (C3). Indeed, it corresponds to dependence between quality indexes/missing observation. In the case (C3), the wrong values $\mathbf{Z}_j$ are correlated to the real values $\mathbf{X}_j^{real}$. A particular case is when $(\mathbf{Z}_j - \mathbf{X}_j^{real})$ follows a centred distribution and is related to mismeasurement theory.

The last case (C4) is closely linked to MNAR setting, where no independency exists between each variable. The frequency of errors Ω and the wrong values $\mathbf{Z}$ can depend on the real values $\mathbf{X}^{real}$; the errors are informative which highly complexify the analysis. The different cases are discussed in section 5.

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

In this paper, only the case (C1) and (C2) will be studied through GLM and linear regression. By default, the result are under (C1) or will be mentioned otherwise. We also suppose that the information brought to the predictor from $\mathbf{Z}$ is not distinct from $\mathbf{X}^{real}$; $\mathbf{Z}$ is informative only through it correlation with $\mathbf{X}^{real}$.

3. Estimation Process

3.1. Reducing the error by mitigating on quality pattern

In this work, $\mathbf{X}$ is governed by the underlying process generating the covariates, as in the equation (1). In OLS regression, the solution $\hat{\beta}$ minimizes the Residual Squared Error (RSE) calculated on the dataset $\mathbf{X}$. In GLM regression, 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 define a quality pattern. Denote $P(\mathbf{Q})$ the set of all quality patterns present in the data. By taking it into account, the cost metric can be improved since

$$(1/n)Dev(\hat{\beta}|\mathbf{X}, Y) \geq (1/n) \sum_{K \in P(\mathbf{Q})} \sum_{i \backslash \mathbf{Q}_i = K} Dev(\hat{\beta}^K|\mathbf{X}_i, Y_i), \tag{2}$$

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

3.2. Prediction using quality index

We study in the sequel GLM, given by

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

and the likelihood associated $\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})$ where the optimization is made using the real data set $\mathbf{X}^{real}$. In most cases, the previous model is unknown in our framework. Hereafter, this model is called **"real" model**.

Denote the following naming :

- 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 $Argmax_\beta \mathcal{L}^{M2}(\beta; \mathbf{Y}|\mathbf{X})$. When $\mathcal{L}^{M2}(\beta; \mathbf{Y}|\mathbf{X})$ is estimated using $\mathbf{X}^{real}$, denote it $\mathcal{L}^{M2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q})$, we write $\hat{\beta}^{M_2|\mathbf{X}^{real},\mathbf{Q}}$ the solution of $Argmax_\beta \mathcal{L}^{M2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q})$.
- 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 $\mathcal{Q}$, obtained from Algorithm 3 the models depend on each quality pattern:

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

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

For all the proofs, we will suppose all the input centred. Indeed, in linear regression and in GLM, a simple covariates translation only impacts the intercept's coefficient β_0.

3.3. Algorithm 3 for linear regression and GLM

For linear regression (see [4]), algorithm displayed in figure 3 associated with the Model M_3 first assesses the Naive model M_2 from $\mathbf{X}$. Using $\hat{\beta}^{M_2}$ and the correlation matrix empirical $\hat{\Sigma}$, an estimator of the "perfect quality" correlation matrix Σ^{real} (see A.1) and then coefficients of real model - $\hat{\beta}^{M_1}$ - are evaluated thanks to the quality index $\mathcal{Q}$. Finally, the algorithm find $\hat{\beta}^K$ which minimizes the Residual Squared Errors for each pattern of quality K.

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

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

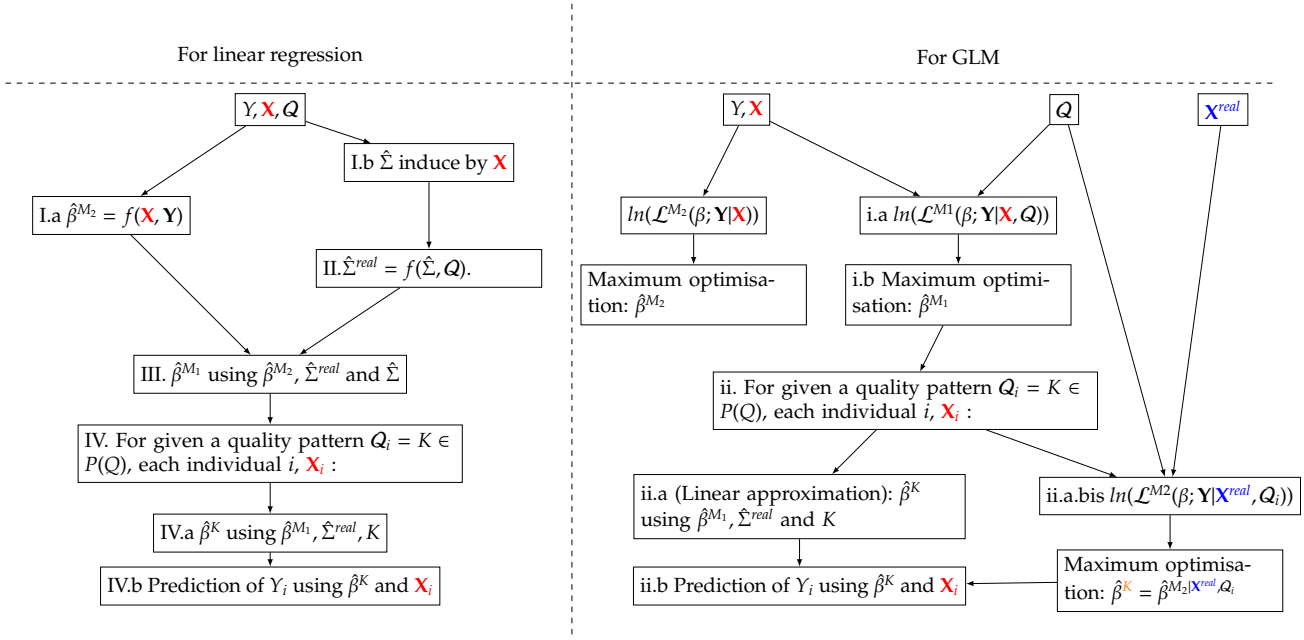


Figure 3.: Process to take into account the quality index for linear regression and an approximation for GLM log-Poisson. In some way, this process adjusts the coefficient to each quality pattern.

3.4. *Two assumptions under study*

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

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

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

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

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

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

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

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

3.5. The likelihood of the model with quality index

For actuarial pricing, most of the model used are not Gaussian but often Poisson distribution or Gamma one. In the GLM case, the set of coefficient $\beta = (\beta_0, ..., \beta_p)^T$ is found by maximizing the likelihood or log-likelihood (ML-Maximum likelihood);

$$\underset{\beta \in \mathbb{R}^p}{argmax}\ \mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}) = \underset{\beta \in \mathbb{R}^p}{argmax} \sum_{i=1}^{n} ln(f_Y(y_i|\mathbf{X}_i; \beta)), \tag{3}$$

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

Because the observations are independent and identically distributed, the previous log likelihood is the sample analogue of $\mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$.We assume mild regularity conditions (see A.3) for the proper converges of our models. In our framework, these regularity conditions leads to existence of the moment generating function for each imperfectly observed covariates when needed. More detail on the theoretical part can be found in the appendixes.

Model GLM	Case		Hyp $\mathbf{X}^{real}$	Hyp $\mathbf{Z}$	Estimator convergence	Log-likelihood (M_1)
Log-Gaussian	(C1)-(C2)	$p = 1$	No Hyp	No Hyp	$\hat{\beta}_0^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} \beta_0$ and $\frac{\hat{\beta}_1^{M2}}{Q_1} \xrightarrow{\mathbb{P}} \beta_1$	-
	(C1)	$p > 1$	X-A1	Z-A1	$\hat{\beta}_0^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} \beta_0$, $\frac{\hat{\beta}_j^{M2}}{Q_j} \xrightarrow{\mathbb{P}} \beta_p$, $j = 1, ..., p$.	-
	(C1)	$p > 1$	X-A3	Z-A1	$\hat{\beta}_0^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} \beta_0$, $\frac{\hat{\beta}_j^{M2}}{Q_j} \xrightarrow{\mathbb{P}} \beta_p$, $j = 1, ..., p$.	-
	(C2) $\Omega_j = \Omega_k$	$p > 1$	No Hyp	Z-A2	$\hat{\beta}_0^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} \beta_0$, $\frac{\hat{\beta}_j^{M2}}{Q_j} \xrightarrow{\mathbb{P}} \beta_p$, $j = 1, ..., p$.	-
Log-Poisson	(C1)-(C2)	$p = 1$	No Hyp	No Hyp	$\hat{\beta}_0^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} \beta_0$ and $\hat{\beta}_1^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} [0; \beta_1]$	Yes
	(C1)	$p > 1$	X-A3	Z-A1	$\hat{\beta}_j^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} \beta_j$, $j = 0, ..., p-1$ $\hat{\beta}_p^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} [0; \beta_p]$.	Yes
	(C2) $\Omega_j = \Omega_k$	$p > 1$	No Hyp	Z-A2	No bounded convergence	Yes
Log-Gamma	(C1)-(C2)	$p = 1$	No Hyp	No Hyp	$\hat{\beta}_0^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} \beta_0$ and $\hat{\beta}_1^{\mathcal{M}_2} \xrightarrow{\mathbb{P}} [0; \beta_1]$	Yes
	(C1)	$p > 1$	X-A3	Z-A1	No bounded convergence	Yes
	(C2) $\Omega_j = \Omega_k$	$p > 1$	No Hyp	Z-A2	No bounded convergence	Yes
Inv-Gamma	(C1)	$p > 1$	X-A3	Z-A1	No bounded convergence	No
Probit	(C1)	$p > 1$	X-A3	Z-A1	No bounded convergence	No

Previous table shows different results for different GLM. For the most commonly used in pricing (Log-Gaussian, Log-Poisson and Log-Gamma GLM) some interesting results can be found whith this framework thanks to the additive or multiplicative structure. For Probit or Inv-Gamma GLM, no results can be found without approximation. Log-Gaussian GLM leads to explicit relation between β and β^{M2}. Therefore, the M_1 log-likelihood is not needed to be calculated. Because Log-Gaussian GLM and linear regression are equivalent, we logically

found the same results. It is important to notice that β_j^{M2} only depends on Q_j and β_k and Q_j for all X_k correlated to X_j.

In the case (C1), Log-Poisson and Log-Gamma GLM's multiplicative structure permit determining $ln(\mathcal{L}^{M1}(\hat{\beta};\mathbf{Y}|\mathbf{X},\mathbf{Q}))$. However, in multivariate case, under (X-A3) and (Z-A1), $\beta_p^{M_2}$ depends on Q_p and $M_{X_p}(t)$ and β_j for $j = 1,...,p-1$. The main difference is that β_j^{M2} depends on the distribution of X_p. Under (X-A3) and (Z-A1), if, for Log-Poisson model, $\beta_j^{M2} = \beta_j$ and $\beta_p^{M_2}$ has a bounded convergence in probability, for Log-Gamma model β_p^{M2} is not bounded and moreover the other coefficients, β_j^{M2}, are also impacted by Q_p.

In the case (C2) with fully correlated quality variable under (Z-A2), i.e. $\Omega_j = \Omega_k$ for all j and k, Log-Gaussian coefficients $\beta_j^{M_2}$ have a simple affine linear-ship with $\beta_j^{M_1}$ for $j = 1,...,p,$. Regrettably, no proprieties on the estimator can be state for Log-Gamma and Log-Poisson GLM.

3.6. Deduce β^{M3}

As already mentioned in Section 3.2, the vector β^K exactly matches β^{M_1} when all individualized quality indexes equal 1, i.e. when $K = J_{1,p+1}$. In full generality, when $K = \mathbf{Q}_i$ is made of terms $Q_{ij} \neq 1$, the coefficients $\hat{\beta}^K$ need to be calculated, $\hat{\beta}^{K=\mathbf{Q}_i}$ is an estimator of 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 coefficient $\hat{\beta}^K$ is the one minimizing the mean $Dev(\hat{\beta}^K|\mathbf{X},Y)$ for a given pattern of quality K as we wanted (see the equation 2).

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

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

$$\begin{aligned}&\bar{Q}_1 \sum_{i=1}^{n} ln(f_Y(y_i|X_{i1}^{real};\beta))\\ &+ (1-\bar{Q}_1) \times \sum_{i=1}^{n} \frac{1}{n} \sum_{h=1}^{n} ln(f_Y(y_i|X_{h1}^{real};\beta)).\end{aligned} \tag{4}$$

This estimator which converges almost surely, is denoted $ln(\mathcal{L}^{M2}(\beta;\mathbf{N}|X^{real},\mathbf{Q}))$. The associated maximum likelihood estimator $\hat{\beta}^{M2|X^{real},\mathbf{Q}}$ converges in probabilities into β, i.e.

$$\hat{\beta}^{M2|X^{real},\mathbf{Q}} \xrightarrow{P.} \beta.$$

Proof. See A.11. □

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

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

$$\begin{aligned}&\bar{Q}_p \sum_{i=1}^{n} ln(f_Y(y_i|X^{real}_{i;(*p)}, \mathbf{X}_{i;p} = \mathbf{X}^{real}_{i;p};\hat{\beta})\\ &+ (1-\bar{Q}_p)\sum_{i=1}^{n}\frac{1}{n}\sum_{h=1}^{n} log(f_Y(y_i|X^{real}_{i;(*p)}, \mathbf{X}_{i;p} = X^{real}_{h;p};\hat{\beta})).\end{aligned} \tag{5}$$

is consistent. The associated maximum likelihood estimator $\hat{\beta}^{M2|X^{real},\mathbf{\Omega}}$ *converges in probabilities into* β*, i.e.*

$$\hat{\beta}^{M2|X^{real},\mathbf{\Omega}} \xrightarrow{\mathbb{P}} \beta.$$

Remark 3.2. *In fact, for any correlation structure in between* $\mathbf{X}^{real}$*,* $\mathbf{\Omega}$*,* $\mathbf{Z}^{real}$*, we could find estimate of the expected likelihood of* M_2 *easily. 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 our case. Nonetheless, if $\mathbf{X}$ has the same the correlation structure than $\mathbf{X}^{real}$[3], a solution would be to simulate Y^{new} from $\mathbf{X}$ using $\hat{\beta}$ and therefore calculate the previous estimator.

3.7. *Deduce* β^{M1} *for log-Poisson GLM*

In this part we 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. We denote $Y = N$ and V the exposure to have more traditional notations for count distributions.

Remind that only X_p has a heterogeneous quality. Using the equation A9, an estimator of $ln(\mathcal{L}(\hat{\beta};\mathbf{N}|\mathbf{X}^{real}))$ can be found as follows :

$$\begin{aligned}ln(\mathcal{L}^{M_1}(\hat{\beta};\mathbf{N}|\mathbf{X},\mathbf{Q})) = \frac{1}{\bar{Q}_p}\Big[&ln(\mathcal{L}^{M2}(\hat{\beta};\mathbf{N}|\mathbf{X})\\ &- (1-\bar{Q}_p)\times ln(\mathcal{L}(\hat{\beta}^{*p};\mathbf{N}|\mathbf{X}^{real}_{(*p)}))\\ &- (1-\bar{Q}_p)\times Ve^{\hat{\beta}^{*p}\mathbf{X}^{real}_{(*p)}}(1-M_{X_p}(\hat{\beta}_p))\Big].\end{aligned} \tag{6}$$

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

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

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

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

$$
\begin{aligned}
ln(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{N}|\mathbf{X}, \mathbf{Q})) = \sum_{K_p \in P(Q_p), K_p \neq 0} \frac{1}{K_p} \Big[& ln(\mathcal{L}^{M2}(\hat{\beta}; \mathbf{N}|\mathbf{X}_{Q_p=K_p}) \\
& - (1 - K_p) \times ln(\mathcal{L}(\hat{\beta}^{*p}; \mathbf{N}|\mathbf{X}^{real}_{(*p);Q_p=K_p})) \\
& - (1 - K_p) \times Ve^{\hat{\beta}^{*p}\mathbf{X}^{real}_{(*p);Qp=Kp}}(1 - M_{X_p}(\hat{\beta}_p)) \Big].
\end{aligned} \tag{7}
$$

where $\mathbf{X}_{Q=K_p}$ represents the dataset where only the individual i such as $Q_{i;p} = K_p$ are kept. The second estimator $ln(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{N}|\mathbf{X}, \mathbf{Q}))$ from the equation 7 is often more precise by construction than the equation, 6 however individual having null quality index are not taken into account. Therefore, in the following part, we will use the second estimator. These two estimators converge in probabilities to $ln(\mathcal{L}(\hat{\beta}; \mathbf{N}|\mathbf{X}^{real}, \mathbf{Q}))$. In the same way, the solution of the maximum likelihood converges in probability.

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

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

4. Simulation study - M1 estimator

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

The study is performed using R ([13]) statistical software.

4.1. Find $\hat{\beta}^{M_1}$ model

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

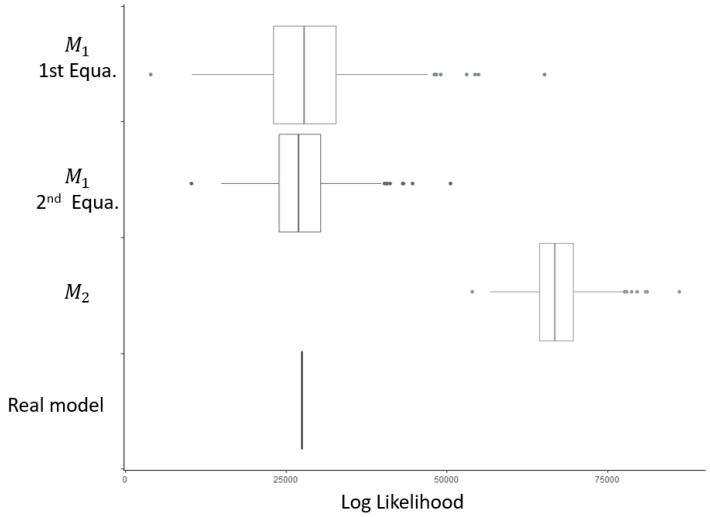


Figure 4.: Estimation of the M_1 log-likelihood for log-Poisson GLM using equation 6 for a given $\mathbf{X}$ and $\mathbf{Q}$. The moment function is estimated using its empirical estimator. The true function leads to the same graph but with a smaller variance. 2000 simulations are done for a given $\mathbf{X}^{real}$ and $\mathbf{Q}$.

Using the precedent result, M_1 likelihood can be estimated as shown figure 4. The use of imperfectly observed dataset implies a wider variance of the estimator M_1 than the real model one. Here, the first estimator has wider variance than the second estimator. As shown by equation A11, the coefficients of $\beta_1, \beta_2, \beta_3$ did not change due to the independence in between the variables - figure 5 - and the coefficient associated to X_4 is corrected - figure 6.

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. We suggest using the linear approximation, e.i. $\hat{\beta}_p^{K=Q_p} = Q_p \times \hat{\beta}_p^{M_1}$. Indeed, as shown on the figure 7, for small values of β_4 (≈ 0.07), the impact of the moment on the likelihood is lower than for higher value of $\beta_4 = 4$. Therefore, the coefficient could be estimated linearly only for $\mathbb{E}(Y)$ small, but would overestimate the coefficient for higher values.

5. Discussion

The following example comes straight up from a project on household pricing using the geolocalised address to add external data.

5.1. Example for each case

In this part, we will discuss the different cases from the scope of a pricing case using house geolocalisation. Here, the goal is to model the frequency or the claim of a household insurance

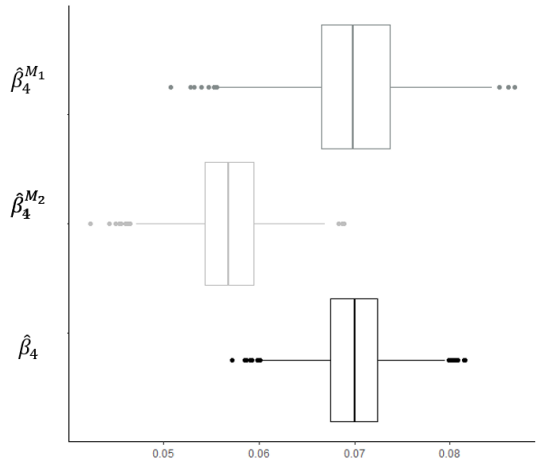


Figure 5.: The quality of the variable X_4 does not impact the estimation of β_1, β_2, β_3; here, highlighted by β_1 with X_i^{real} standard normal distribution and Y following a Poisson distribution. Other distributions of X_i^{real} have also been tested and leads to the same results. 2000 simulations are done for a given $\mathbf{Q}$.

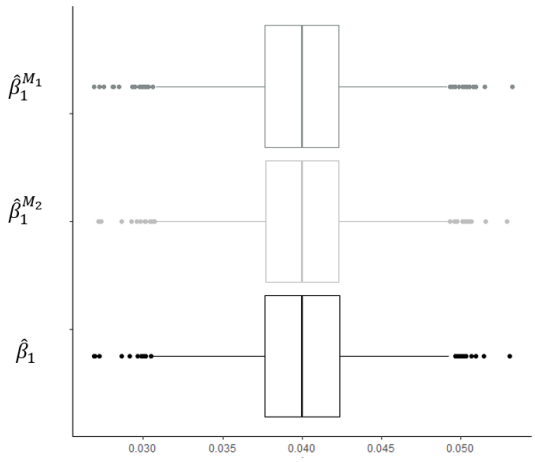


Figure 6.: $\beta_4^{M_2}$ is smaller than $\hat{\beta}$ because of the quality of the variable and $\beta_4^{M_1}$ is unbiais but have a wider variance than the real coefficients.

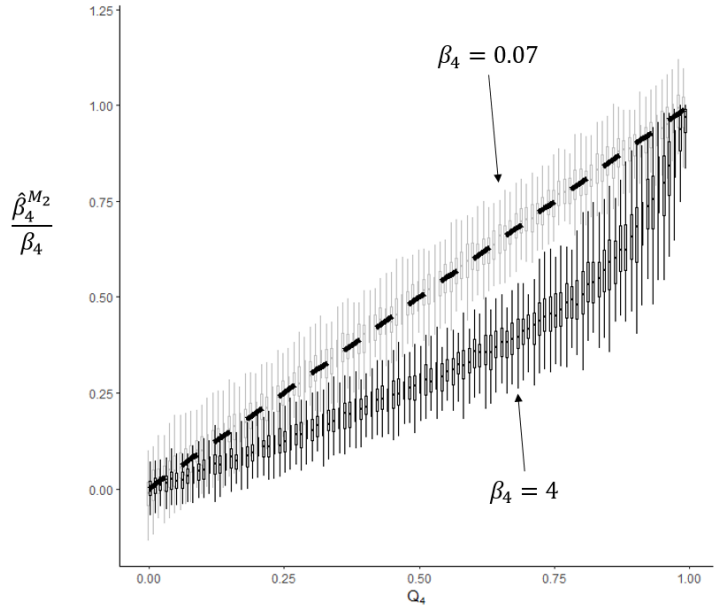


Figure 7.: 1000 simulation for each quality. As for linear regressions, a linear evolution through the quality can be seen for low coefficient however for higher values the relationship is not proportional to the quality.

using only the address and external data. To find the different covariates associate to characteristics of the individual, the first step is to link the address with its geocoding, then to link the geocoding to the right parcel or/then with the building. Finally, the goal is to evaluate the different characteristics thanks to external data or the picture analysis. The output to model is given by insurers departments. It corresponds to the frequency or claims cost and is supposed to be perfectly observed.

Here, the collected data's quality is mainly looked through the credibility dimension. If the geocoding is wrong, all the observations would be taken on another building. Finally, the consistency of the variable and the way it is collected change also the data quality. However, all the variables are not impacted in the same way.

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

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

5.1.0.2. Example of case C2. The living surface, the number of rooms and the footprint are globally one of the most segmenting features in household pricing. Different data-sets and methods are available in France to collect them such as DVF [4]. This database geolocates the parcel and contains different features such as the value of property values, the number of rooms, the surface of the parcel or the living surface among others. The database is updated after a property transfer since 2015. On the uncertainty dimension, errors are coming from the link between geocoding and the address or between the address and the building, each of these steps impact the data's quality depending on the feature. A wrong geocoding would imply that the observations are taken from another building. For all these variables, the case (C2) and the assumption (Z-A2) would be appropriate since they are collected from the same building.

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

5.1.0.4. Example of a case C4. All variables mentioned earlier can fit in this category due to spatial correlation. For instance, if in megalopolis the detection of the house size may be difficult due to the building's density, a systematic uncertainty could appear on this variable for urban

[4]This data-set comes from a certified public service relating to the property values declared during property transfers available in open data at `https://www.data.gouv.fr/fr/datasets/5c4ae55a634f4117716d5656/`

houses - globally smaller. Then $\mathbf{Z}$ would be correlated with the $\mathbf{X}^{real}$ automatically but also with Q. The same analysis could be done on high buildings for the number of floor for instance.

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

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

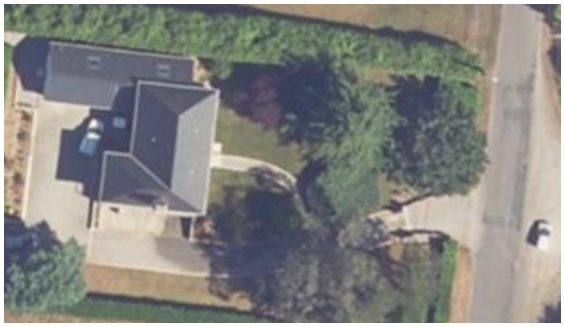


Figure 9.: The detection of a window on a roof is difficult because of the dark slate roof from the IGN cartography (47.179068, -1.814216) - © IGN 2018

5.2. *Actuarial justification of this framework assumptions*

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

5.2.0.2. 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. Moreover, adding some imperfect variables correlated to others also bias the coefficients of these perfectly observed variables.

5.2.0.3. 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 0.1 % (by example for water damage or fire damage coverage. The other benefit is that only one model is fitted.

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 completly 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 tariff. Logically, information from underwriting questionnaire are much better quality and are often suppose perfectly observed. So the tariff muss use β, adapted to perfectly observed variables, and not β^{M2}.

6. Operational applications

6.1. *Quality impact and attenuation*

The different results show that the "attenuation"[5] on $\hat{\beta}$ due to data quality can be explained. However, the quality impacts might not always decrease the coefficients as shown in Chatelain and Milhaud [4]. The quality of a variable impacts all coefficients related to other correlated variables, as well as the intercept. Figure 10 under (X-A2) and (Z-A1) in linear regression shows that even in the simple case the "attenuation" is not always true. With some correlation, the coefficient can be higher than the usual one (the true coefficient equal to 1 and is represented the line on Figures 10 and 11) and even might change sign.

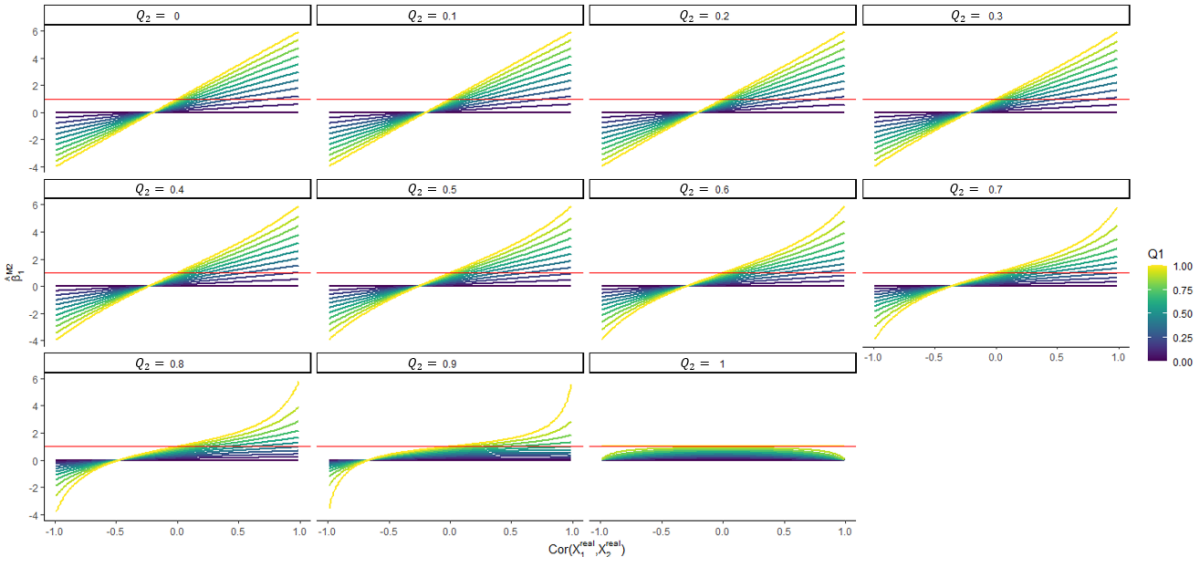


Figure 10.: Log-Gaussian GLM: Value of $\beta_1^{M_2}$ wrapped by Q_2 and grouped by Q_1. The coefficient β are all equal to 1 and the ratio of the standard deviation $\sqrt{Var(X_1^{real})/Var(X_2^{real})}$ equals to 6. The red straight line represents, β_1 which is equal to 1.

This is especially harmful to insurance pricing where covariates' choice must be justified by their impacts. Indeed, some coefficients may seem counter-intuitive due to quality impacts. Figures 10 and 11 provide an 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 with the correlation. Figure 10 shows that if $\rho < 0$, $\beta_1^{M_2}$ could be negative, even if $\beta_1 > 0$. Another point is that the coefficients could be considered as null even if the variable's quality is not low. For instance, for $Q_2 = 0.7$ and $\rho \approx -0.4$, $\beta_1^{M_2} \approx 0$ and $\beta_2^{M_2} \neq 0$. In this case, dropping the variable X_1 would not have any impact on $\beta_2^{M_2}$ even if the true coefficient is different from 0. Moreover, by finding the β^{M_1} - thanks to $\mathbf{X}$ and $\boldsymbol{Q}$, the modeler can find the "real" impact of a variable in models, thus justifying it.

[5] As called in the econometric literature.

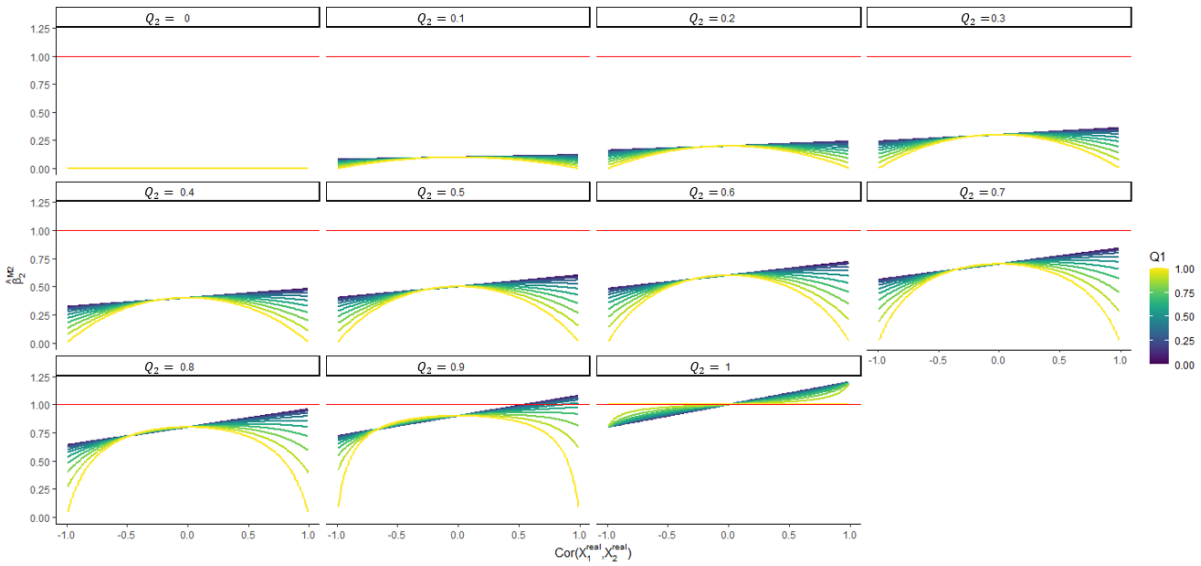


Figure 11.: Log-Gaussian GLM: Value of $\beta_2^{M_2}$ wrapped by Q_2 and grouped by Q_1. The coefficient β are all equal to 1 and the ratio of the standard deviation $\sqrt{Var(X_1^{real})/Var(X_2^{real})}$ equals to 6.

Remark 6.1. *Here, the discussion was done with the simplest hypothesis under the case (C1) and for Gaussian distribution where the variable quality does not impact others independent variables coefficients. For other distributions, the quality impacts would complicate the whole issue further.*

6.2. 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. 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 :

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

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 not centered and Q_1 only, i.e,

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

would be the best option only if $\rho = 0$. The interaction should be on both variables :

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

Obviously, with more covariates and quality indexes, it adds a lot more parameters to fit exactly $n \times 2^{h-n}$ 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. Such as the remark of the previous section 6.1 state, for other distributions the whole issue is much more complex. Therefore, in

such case limiting the correlation in between the variable should be the priority and then to limit the number of variable.

6.3. *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 the equivalent to the process explained in this paper. Denote the following model $E[Y|\mathbf{X} = \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ 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

$$y_i = \beta_0 + \beta_1 \mathbb{E}(X_1^{real}) + \beta_2 x_{i;2}^{real},$$

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

$$y_i = \beta_0^K + \beta_1^K z_1 + \beta_2^K x_{i;2}^{real},$$

where K is the pattern of the quality - here $K = (0, 1)$, $\hat{\beta}_j^K$ is the estimator found thanks to the process 3, $j \in \{0; 1; 2; 3\}$ and z_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 coefficient can be written as

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

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

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

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

6.3.0.1. Multivariate case. By extrapolating these results, it seems that for one missing observation the equivalent imputation should be the prediction of the linear regression of the other covariates. However, this remark does not take into account the issue with other covariates' quality. To go further than the case (C1), the credibility of other covariates may be also correlated with the fact that a value is missing. This remark is close to the analysis of Seaman 2014 [16] about how to impute with fully conditional specifications.

For Log-Poisson GLM, we showed that under (X-A3) and (Z-A1) the mean imputation is also equivalent, which is not always true for other assumptions. For instance, in Log-Gamma model

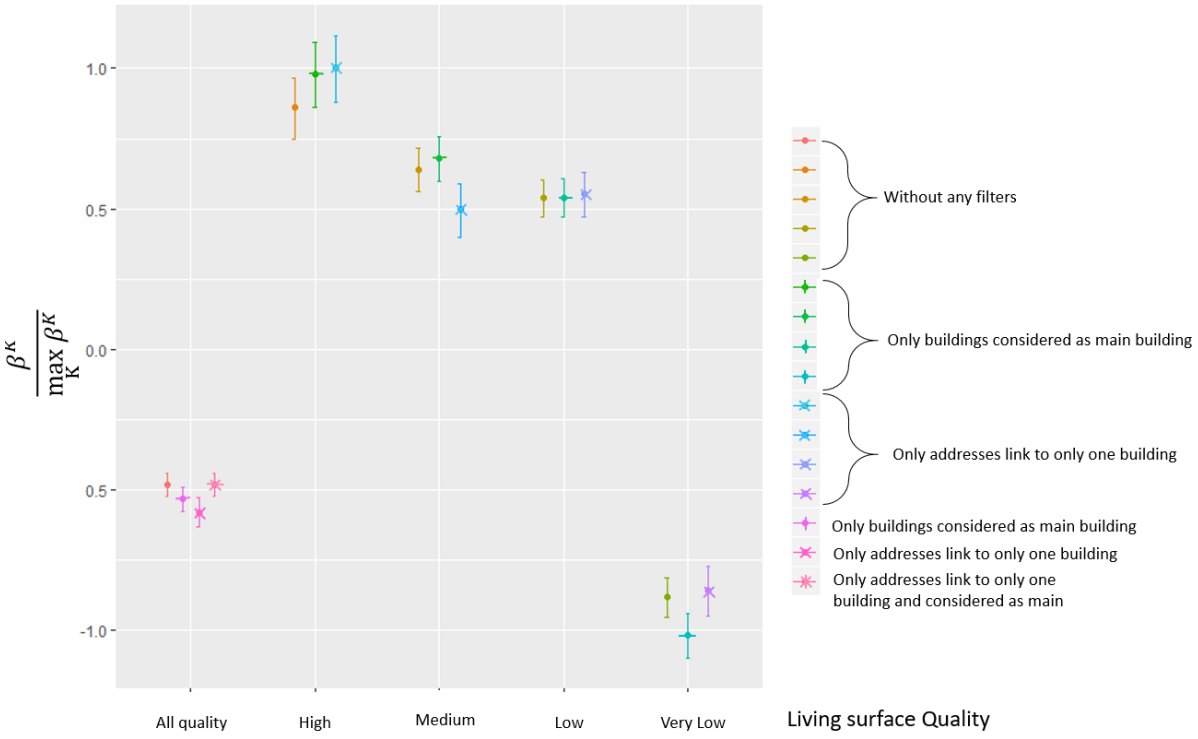


Figure 12.: Ratio of the living surface coefficient of Log-Poisson. Because they are too few "very high" quality observations, there were regrouped with "high" quality observations. Different filters based the building geolocalisation information are done on the data set to challenge the quality index.

the other coefficients are also impacted. To find the best solution into account, one should find β^K and modify all the other coefficients.

6.4. 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"). Is it 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. In this case, a modality may regroup different levels of quality. In other word, quality index is not perfectly determine. Fitting an univariate linear model with variables centred and with an interaction with the covariate $K(X_1)$ with M modalities and a :

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

If we assume the modality $K(X_1) = 1$ corresponds to perfect quality observations, the quality index value would be of the m modality is equal to $\hat{Q}_m = \beta_1^{K(X_1)=m} / \beta_1^{K(X_1)=1}$.

Figure 12 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. Then to complete the missing information, a Machine Learning method is 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 confidence of geocoding property sales database. On the other hand, Machine Learning values are associated with a maximum of "medium" and the confidence is degraded depending on the quality of the covariates and the score associated to each result. Two filters are considered on the addresses geolocalisation : 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 12 helps to evaluate the quality indexes values. Supposing $\beta^{High} = \beta$ perfect and using a linear approximation, "medium" quality value would be estimate by 0.6, "low" quality value by 0.5. However, the coefficient of very low quality values has an opposite sign. In fact, very low quality values are link to rural zone. Therefore, the case (C4) is the most appropriate and our evaluation method can not be used. In the same way, "low" quality values observations are also a bit more link to rural density than "medium" one or "high"[6]. In consequence, the associated value to medium quality 0.5 can be debated. Indeed, the "low" and "very low" quality are correlated with others characteristics impacting the risks. The coefficients calculated on the database are therefore highly impacted. In such case, using a threshold to set aside "very low" quality observations is recommended so that the dataset verify our assumptions. Finally, the different filters on geocoding show that leaner detail could be added within a value of the quality index.

In practice, a modality might regroup observation of different quality. Looking into the case of modalities regrouping two type observations with distinct quality index, denote $K(X_1) = m$ a modality of n observations which regroups n_α (n_κ) number observations with the quality Q_α (Q_κ respectively). The difference between model's coefficients deduce and the real model can be expressed as a barycenter of the sum of the group's quality under (X-A1):

$$\beta_1^{M2,\hat{Q}_m} - \beta_1 = \frac{n_\alpha}{n}(Q_\alpha - 1)\beta_1 + \frac{n_\kappa}{n}(Q_\kappa - 1)\beta_1. \tag{8}$$

Equation 8 can be easily extended to higher dimension. If groups of different quality are mixed together and are given the same value Q, the best one should be the pondered mean of each quality in a context of linear regression with the Assumption (X-A1). However, with the Assumption (X-A2), the aggregation of the quality influence the coefficients value of other correlated covariates.

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

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

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

Proof. See appendix B. □

[6]Here, the mismearsurment side is set aside.

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

$$\forall Q_\alpha, Q_\kappa \in [0,1], \quad \beta_k^{M_2}(\frac{n_\alpha}{n}Q_\alpha + \frac{n_\kappa}{n}Q_\kappa) \leq \frac{n_\alpha}{n}\beta_k^{M_2}(Q_\alpha) + \frac{n_\kappa}{n}\beta_k^{M_2}(Q_\kappa). \tag{9}$$

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

In a nutshell, because the evaluation of a quality index is never perfect, a practical recommendation with the use of quality indexes would be to prefer covariates with low correlation in between them.

7. Conclusion

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

To use data's quality index in multivariate correlated covariate, further research is ongoing to adapted decision trees to this use. Several issues remain generalizing for penalized likelihood optimization and quality index evaluation.

References

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

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

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

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

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

[6] Sam Efromovich. *Missing and modified data in nonparametric estimation: with R examples*. CRC Press, 2018.

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

[8] General Committee of the Actuarial Standards Board and Applies to All Practice Areas, GCASB. Data quality - revised edition. Technical Report 23, Actuarial Standard of Practice, May 2014.
[9] Ali S Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):761–771, 1992.
[10] Daniel F Heitjan and Srabashi Basu. Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3):207–213, 1996.
[11] EL Lehmann and George Casella. Unbiasedness. *Theory of Point Estimation*, pages 83–146, 1998.
[12] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.
[14] Marco S Reis and Pedro M Saraiva. Integration of data uncertainty in linear regression and process optimization. *AIChE journal*, 51(11):3007–3019, 2005.
[15] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
[16] Shaun Seaman and Ian White. Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, 43(16):3499–3515, 2014.
[17] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.
[18] Myriam Tami, Marianne Clausel, Emilie Devijver, Adrien Dulac, Eric Gaussier, Stefan Janaqi, and Meriam Chebre. Uncertain trees: Dealing with uncertain inputs in regression trees. *arXiv preprint arXiv:1810.11698*, 2018.
[19] Ion-George Todoran, Laurent Lecornu, Ali Khenchaf, and Jean-Marc Le Caillec. Toward the quality evaluation of complex information systems. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, volume 9091, page 90910N. International Society for Optics and Photonics, 2014.
[20] Asma Trabelsi, Zied Elouedi, and Eric Lefevre. Handling uncertain attribute values in decision tree classifier using the belief function theory. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 26–35. Springer, 2016.
[21] Stef Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.
[22] Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.

Appendix A. Theoretical framework

A.1. The covariance impacted by quality index

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

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

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

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

In case (C2), if X_j^{real} *and* X_k^{real} *are independent, both* (A3) *and* (A4) *hold true. Otherwise, if the joint quality are completely and positively dependant, we have :*

$$\textbf{\textit{Under (Z-A1)}}\ Cov_{jk} = Q_j \times Cov_{jk}^{real}, \tag{A3}$$

$$\textbf{\textit{Under (Z-A2)}}\ Cov_{jk} = Cov_{jk}^{real}. \tag{A4}$$

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

A.2. Regression model under consideration

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

$$\mu = E[Y|\mathbf{X} = \mathbf{x}] = g^{-1}(\mathbf{x}\beta). \tag{A5}$$

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

A.3. Univariate analysis in GLM

In this section we focus on the univariate case (p = 1). For β in $\mathbb{R}^2$, the model M_2 maximizes the following log-likelihood

$$\begin{aligned}\mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X}; \beta))) = \mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))|\Omega = 1) \times \mathbb{P}(\Omega = 1) \\ + \mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{Z}; \beta))|\Omega = 0) \times \mathbb{P}(\Omega = 0).\end{aligned} \tag{A6}$$

In the equation A6,the expected value of $\Omega = 0$, equals to Q_1, is known and when $\Omega = 0$, X_1^{real} is not observed and Z_1 is given.

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

$$\mathbb{E}(ln(f_Y(Y|\mathbf{X}^{real};\beta)))|\Omega = 1) = \mathbb{E}(ln(f_Y(Y|\mathbf{X}^{real};\beta))),$$
$$\mathbb{E}(ln(f_Y(Y|\mathbf{Z};\beta))|\Omega = 0) = \mathbb{E}(ln(f_Y(Y|\mathbf{Z};\beta))).$$

How could we estimate $\mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{Z}_1;\beta)))$ with $\mathbf{X}_1^{real}$? In the multivariate case, the value z_{i1} could be estimated using the other covariables. If X_1^{real} and Z_1 are correlated or dependent, a function g could exist such as $g(X_1^{real})$ is a good estimator of Z_1. Under the case (C1), none of these solutions can be applied. Indeed, the quality index Q_1, the real data-set X_1^{real} and the wrong values Z_1 are completely independent.

Beforehand, recall one main regularity condition which are mandatory for the MLE convergence of the exponential family based $\mathbf{X}^{real}$ (see section 6.2 of [11] for all the conditions needed):

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

Assumption A.2 and the other one for the exponential family lead to the *Bartlett identities* and both derivatives can be passed under the integral sign ([2]). The Bartlett identities are :

$$\mathbb{E}\Big(\frac{\delta}{\delta\beta}ln(f_Y(Y|X^{real};\beta))\Big) = 0,$$
$$Var\Big(\frac{\delta}{\delta\beta}ln(f_Y(Y|X^{real};\beta))\Big) = -\mathbb{E}\Big(\frac{\delta^2}{\delta\beta^2}ln(f_Y(Y|X^{real};\beta))\Big).$$

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

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

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 $\mathbf{Z}$ from Y. Therefore, the *Bartlett identities* are still verified.

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

$$\sum y_i(\beta_0 + \beta_1 x_1) - b(\beta_0 + \beta_1 x_1) + C^{st},$$

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

$$\int_{\mathbb{R}} |ln(f_Y(y|z;\beta))| dF_{Z_1}(z) \underset{\text{Triangle ineq.}}{\leq} \int_{\mathbb{R}} y_i|\beta_0 + \beta_1 z_1| dF_{Z_1}(z) + \int_{\mathbb{R}} |ln(1 + exp(\beta_0 + \beta_1 z_1))| dF_{Z_1}(z),$$
$$\underset{\frac{x}{1+x} \leq ln(x) \leq x}{\leq} \int_{\mathbb{R}} y_i|\beta_0 + \beta_1 z_1| dF_{Z_1}(z) + \int_{\mathbb{R}} exp(\beta_0 + \beta_1 z_1) dF_{Z_1}(z),$$

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 :

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

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

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

$$\begin{aligned} \mathbb{E}(ln(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p \, \mathbb{E}(ln(f_Y(Y|\mathbf{X}^{real}_{(*p)}, \mathbf{X}_p = \mathbf{X}^{real}_p; \beta))) \\ &\quad + (1 - Q_p) \int_{\mathbb{R}} \mathbb{E}(ln(f_Y(Y|\mathbf{X}^{real}_{(*p)}, \mathbf{X}_p = z; \hat{\beta}))) f_{Z_p}(z) dz. \end{aligned} \tag{A7}$$

can be written in a similar way than in the univariate case, using Fubini's theorem 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} .

A.4. Example 1: Log-Gaussian GLM

In this section, let focus on Log-Gaussian GLM case. Without loss of generality, remind that we can assume the covariate centered. (See the paper for uncentered case)

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

$$ln(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})) \propto \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i\beta)^2, \tag{A8}$$

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

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

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

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

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

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

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

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

The proof of these results are in A.8.

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 moment generating function $M_{X_p^{real}}(t) = M_{X_p}(t) = M_{Z_p}(t)$ for all $t \in \mathbb{R}$ and its derivatives' existence are ensured by the mild regularity condition A.2. We denote $Y = N$ and v the exposure to have more traditional notations. Let $\beta \in \mathbb{R}^{p+1}$. The sample estimator of the expected likelihood is equal to

$$\begin{aligned} ln(\mathcal{L}^{M_2}(\beta; \mathbf{N}|\mathbf{X})) = Q_p \, ln(\mathcal{L}(\beta; \mathbf{N}|X^{real}) + (1 - Q_p) \, ln(\mathcal{L}(\beta^{*p}; \mathbf{N}, X_{(*p)}^{real})) \\ + (1 - Q_p) \sum_{i=1}^{n} v_i e^{\beta^{*p} \mathbf{X}_{i;(*p)}^{real}} (1 - M_{X_p^{real}}(\beta_p)). \end{aligned} \tag{A9}$$

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

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

The expected likelihood can be bounded :

$$\begin{aligned} Q_p \, ln(\mathcal{L}(\beta; \mathbf{N}|X^{real})) + (1 - Q_p) \, ln(\mathcal{L}(\beta^{*p}; \mathbf{N}, X_{(*p)}^{real})) \\ + (1 - Q_p) \sum_{i=1}^{n} v_i e^{\beta^{*p} X_{(i);(*p)}^{real}} \left(1 - \exp\left(\frac{\beta_p^2 \mathbb{V}(X_p^{real})}{2}\right)\right) \\ \leq ln(\mathcal{L}^{M_2}(\beta; \mathbf{N}|\mathbf{X})) \leq Q_p \, ln(\mathcal{L}(\beta; \mathbf{N}|X^{real})) + (1 - Q_p) \, ln(\mathcal{L}(\beta^{*p}; \mathbf{N}, X_{(*p)}^{real})). \end{aligned} \tag{A10}$$

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

Proof. See A.9.1. □

Lemma A.2. *Let $\beta \in \mathbb{R}^{p+1}$. Under the assumption (X-A3), the derivative of the M2 log-likelihood for j in $\{0, ..., p-1\}$ are equal to :*

$$\begin{aligned}\frac{\delta}{\delta\beta_p}\mathbb{E}(ln(f_N(\beta;\mathbf{N}|\mathbf{X}))) &= Q_p\, d_p(\beta) \\ &\quad - (1-Q_p)\int_{\mathbb{R}^{p-1}} ve^{\beta^{*p}x^{real}_{(*p)}} dF_{\mathbf{X}^{real}_{(*p)}}(x^{real}_{(*p)})\, M'_{X_p}(\hat{\beta}_p); \\ \frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_N(\beta;\mathbf{N}|\mathbf{X}))) &= d_j(\beta),\end{aligned} \tag{A11}$$

where d_i is the derivative according to β_i of $\mathbb{E}(ln(f_N(\beta;\mathbf{N}|\mathbf{X}^{real})))$.

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

Proof. See A.9.1. □

When $\hat{\beta}_p \to 0$, $M'_{X_p}(\hat{\beta}_p) \to 0$. It can be easily shown that the derivative - equation A11 is a constant function of the mean quality. Therefore, the following proposition can be deduced.

A.5.0.1. Proposition 1. *Suppose the framework of this paper with log-Poisson distribution. Under the Assumptions (X-A3), e.g. $Q_j = 1$ for $j \in \{1, \cdots, p-1\}$ and $Q_p \in (0,1)$,*

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

is a monotonic function of the quality.

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

Theorem A.3. *Under the assumption (X-A3) and (Z-A1),*

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

Proof. See A.9.1. □

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. 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 A.9.4) :

$$\mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{X}^{real};\beta))) = \frac{1}{Q_1}\Big(\mathbb{E}(ln(f(\mathbf{N}|\mathbf{X},\beta)) - (1-Q_1)\Big(\mathbb{E}(ln(f(\mathbf{N}|\beta_0)) + \ vexp(\beta_0)\,(1 - M_{\mathbf{X}^{real}}(\beta_*))\Big)\Big), \tag{A12}$$

where $M_{\mathbf{X}^{real}}(\beta_*)$ is the multivariate generating function of $X_1^{real}, ..., X_p^{real}$ and $\beta_* = {}^t(\beta_1, ..., \beta_p)$. Unfortunately, no bounds can be state.

A.6. Example 3: Log-Gamma GLM

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

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

Here, we are only interested to maximize the log likelihood according to β for a known ν. Therefore, we will study

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

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

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

The M_1 estimator can be calculated

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

Lemma A.4. *Let $\beta \in \mathbb{R}^{p+1}$. Under the assumption (X-A3), the derivative of the M_2 log-likelihood for j in $\{0, ..., p-1\}$ are equal to :*

$$\begin{aligned}
\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_j(\beta) \\
&\quad + (1 - Q_p)\ \frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_Y(\beta^{*\mathbf{p}}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))M_{X_p}(\beta_p), \\
\frac{\delta}{\delta\beta_p}\mathbb{E}(ln(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_p(\beta) \\
&\quad + (1 - Q_p)\ \mathbb{E}(ln(f_Y(\beta^{*p}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))M^{'}_{X_p}(\beta_p),
\end{aligned} \tag{A13}$$

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

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

Proof. See A.10.1 □

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

$$Q_p\ \mathbb{E}(ln(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1 - Q_p)M_{\mathbf{X}^{real}}(-\beta)\mathbb{E}(ln(f_Y(\hat{\beta}_\mathbf{0}; \mathbf{Y}))).$$

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

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

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

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

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

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

Because of $ln(\mathbf{x}^{real}\beta^{*p} + z_p\beta_p)$, the sample analogue can not be estimated using only $\mathbf{X}^{real}$ and X which will not allowed 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 :

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

can not be calculated using only $\mathbf{X}^{real}$. $ln(1+exp(x)) = ln(2+exp(x)-1) = ln2+ln(1+(exp(x)-1)/2) \sim ln2 + (exp(x)-1)/2 + o((exp(x)-1))$ when $exp(x)$ is close to 1. Therefore, when $\mathbf{x}\beta$ is close to 0, a fine approximation with a multiplicative structure can be state.

A.8. Log-Gaussian proofs

A.8.1. Proof of equation A.4

Proof. We remind that X_1^{real} is supposed centered. The likelihood maximization solution can be found as the solution of the derivative equal to 0. Deriving by β, the sample estimator can be written as follows :

$$\begin{aligned}\frac{\delta\mathbb{E}(ln(f(\mathbf{N}|\mathbf{X},\hat{\beta}^{M2}))}{\delta\beta} =& \frac{\delta\mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta)))}{\delta\beta}\times Q_1 + \frac{\delta\mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{Z};\beta)))}{\delta\beta}\times(1-Q_1)\\ =& Q_1\frac{\delta}{\delta\beta}\int_{\mathbb{R}^2}(y - x\beta_1+\beta_0)^2 dF_{X_1^{real},Y}(x,y)\\ &+ (1-Q_1)\frac{\delta}{\delta\beta}\int_{\mathbb{R}}\int_{\mathbb{R}}(y - z\beta_1-\beta_0)^2 dF_{Z_1}(z) dF_Y(y) = 0.\end{aligned} \tag{A14}$$

We 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, we can apply the theorem of differential under the integral

and use the first *Bartlett identity*.

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

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

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

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

A.8.2. Proof of equation A.4

Proof. The first *Bartlett identity* under the assumption (X-A3) are equal:

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

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

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

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

A.8.3. Proof of equation A.4

Proof. Just for this proof, denote $\beta_* = (\beta_1, ..., \beta_p)$. In the case (C2) with perfectly correlated quality variable , i.e., $\Omega_j = \Omega_k \rightarrow Q_j = Q_k$ $(j \neq k)$, we can write :

$$\begin{aligned}
\frac{\delta \mathbb{E}(ln(f(\mathbf{N}|\mathbf{X}, \beta^{M_2}))}{\delta \beta} &= 0 \\
&= \frac{\delta \mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta^{M_2})))}{\delta \beta} \times Q_1 + \frac{\delta \mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{Z}; \beta^{M_2})))}{\delta \beta} \times (1 - Q_1) \\
&= Q_1 \frac{\delta}{\delta \beta} \int_{\mathbb{R}^{p+1}} (y - x^{real} \beta_*^{M_2} - \beta_0^{M_2})^2 dF_{X_1^{real}, ..., X_p^{real}, Y}(x, y) \\
&\quad + (1 - Q_1) \frac{\delta}{\delta \beta} \int_{\mathbb{R}} \int_{\mathbb{R}^p} (y - \mathbf{z} \beta_*^{M_2} - \beta_0^{M_2})^2 dF_{Z_1, ..., Z_p}(\mathbf{z}) dF_Y(y).
\end{aligned} \tag{A15}$$

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

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

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

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

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

A.9. *Proof for the GLM Log-Poisson*

A.9.1. *Proof of equations A9 and A10*

Proof. Under the assumption (X-A3), using the Fubini's theorem, the expected likelihood (without the constant) is equal to

$$\begin{aligned}
\mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{X};\beta))) \propto & Q_p \mathbb{E}(ln(f_N(N|\mathbf{X}^{real};\beta))) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^p} -ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z} + n(\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z) dF_{X^{real}_1,\dots X^{real}_{p-1},N}(\mathbf{x}^{real}_{(*p)},n) dF_{Z_p}(z) \\
\propto & Q_p \mathbb{E}(ln(f_N(N|\mathbf{X}^{real};\beta))) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^p} +ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} - ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z} dF_{X^{real}_1,\dots X^{real}_{p-1},N}(\mathbf{x}^{real}_{(*p)},n) dF_{Z_p}(z) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^p} -ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} + n(\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z) dF_{X^{real}_1,\dots X^{real}_{p-1},N}(\mathbf{x}^{real}_{(*p)},n) dF_{Z_p}(z) \\
= & Q_p \mathbb{E}(ln(f_N(N|\mathbf{X}^{real};\beta))) + (1-Q_p)\mathbb{E}(ln(f_N(N|\mathbf{X}^{real}_{(*p)};\beta^{*p}))) \\
& + (1-Q_p)\int_{\mathbb{R}^p} ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} dF_{X^{real}_1,\dots X^{real}_{p-1},N}(\mathbf{x}^{real}_{(*p)},n)(1-M_{X_p}(\beta_p)).
\end{aligned} \tag{A16}$$

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

$$\begin{aligned}
& \bar{Q}_p \; ln(\mathcal{L}(\beta|\mathbf{N},\mathbf{X}^{real})) + (1-\bar{Q}_p)\; ln(\mathcal{L}(\beta^{*p}|\mathbf{N},\mathbf{X}^{real}_{*p})) \\
& + (1-\bar{Q}_p)\sum_{i=1}^{n} v_i e^{\beta^{*p}\mathbf{X}^{real}_{i;(*p)}}(1-M_{X_p}(\beta_p)).
\end{aligned} \tag{A17}$$

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

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

With Hoeffding inequality, another bound can be deduced without needing a bounded variable[7] :

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

These inequalities leads to the equation A10. □

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

A.9.2. Proof of the Lemma A.2

Proof.

$$\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{X};\beta))) = Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_N(N|\mathbf{X}^{real};\beta))) + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^p} -v\mathbf{x}_j^{real}e^{\beta^{*p}\mathbf{x}_{(*p)}^{real}+\beta_p z} + n\mathbf{x}_j^{real}dF_{X_1^{real},\dots X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)dF_{Z_p}(z) \tag{A18}$$

We 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), $dF_{X_1^{real},\dots X_p^{real}}(\mathbf{x}^{real}) = dF_{X_1^{real},\dots X_{p-1}^{real}}(\mathbf{x}_{(*p)}^{real})dF_{X_p^{real}}(x_p^{real})$ for j in $\{1,\dots,p-1\}$. By replacing these values in the previous equation, we have :

$$\begin{aligned}\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{X};\beta))) &= Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_N(N|\mathbf{X}^{real};\beta))) \\ &\quad + (1-Q_p)\int_{\mathbb{R}^{p+1}} -vx_j^{real}e^{\beta\mathbf{x}^{real}} + n\mathbf{x}_j^{real}dF_{X_1^{real},\dots X_p^{real},N}(\mathbf{x}^{real},n) \\ &= Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_N(N|\mathbf{X}^{real};\beta))) \\ &\quad + (1-Q_p)\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_N(N|\mathbf{X}^{real};\beta))) = d_j(\beta),\end{aligned} \tag{A19}$$

The derivative according to β_p is calculated thanks to equation A17 without difficulty. This end the proof for equation A11. □

A.9.3. Proof of the theorem A.3

Proof. We know that the solution β^{M_2} exists and is unique. Moreover, the solution β^{M2} 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

$$\begin{aligned}d_j(\beta) &= \int_{\mathbb{R}}\int_{\mathbb{R}^p} -vx_j^{real}e^{\beta^{*p}\mathbf{x}_{*p}^{real}+\beta_p x^{real}}dF_{X_1^{real},\dots X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)dF_{X_p^{real}}(x^{real}) \\ &= \int_{\mathbb{R}^p} -vx_j^{real}e^{\beta^{*p}\mathbf{x}_{*p}^{real}+\beta_p x^{real}}dF_{X_1^{real},\dots X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)\underbrace{\int_{\mathbb{R}} e^{\beta_p x^{real}}dF_{X_p^{real}}(x^{real})}_{>0} = 0\end{aligned} \tag{A20}$$

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

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

$$\begin{aligned} d_j^{M_2}(b) = d_j(b) &= \int_{\mathbb{R}} \int_{\mathbb{R}^p} -v x_j^{real} e^{\beta^{*p} \mathbf{x}_{(*p)}^{real} + b_p x^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n) dF_{X_p^{real}}(x^{real}) \\ &= \underbrace{\int_{\mathbb{R}^p} -v x_j^{real} e^{\beta^{*p} \mathbf{x}_{*p}^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n)}_{=0} \int_{\mathbb{R}} e^{b_p x^{real}} dF_{X_p^{real}}(x^{real}) = 0. \end{aligned} \tag{A21}$$

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

$$d_p^{M_2}(\beta^{M_2}) = Q_p \, d_p(\beta^{M_2}) - (1 - Q_p) \int_{\mathbb{R}^p} v e^{\beta^{*p} X_{i;*p}^{real}} \, dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n) M_{X_p}^{'}(b_p). \tag{A22}$$

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

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

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

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

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

For $j = 1, ..., p$, 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 $\bar{Q}_1$ using the strong law of large number,

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

□

A.9.4. Proof of the Log poisson results for (C2) assumption

Proof. Just for this proof, denote $\beta_* = (\beta_1, ..., \beta_p)$. In the case (C2) with perfectly correlated quality variables, i.e., $\Omega_j = \Omega_k \rightarrow Q_j = Q_k$ $(j \neq k)$, we can write under (Z-A2):

$$\begin{aligned} \mathbb{E}(ln(f(\mathbf{N}|\mathbf{X}, \beta)) =& Q_1 \mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{X}^{real}; \beta))) + (1 - Q_1) \mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{Z}; \beta))) \\ =& Q_1 \mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{X}^{real}; \beta))) + (1 - Q_1) \mathbb{E}(ln(f_N(\mathbf{N}|\beta_0))) \\ &+ (1 - Q_1) \int_{\mathbb{R}} \int_{\mathbb{R}^p} -v e^{\beta_0 + \beta_* \mathbf{z}} + n(\beta_0 + \beta_* \mathbf{x}_{(*p)}^{real} + \beta_p z) dF_{Z_1,...Z_p}(\mathbf{z}) dF_N(n) \\ =& Q_1 \mathbb{E}(ln(f_N(\mathbf{N}|\mathbf{X}^{real}; \beta))) + (1 - Q_1) \mathbb{E}(ln(f_N(\mathbf{N}|\beta_0))) \\ &+ (1 - Q_1) \, v exp(\beta_0) \, M_{\mathbf{Z}}(\beta_*), \end{aligned} \tag{A23}$$

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

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

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

$$\mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))) = \frac{1}{Q_1} \Big(\mathbb{E}(ln(f(\mathbf{N}|\mathbf{X}, \beta)) - (1 - Q_1) \mathbb{E}(ln(f_Y(\mathbf{Y}|\beta_0))) - (1 - Q_1)\ vexp(\beta_0)\ M_{\mathbf{X}^{real}}(\beta_*) \Big). \tag{A24}$$

□

A.10. Proof for GLM log gamma

A.10.1. Proof of the lemma A.6

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

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

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

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

Under the assumption (X-A3), the derivative of the M2 log-likelihood for j in $\{0, ..., p-1\}$ are equal to

$$\begin{aligned}
\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) &= Q_p\, d_j(\beta) + (1-Q_p) \\
&\quad \frac{\delta}{\delta\beta_j}\int_{\mathbb{R}}\int_{\mathbb{R}^p} -y\, exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p) - \mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real}, y) dF_{Z_p}(\mathbf{z}_p) \\
&\quad = Q_p\, d_j(\hat{\beta}) + (1-Q_p) \\
&\quad \int_{\mathbb{R}}\int_{\mathbb{R}^p} -y x_j^{real} exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p) dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real}, y) dF_{Z_p}(\mathbf{z}_p) \\
&\quad = Q_p\, d_j(\beta) + (1-Q_p)\frac{\delta}{\delta\beta_j}\mathbb{E}(ln(f_Y(\beta^{*p};\mathbf{Y}|\mathbf{X}^{real})))M_{X_p}(-\beta_p), \\
\frac{\delta}{\delta\beta_p}\mathbb{E}(ln(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) &= Q_p\, d_p(\beta) + (1-Q_p) \\
&\quad \int_{\mathbb{R}}\int_{\mathbb{R}^p} -y z_p exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p) dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real}, y) dF_{Z_p}(\mathbf{z}_p) \\
&\quad = Q_p\, d_j(\beta) - (1-Q_p)\mathbb{E}(ln(f_Y(\beta^{*p};\mathbf{Y}|\mathbf{X}^{real})))M'_{X_p}(-\beta_p).
\end{aligned} \tag{A25}$$

□

A.11. Proof of the Theorem 3.1

Proof. Let $(\mathbf{Y}, \mathbf{X}, \boldsymbol{Q})$ 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,

$$\begin{aligned}
\mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X};\beta))) &= \mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))|\Omega = 1) \times \mathbb{P}(\Omega = 1) \\
&\quad + \mathbb{E}(ln(f_Y(\mathbf{Y}|\mathbf{Z};\beta))|\Omega = 0) \times \mathbb{P}(\Omega = 0).
\end{aligned} \tag{A26}$$

The first term is known and as Z is independent of Y, the second can rewritten, using Fubini's theorem :

$$\mathbb{E}(ln(f_Y(Y|\mathbf{Z};\hat{\beta}))) = \mathbb{E}_Y \int_{\mathbb{R}} ln(f_Y(Y|z;\beta)) dF_{Z_1}(z). \tag{A27}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{h=1}^{n} log(f(y_i|X_{h,1}^{real};\hat{\beta})). \tag{A28}$$

Having $\mathbb{E}(|log(f_Y(Y|\ \mathbf{Z}, \beta)|) < \infty$ from the strong law of large numbers, this sample estimator converges almost surely. The sample estimator $\sum ln(f_Y(y_i|X_{i;1}^{real};\hat{\beta}))$ converges almost surely

$\mathbb{E}(ln(f_Y(Y|\mathbf{X}^{real};\hat{\beta}))$. Using the strong law of large number, $\bar{Q}_1$ converges in probability towards Q_1. Thus,

$$\bar{Q}_1 \sum_{i=1}^{n} ln(f_Y(y_i|X_{i,1}^{real};\hat{\beta})) + (1-\bar{Q}_1) \times \sum_{i=1}^{n} \frac{1}{n} \sum_{h=1}^{n} ln(f_Y(y_i|X_{h,1}^{real};\hat{\beta})). \tag{A29}$$

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

Finally, we know that :

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

Therefore, using the Cramer-Rao conditions - Collorary 3.8 of [11], the global maximum exists, is unique and converges in probability to β^{M2}, i.e.,

$$\hat{\beta}^{M2|X^{real},\mathbf{Q}} \xrightarrow{P} \beta^{M2},$$

meaning that the estimator is consistent. □

Appendix B. Convexity: Propositions 1 and 2

Proof. Denote X_j, X_k $(i \neq j)$ the covariates with a Pearson correlation of $|\rho| \neq 1$ and suppose $\beta_k^{M_1}$ and $\beta_j^{M_1}$ non null. Using the Corollary 3.5.1 of [4], the following derivatives are found :

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

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

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

$$0 \leq \beta_k(1 - Q_j^2\rho^2) + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j^{M_1}\rho(1 - Q_j^2)$$

$$0 \leq \beta_k + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j\rho - Q_j^2(\rho^2\beta_k - \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j\rho),$$

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

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