Regression-based imputation of explanatory discrete missing data

G. Hernández-Herrera[1,2], A. Navarro[1], and D. Moriña*[3]

[1]Research Group on Psychosocial Risks, Organization of Work and Health (POWAH), Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona
[2]Instituto de Investigaciones Médicas, Facultad de Medicina, Universidad de Antioquia
[3]Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona (UB)

Abstract

Imputation of missing values is a strategy for handling non-responses in surveys or data loss in measurement processes, which may be more effective than ignoring the losses and omitting them. The characteristics of variables presenting missing values must be considered when choosing the imputation method to be used; in particular when the variable is a count the literature dealing with this issue is scarce. If the variable has an excess of zeros it is necessary to consider models including parameters for handling zero-inflation. Likewise, if problems of over- or under-dispersion are observed, generalisations of the Poisson, such as the Hermite or Conway-Maxwell Poisson distributions are recommended for carrying out imputation. The aim of this study was to assess the performance of various regression models in the imputation of a discrete variable based on Poisson generalisations,

*Corresponding Author: David Moriña (dmorina@mat.uab.cat)

in comparison with classical counting models, through a comprehensive simulation study considering a variety of scenarios and a real data example. To do so we compared the results of estimations using only complete data, and using imputations based on the most common count models. The COMPoisson distribution provides in general better results in any dispersion scenario, especially when the amount of missing information is large.

Keywords: missing data, count data, Hermite, COMPoisson, zero-inflated, multiple imputation

1 Introduction

Missing data are practically unavoidable in research in any field, however their consequences for the validity of research findings are not considered in the majority of cases. Nowadays, many scientific journals emphasise the importance of including information about missing data and the strategy used to handle them ([6]), and yet it is still not common to find this, and very few publications explicitly describe missing data and the methods used to deal with them. Data can be missing due to non-responses or be caused by problems in study design, or even created deliberately by researchers as part of privacy policies. In order to identify the behaviour of missing data in a sample, in terms of quantity and mechanisms of data loss, the first step is to carry out a statistical description of the observed variables, identify the data lost in each one, and then identify the patterns of data loss which will help to make a decision about a method for handling the missing data. In many cases, researchers resort to using only data which are completely available, ignoring those which are missing when performing analyzes; but this decision can generate problems of bias and lack of precision in estimates and affect the power of the study due to the reduction in sample size.

In order to adequately deal with missing data it is necessary to identify the mechanism of data loss (process of non-response) defined as the origin, causes, moment, relationships or characteristics which give rise to the lack of information. Moreover, it is important to establish whether the observations have been lost randomly, or whether their loss is associated with definable causes, and to determine the percentage of missing data in the sample, since depending on these factors, the levels of uncertainty in working with imputed data can vary significantly ([15]).

The mechanism of data loss is classified depending on the probability of response: if this is independent of the observed and unobserved data ($P(Y \text{ missing} \mid X, Y) = P(Y \text{ missing})$), we say the non-response process is MCAR, missing completely at random. If however it is dependent on the observed data ($P(Y \text{ missing} \mid X, Y) = P(Y \text{ missing} \mid X)$), we say the non-response process is MAR, missing at random. When the process is neither MAR nor MCAR ($P(Y \text{ missing} \mid X, Y) = P(Y \text{ missing} \mid X, Y)$) it is termed *informative*, NMAR, not missing at random ([21]).

The advantage, statistically speaking, of the mechanism of loss being MCAR is that conclusions obtained from the analysis of these data can still be valid. The power of the study can be affected by the design, but the parameters estimated are not biased by the absence of data. However, MAR is the most common mechanism and thus it is important to take into account that the probability of data loss is conditionally independent; but if the mechanism is NMAR, this represents a difficulty for imputation, since the estimation of parameters requires knowledge of the model of data loss in order to achieve unbiased estimates.

[18], [25] and [9] have constructed a taxonomy of the most popular methods for handling missing data, which shows that when the data loss mechanism is completely random (MCAR) the methods stochastic regression, multiple imputation, maximum likelihood with Expectation-Maximization (EM) algorithm, Bayesian imputation, and weighted methods produce consistent estimates; when the mechanism of data loss is random (MAR) multiple imputation produces consistent estimates but only under certain conditions, just as with weighted imputation methods, whereas for the non-random mechanism (NMAR) it is more complicated to obtain consistent estimates as the correct probability model of data loss needs to be properly identified ([34]). This issue has been studied in detail in specific situations like zero inflation in [19].

In many studies, counting variables appear, for example in health research it is common to study episodes of a particular disease and these form the basis of estimates of incidence, when the time-period in which they are observed is also taken into account. Although in the presence of missing data of this type, it would seem logical that the imputation model be based on discrete distributions, such as Poisson or Negative Binomial, in a biomedical and epidemiological literature review, it was found that very few studies employed these specific methods to impute discrete variables. In practice, the procedures used to handle such missing data were as though the variable had been continuous, or treating it as categorical or ordinal, using polytomic regression techniques, and in other cases using the strategy of applying some normalizing transformation to the data, and subsequently using imputation methods for normal data ([16]).

Much less common is the use of other more sophisticated techniques which tackle specific situations, such as the problem of inflated zeros ([24]), a situation which may be explained by the lack of software available to carry out imputations of this type. However, some packages have now been developed, for example in R, which allow imputation of these kinds of data ([15]) and some studies have stressed the utility of Poisson and Negative Binomial zero-inflated models for epidemiological studies with both cross-sectional and longitudinal designs ([17]). Few studies use generalized distributions, which are more flexible and take into account problems of over- and under-dispersion, also common in health data: for example no studies are found employing the Hermite distribution, a generalization of the Poisson distribution, more flexible when there is over-dispersion for the imputation variables of count data. Nor are studies found employing the Conway Maxwell Poisson distribution which permits modeling count data in the presence of over- and under-dispersion.

It is also not common in scientific literature to find an exhaustive examination of missing information in a discrete variable acting as covariate in an analysis. However, this point is crucial in certain contexts such as survival analysis in the presence of recurrent events, where the usual analysis techniques introduced in [1] and [26] require knowledge of the number of previous events suffered by individuals, something which of often unknown (for example events which occurred prior to initiation of a cohort study). The aim of the present study is precisely to assess the performance of methods of imputation of missing data in a discrete covariate, based on generalizations of the Poisson distribution, in comparison to the classical Poisson and Negative Binomial counting methods, in different scenarios of dispersion and nature of response variable and within a framework of multiple imputation, following the recent recommendations in many scenarios like confirmatory clinical trials ([3]). Although applied researchers were reluctant to using multiple imputation methods until recently, the implementation in most used data analysis software has increased their popularity in the latest years. [14] present an alternative based on regression models, but accounting only for continuous covariates. The considered regression models are described in the next section and their performance on real lung cancer clinical trial data and on a comprehensive simulation study are analyzed in Section *Results*.

2 Methods

In this section we present some discrete variable regression models, on the basis of which to carry out imputation of missing data. The classical counting models including Poisson, negative binomial (parameterized in terms of its mean μ and dispersion index d) and their zero-inflated versions are described in the supplementary material, and only the less known distributions (Hermite and Conway-Maxwell Poisson) are presented here. These distributions are very flexible and may be adapted and used in any scenario of dispersion. A comprehensive description of most common count data modeling strategies with special focus on dispersion issues can be found in [8]. In our context, the main interest is in fitting a generalized linear model (GLM, [20]) according to the nature of the response variable Y (linear for continuous, logistic for binary, Poisson or other discrete regression models for count data) using a discrete explanatory variable X with missing observations. In order to carry out the imputation of missing data using the regression models described in this section, two phases are required; firstly, a generalized linear model is fitted using the covariate X as response and the response Y as covariate, based on the corresponding distribution (Poisson, NB, ZIP, ZINB, Hermite or COMPoisson). On the second phase, imputed values are randomly sampled from the corresponding distribution with the parameters obtained in the previous step, including random noise generated from a normal distribution in all cases (as the parameters are estimated by maximum likelihood). Let's assume we are interested in fitting the model $Y = \beta_0 + \beta_1 \cdot X$, where X is a discrete explanatory variable, and consider a vector of the parameters of the model $\beta = (\beta_0, \beta_1)$. In order to produce proper estimation of uncertainty, the described methodology can easily be extended to a multiple imputation framework. The results reported in Section *Results* correspond to the combination of $m = 50$ imputed data sets, according to the well known Rubin's rules ([31]) and based on the following steps in a Bayesian context:

1. Fit the corresponding count data model and find the posterior mean and variance $\hat{\beta}$ and $V(\hat{\beta})$ of model parameters β.

2. Draw new parameters β^* from $N(\hat{\beta}, V(\hat{\beta}))$, where $\hat{\beta}$ is the maximum likelihood estimate of a parameter β.

3. Compute predicted scores p using the parameters obtained in the previous step (the actual expression depends on the count data model).

4. Draw imputations from the corresponding count data distribution and scores obtained in the previous step.

For a recent review of multiple imputation methods with special focus on medical research, see [13].

The performance of these models in different scenarios will be compared in the next section. The simulation strategy is described in detail in Section *Simulation study*.

2.1 Hermite distribution

Several generalizations of Poisson distribution have been considered in literature. A remarkable approach to Poisson generalizations are two-parameter discrete distributions closed under convolutions and satisfying that the sample mean is the maximum likelihood estimator of the population mean, which are characterized in [27]. Within this family of distributions, a case of special interest are the so-called *compound-Poisson* or *contagious* distributions. They are families with probability generating function (PGF) defined by

$$P(s) = exp(\lambda(f(s){-}1)) = \exp(a_1(s{-}1){+}a_2(s^2{-}1){+}\ldots{+}a_m(s^m{-}1){+}\ldots), \quad (1)$$

where $f(s)$ is also a PGF and $\sum_{i=1}^{m} a_i = \lambda$. One of these families is the Generalized Hermite distribution, first introduced in [5] as the situation where a_m is significant compared to a_1 in (1), while all the other terms a_i are negligible, resulting in the PGF

$$P(s) = \exp(a_1(s-1) + a_m(s^m - 1)). \quad (2)$$

After fixing the value of the positive integer $m \geq 2$, the *order* or *degree* of the distribution, the domain of the parameters is $a_1 > 0$ and $a_m > 0$. Note that when a_m tends to zero, the distribution tends to a Poisson. Otherwise, when a_1 tends to zero it tends to m times a Poisson distribution. It is immediate to see that the PGF in (2) is the same than the PGF of $X_1 + mX_2$, where X_i are independent Poisson distributed random variables with population mean a_1 and a_m respectively. From here, it is straightforward to calculate the population mean, variance, skewness and excess kurtosis of the Generalized Hermite distribution:

$$\begin{aligned}
\mu &= a_1 + m a_m, \\
\sigma^2 &= a_1 + m^2 a_m, \\
\gamma_1 &= \frac{a_1 + m^3 a_m}{(a_1 + m^2 a_m)^{3/2}}, \\
\gamma_2 &= \frac{a_1 + m^4 a_m}{(a_1 + m^2 a_m)^2}.
\end{aligned} \tag{3}$$

A useful expression for the probability mass function of the Generalized Hermite distribution in terms of the population mean μ and the population index of dispersion $d = \sigma^2/\mu$ is provided in [27].

$$P(Y=k) = P(Y=0)\frac{\mu^k (m-d)^k}{(m-1)^k} \sum_{j=0}^{[k/m]} \frac{(d-1)^j (m-1)^{(m-1)j}}{m^j \mu^{(m-1)j} (m-d)^{mj} (k-mj)!j!}, k = 0, 1, \ldots \tag{4}$$

where $P(Y=0) = \exp(\mu(-1 + \frac{d-1}{m}))$ and $[k/m]$ is the integer part of $\frac{k}{m}$. Note that m can be expressed as $m = \frac{d-1}{1+\log(p_0)/\mu}$.

The probabilities can be also written in terms of the parameters a_1, a_m using the identities given in (3).

The case $m = 2$ in (2) is covered in detail in [12] and [11] and the resulting distribution is simply called Hermite distribution. In that case, the probability mass function, in terms of the parameters a_1 and a_2, has the expression

$$P(Y=k) = e^{-a_1-a_2} \sum_{j=0}^{[k/2]} \frac{a_1^{k-2j} a_2^j}{(k-2j)!j!}, k = 0, 1, \ldots \tag{5}$$

The Hermite distribution is said to be zero-inflated with respect to the Poisson distribution, because the probability of the variable taking value zero under Hermite is greater than under Poisson, when the two distributions have the same mean. This characteristic of the Hermite distribution allows proposing the use of a Hermite regression model for count variable with an excess of zeros, instead of using a classical Poisson regression model.

Given a sample $X = x_1, \ldots, x_n$ of a population coming from a generalized Hermite distribution with mean μ, index of dispersion d and order m, the log-likelihood function is

$$l(X;\mu,d) = n{\cdot}\mu{\cdot}\left(-1+\frac{d-1}{m}\right)+log\left(\frac{\mu(m-d)}{m-1}\right)\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}log(q_i(\theta)), \quad (6)$$

where $q_i(\theta)=\sum_{j=0}^{[x_i/m]}\frac{\theta^j}{(x_i-mj)!j!}$ and $\theta=\frac{(d-1)(m-1)^{(m-1)}}{m\mu^{(m-1)}(m-d)^m}$.

The maximum likelihood equations do not always have a solution. It is due to the fact that this is not a regular family of distributions because its domain of parameters is not an open set. The following result gives a sufficient and necessary condition for the existence of such a solution [27]:

Proposition 1. Let $x_1,\ldots,x_n$ be a random sample from a generalized Hermite population with fixed m. Then, the maximum likelihood equations have a solution if and only if $\frac{\mu^{(m)}}{\bar{x}^m}>1$, where $\bar{x}$ is the sample mean and $\mu^{(m)}$ is the m-th order sample factorial moment, $\mu^{(m)}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i(x_i-1)\cdots(x_i-m+1)$.

If the likelihood equations do not have a solution, the maximum of the likelihood function (6) is attained at the border of the domain of parameters, that is, $\hat{\mu}=\bar{x}$, $\hat{d}=1$ (Poisson distribution), or $\hat{\mu}=\bar{x}$, $\hat{d}=m$ (m times a Poisson distribution). The case $\hat{\mu}=\bar{x}$, $\hat{d}=m$ corresponds to the very improbable situation where all the observed values were multiples of m. Then, in general, when the condition of Proposition 1 is not satisfied, the maximum likelihood estimators are $\hat{\mu}=\bar{x}$, $\hat{d}=1$. It means that data is fitted assuming a Poisson distribution.

The R implementation of the Hermite distribution basic functions and regression model (package *hermite*) is detailed in [22].

2.2 Conway-Maxwell Poisson distribution

Let x be a random count variable with the following probability distribution function:

$$P(x,\lambda,\nu)=\frac{\lambda^x}{(x!)^\nu Z(\lambda,\nu)}, \quad (7)$$

where $\lambda=E(x^\nu)$ with ν the dispersion parameter and $Z(\lambda,\nu)=\sum_{j=0}^{\infty}\frac{\lambda^\nu}{(j!)^\nu}$ is a normalizing constant. If $\nu=1$ the distribution is equidispersed, and it is overdispersed (underdispersed) if $\nu<1$ ($\nu>1$).

The COMPoisson distribution, characterized by the previous probability distribution function, is a two-parameter extension of the Poisson distribution. It can be thought as a generalization of certain discrete distributions, such as the Poisson distribution when $\nu = 1$, the Bernoulli distribution with probability $\frac{\lambda}{1+\lambda}$ when $\nu \to \infty$ and the geometric when $\nu = 0$ and $\lambda < 1$. It was first suggested in 1962 by Conway and Maxwell in [4] (see [33] for a recent review of its applications). This distribution may also be seen as a weighted Poisson distribution with weighting function $w_y = (y!)^{1-\nu}$. In this sense, [30] compared the COMPoisson with a weighted Poisson where the weights take the following form:

$$W_y = \begin{cases} e^{-\beta_1(\lambda - y)} & \text{if } y \leq \lambda \\ e^{-\beta_2(y-\lambda)} & \text{if } y > \lambda. \end{cases} \tag{8}$$

There is underdispersion when $\beta_1, \beta_2 > 0$, overdispersion when $\beta_1, \beta_2 < 0$ and equidispersion when $\beta_1 = \beta_2 = 0$.

This is a flexible distribution which can take account of the excessive or insufficient dispersion often found in count data ([2]) and allows modeling count data in three dispersion scenarios: equidispersion, overdispersion and underdispersion, and is therefore an interesting alternative from the point of view of the present study. For example, the COMPoisson distribution has been used in linguistics to model word length, to model count data in marketing, and eCommerce, and to model grocery shop sales, among other uses. The capacity to handle different types and levels of dispersion makes the distribution more useful in applications where the level of dispersion may vary ([32]).

2.3 Simulation study

In order to assess the efficiency of the methods considered in this study, we designed a simulation procedure based on the following algorithm.

1. Generate a population of size 1,000,000 with two variables. One variable Y, following a binomial, normal or Poisson distribution will be used as the response, depending on the scenario considered (binary, continuous or discrete response variable), and a second will be used as the explanatory variable X, consisting of a count of the number of events and based on a Poisson, negative binomial or zero-inflated distribution, depending on the scenario of dispersion and zero-inflation.

The covariate X is generated first randomly sampling from the corresponding distribution and then the response Y is generated on the basis of a GLM with different intensities of association with the explanatory variable ($\beta = 0.5$, $\beta = -0.3$ and $\beta = -0.5$) although we only report results corresponding to $\beta = 0.5$ because no relevant differences were observed compared to the other values. These values of β were considered to keep the association between the response variable and the coviarate within the usual ranges, with a moderate and reasonable strength.

2. From the generated population, randomly select 1000 samples of size 2000 and generate missings in the explanatory variable using a MAR or MCAR mechanism, and for percentages of 5% to 30%. Regarding missings generated via MAR, the criteria was that 80% of missings corresponded to values of 0 for the binary variable or values below the mean for discrete or continuous variables and 20% of the missings corresponded to values of 1 for the binary variable or values above the mean for discrete or continuous variables.

3. For each sample, fit logistic, linear or Poisson regression models depending on the response variable, as follows:

    - **Listwise deletion (lw)**. Fit the regression model eliminating missing values, in other words, using only the information available.
    - **Poisson (pois)**. Fit a Poisson regression model using the count variable as response. Based on the estimated coefficients, impute missing values and subsequently fit the regression model, incorporating the set of imputed values.
    - **Negative Binomial (nb)**. Similar to the above procedure, but impute missing values using a Negative Binomial regression model.
    - **Hermite**. In this case, use a Hermite regression model to impute the missing values in the discrete variable.
    - **COMPoisson (cmp)**. Use Conway-Maxwell Poisson regression model to perform imputation of the missing data.
    - **Zero-inflated Poisson (zpois)**. Use a Poisson regression model with zero-inflation to impute the missing values.

- **Zero-inflated negative binomial (znb)**. Use a Negative Binomial regression model to impute the missing values.

The efficiency of the proposed imputation methods was assessed by comparing relative bias with respect to the population parameter $\left(\mid \frac{\bar{\hat{\beta}}-\beta}{\beta} \mid \right)$, where $\bar{\hat{\beta}}$ is the average of the estimate of a parameter β over the imputed data sets), the average length of 95% confidence intervals for the parameter of interest (AIL, computed as the average length of the 95% confidence intervals for the parameter of interest over all imputed data sets) and the coverage probability (computed as the proportion of imputed data sets with 95% confidence intervals including the true value of the population parameter). The same procedure was also carried out using samples of size $n = 200$ but the results showed no differences with those reported, and have been omitted. The following dispersion scenarios were considered:

- Equidispersion: the explanatory variable was generated following a Poission distribution with parameter $\lambda = 2$. The procedure described was performed using other values for λ but the results did not differ from those reported.

- Overdispersion: the explanatory variable was generated following a negative binomial distribution with mean $\mu = 2$ and dispersion index $d = 2$. The procedure described was performed using other values for μ and d but the results did not differ from those reported.

- Underdispersion: the explanatory variable was generated following a Poisson distribution with parameter $\lambda = 2$, and the underdispersion was generated through an iterative procedure, substituting values for the mean at random until a dispersion of $d = 0.5$ was obtained. The procedure described was performed with other values for λ and d but the results did not differ from those reported. In this case, the methods based on the Hermite distribution or on zero-inflated distributions were not considered due to convergence issues.

- Excess zeros: the explanatory variable was generated following a Poisson distribution with parameter $\lambda = 2$ and subsequently a random 10% of values were replaced by zeros. The procedure described was performed using other values for λ and other proportions of zero-inflation

($p = 0.1, 0.3, 0.4$ and 0.6) but the results did not differ from those reported.

The tables including the results for all kinds of response variables and all dispersion scenarios are available as supplementary material.

3 Results

The performance of the considered imputation methods is compared in this section in a real data example from a randomized clinical trial (RCT) of two treatment regimens for lung cancer, first introduced in [10]. The considered variables are survival time (continuous response, Y) and the number of months from diagnosis to randomisation (discrete explanatory variable, X), in which a quantity of random missing values were introduced 10%, 20%, 30% and 40%. Although the main interest in practice in a RCT would be to estimate the treatment effect, we focus here on the effect of the discrete covariate over the continuous response, as it was an observational study like the simulation study presented in Section *Simulation study.*

3.1 Lung cancer data

The explanatory variable number of months from diagnosis to randomisation is clearly overdispersed (the variance is 112.62 while the mean is 8.77). Fitting a linear regression model ($Y = \beta_0 + \beta_1 \cdot X$) to the full data without missing observations in the discrete explanatory variable X, the estimate is $\hat{\beta_1} = -0.69$ ($SD = 1.28$). Table 1 shows three performance measures for each of the considered imputing distributions (all cases included in a multiple imputation framework with $m = 50$ imputed data sets) in each missing data scenario (percentage of missing observations from 10% to 40%). The average interval length (AIL) reveals the uncertainty around the estimate in each case, the relative bias shows the deviation of the average estimate with respect to the reference value of $\hat{\beta_1} = -0.69$, while the p-values are obtained from χ^2 goodness of fit tests comparing the distribution of the original explanatory variable with no missing observations and the distribution of the same variable with imputed missing observations according to each considered distribution.

The 95% confidence intervals include the reference value of $\hat{\beta_1} = -0.69$ in all cases.

Table 1: Average interval length (AIL) and relative bias for each imputation method. p-values from χ^2 goodness of fit test comparing the full data and imputed distributions.

% of missing data	Model	Relative bias	AIL	p-value
10%	Listwise deletion	0.09	5.54	0.050
	Poisson	0.10	5.36	0.136
	Negative binomial	0.08	5.30	0.155
	Hermite	0.12	5.35	0.140
	COMPoisson	0.16	5.27	0.115
20%	Listwise deletion	0.24	5.78	0.012
	Poisson	0.16	5.48	0.192
	Negative binomial	0.04	5.39	0.136
	Hermite	0.24	5.46	0.176
	COMPoisson	0.11	5.34	0.096
30%	Listwise deletion	0.80	8.93	0.048
	Poisson	0.76	8.29	0.239
	Negative binomial	0.82	7.97	0.192
	Hermite	0.89	8.25	0.196
	COMPoisson	0.67	7.96	0.200
40%	Listwise deletion	0.55	9.79	0.134
	Poisson	0.40	8.54	0.246
	Negative binomial	0.44	7.93	0.238
	Hermite	0.50	8.42	0.235
	COMPoisson	0.24	8.05	0.206

As can be seen in Table 1, highest relative biases are obtained when missing data is not imputed (listwise deletion), with largest 95% confidence intervals and significant deviation from the full data distribution (χ^2 test p-values under 0.05 for the 10% - 30% of missing observations). On the other hand, Table 1 also shows that all the considered discrete distributions perform similarly when the proportion of missing observations in the explanatory variable is relatively small (10% - 20%) but missing data imputation based on the COMPoisson distribution is the best performing method in a scenario with a high proportion of missing data. In particular, for the 40% missing observations scenario it can be seen that estimates for β_1 obtained imputing the missing data on the basis of the COMPoisson distribution are clearly closer to the reference value (lowest relative bias) yet with a reasonable average 95% confidence interval length and no significant deviation of the imputed explanatory variable distribution with respect to the corresponding original full data distribution.

3.2 Simulation study results

Below we present the results for estimations of the β_1 coefficient and standard error of the regression models fitted, handling missing values in the cases of continuous response variables by the methods: listwise deletion (lw), imputation with Poisson regression (pois), imputation with negative binomial regression (nb), imputation with Hermite regression (herm), imputation using the COMPoisson model (comp), imputation with zero-inflated Poisson model (zpois) and imputation with zero-inflated negative binomial model (znb). For the other types of response variable, results are provided as supplementary material. We also present biases in the estimations and true coverage indices of the confidence intervals for the same case in all scenarios.

According to these results, we observe a considerable increase in bias when over 20% of values are missing, in all scenarios of dispersion and response variable (see Tables S1, S2, S3 and S4) in the Supplementary material. Figure 1 shows the behaviour of bias for each imputation method under the scenario of equidispersion, and it may be seen that in the estimate using only the information available without imputation (listwise deletion), the bias is greater in comparison to the other methods. The same trends can be seen for the average length of the confidence intervals and their coverage (see Figure 2 and Figure 3).

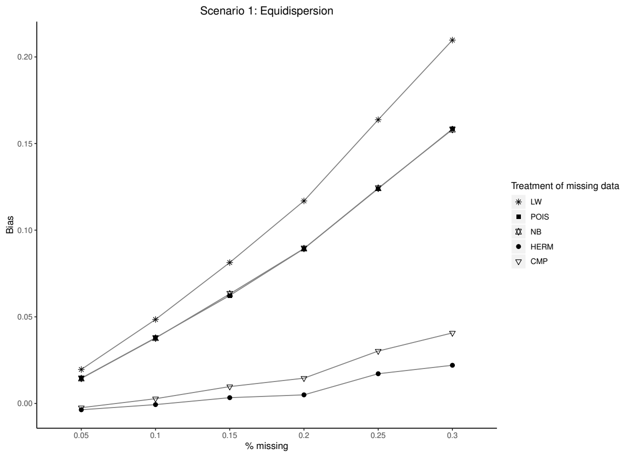

Figure 1: Bias in the coefficient estimate for the scenario of equidispersion with continuous response variable.

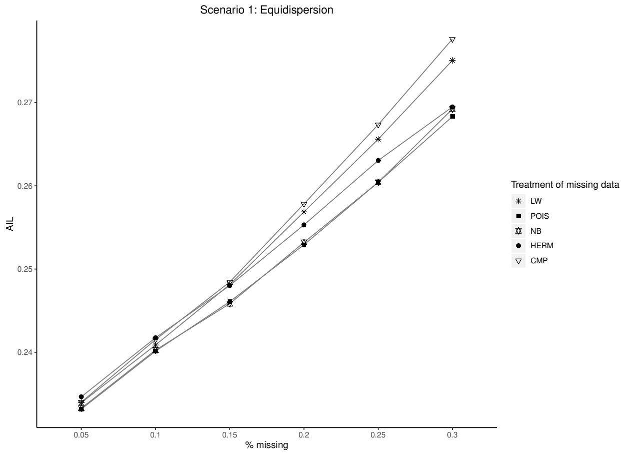

Figure 2: Average length of confidence intervals of the coefficient in the scenario of equidispersion and continuous response variable.

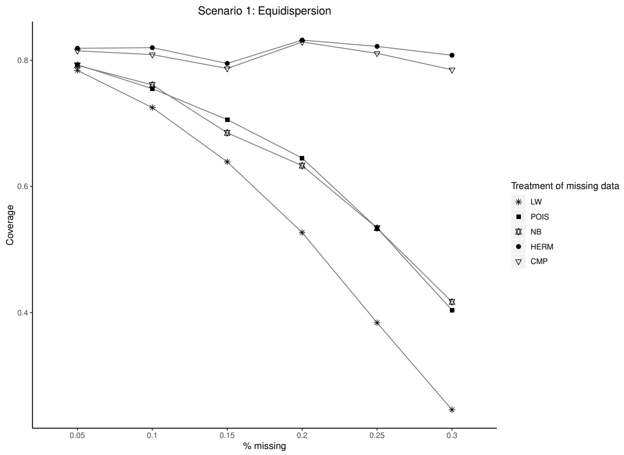

Figure 3: Coverage of confidence intervals for the coefficient in the scenario of equidispersion and continuous response variable.

In the scenario of overdispersion with a continuous response variable, the behaviour of bias is similar when listwise deletion is used, or when using zero-inflated models to carry out the imputation; lower bias may be seen when imputation is done with the negative binomial or COMPoisson models, as Figure 4 shows, however coverage of the confidence intervals is low compared to the other methods (see Figure 6). As expected, the average lengths of 95% confidence intervals increase with the proportion of missing observations (see Figure 5) in a very uniform manner except for the COMPoisson and zero-inflated negative binomial imputation models, with slightly wider confidence intervals in general.

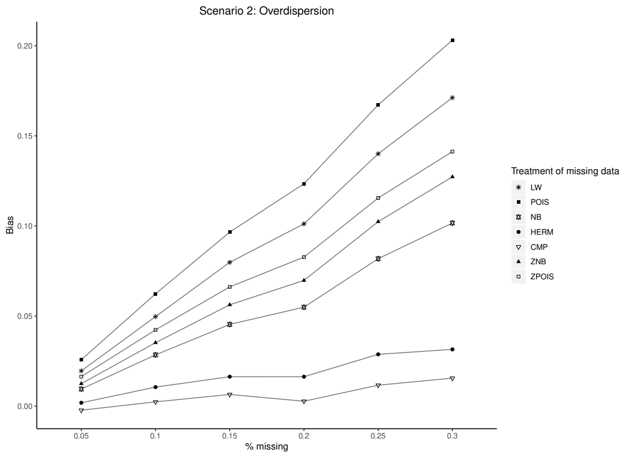

Figure 4: Bias in coefficient estimation when there is overdispersion and the response variable is continuous.

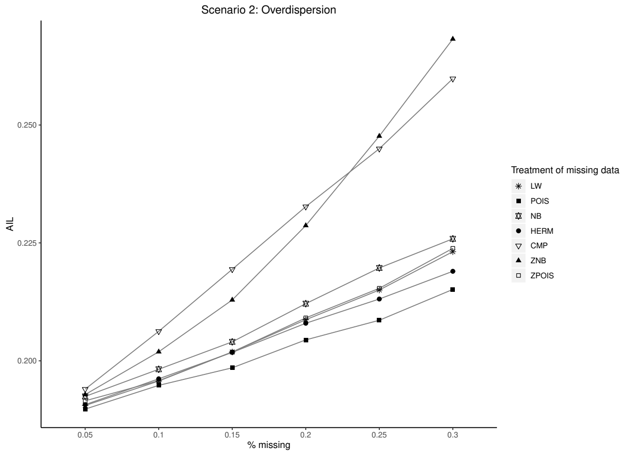

Figure 5: Average length of confidence intervals of the coefficient when there is overdispersion and the response variable is continuous.

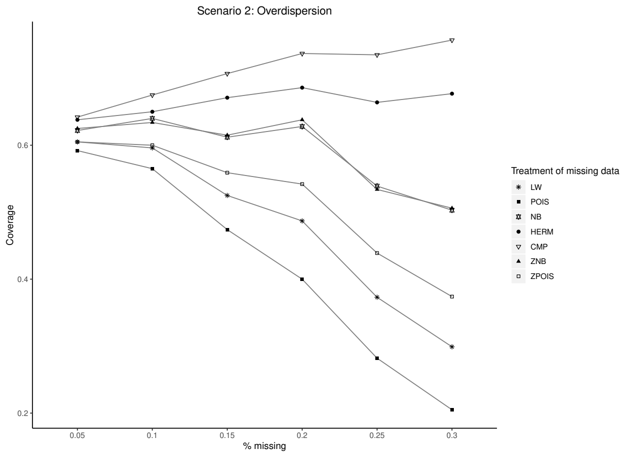

Figure 6: Coverage of confidence intervals of the coefficient in the scenario of overdispersion and continuous response variable.

In the case of underdispersion the results presented in Figure 7 show that in this scenario the biases are lower when imputing using the Poisson and Negative Binomial regression models (in this scenario it is not possible to obtain the maximum likelihood estimators corresponding to the Hermite distribution) and that the coverage of confidence intervals is greater for these two models (see Figure 9). As in the previous scenarios, Figure 8 shows a very uniform increase in the average 95% confidence intervals as the proportion of missing observations increases except the COMPoisson based imputation model, with wider intervals.

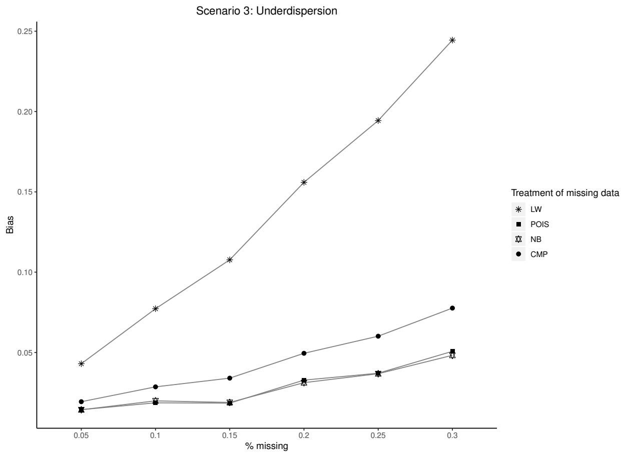

Figure 7: Bias in estimation of the coefficient when there is underdispersion and the response variable is continuous.

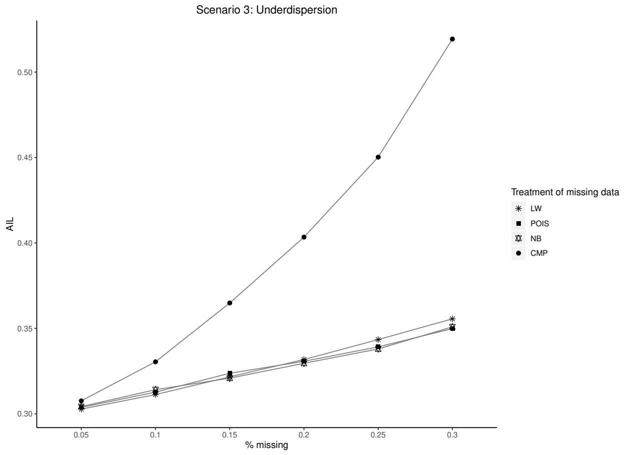

Figure 8: Average length of confidence intervals in the scenario of underdispersion and continuous response variable.

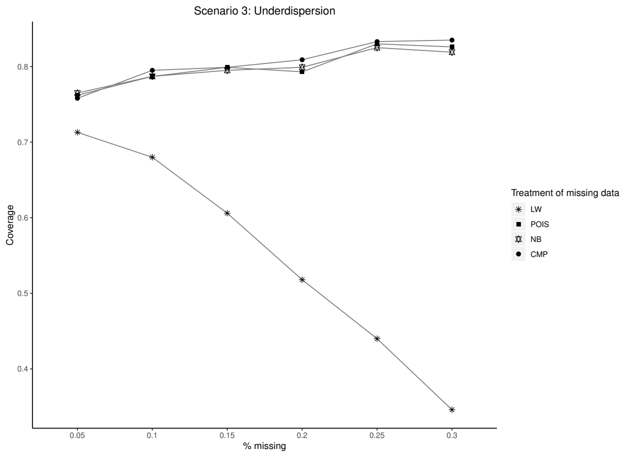

Figure 9: Coverage of confidence intervals in the scenario of underdispersion and continuous response variable.

When the data present, apart from missing values, an excess of zeros, the results show an improved estimation behaviour using the zero-inflated Poisson and Negative binomial models, as is to be expected, although the performance of the imputation methods considered is in general worse than in the other dispersion scenarios, as may be seen in Figure 10. Figure 11 reveals a very uniform behavior regarding the average lengths of the 95% confidence intervals regardless the method used for imputing missing values. When the proportion of missing observations is high, it can be seen (Figure 12) that higher coverages in this scenario are obtained for zero-inflated based imputation methods as expected (both Poisson and negative binomial zero-inflated models), but also with the negative binomial and COMPoisson.

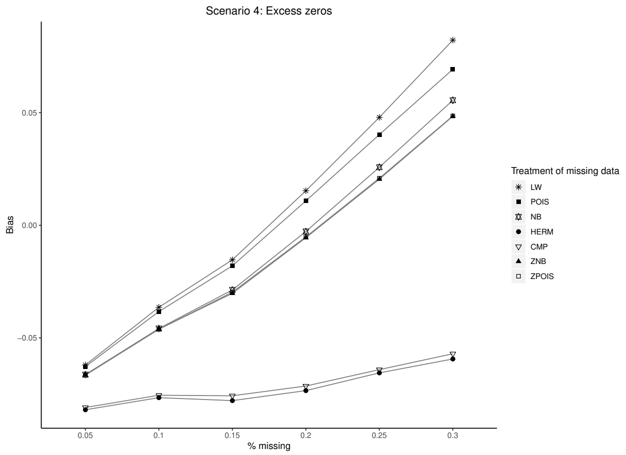

Figure 10: Bias in estimation of the coefficient when there is an excess of zeros and the response variable is continuous.

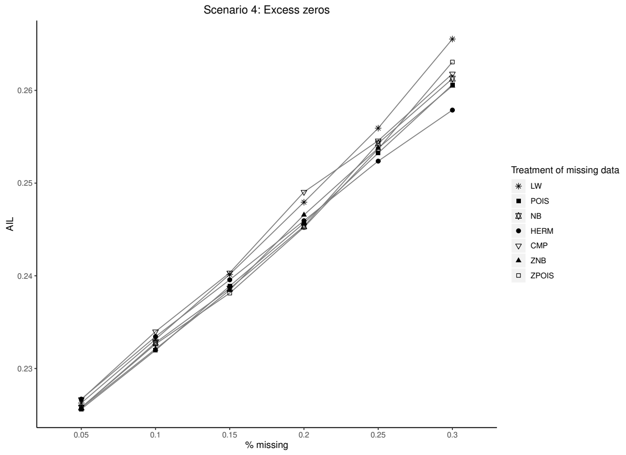

Figure 11: Average length of confidence intervals in the scenario of excess zeros and continuous response variable.

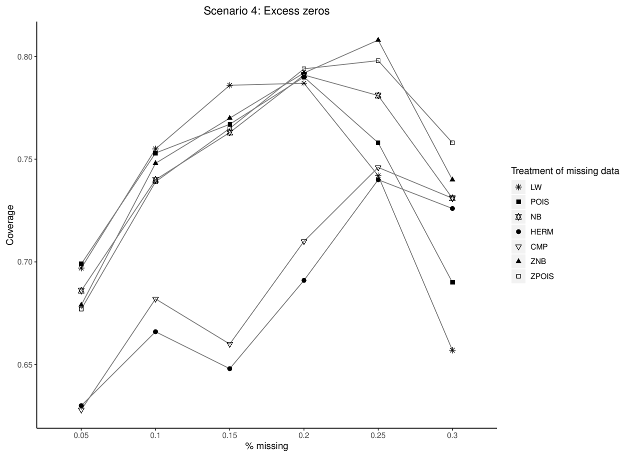

Figure 12: Coverage of confidence intervals when there is an excess of zeros and the response variable is continuous.

4 Discussion

In medical research it is common for data to be partially missing and it is necessary to cope with this problem in order to analyze the information in a coherent and consistent manner. Many methods are available which allow this, and proper handling of information which is lacking will depend on choosing the most appropriate one.

When the variable presenting missing values is a count, there are various alternative ways to impute such missing values, which depend on the distribution characteristics of the count variable, particularly the behaviour of the mean and variance. The most widely used method is to assume that there is equidispersion and that a classical Poisson model is the best alternative to impute the missing counts; however, in real-life research this assumption is not always correct, and it is common to find count variables exhibiting overdispersion or underdispersion, for which the Poisson model is no longer the best to use in imputation. If there is overdispersion the Poisson model underestimates the amount of dispersion. In recent years much work has been done on implementation of other counting models, which may be generalizations of the Poisson model and which take over- and under-dispersion into account, as well as the problem of excess zeros ([28]).

The lung cancer example shows that the COMPoisson distribution provides with a powerful alternative to missing data imputation in a realistic situation of overdispersed discrete explanatory variables, even with a large proportion of missing information, within the framework of multiple imputation, which is now implemented in many standard software and is therefore available to researchers in a straightforward way.

Moreover, in the simulation study we generated different scenarios with a variety of percentages of missing data and three distributions of the response variable: continuous, binary and discrete, and subsequently fitted the respective models with the imputed independent variable using one of the models mentioned. The results of this simulation allow us to affirm that in order to impute a count variable, it is not sufficient to assume the distribution is Poisson; it is necessary to identify the relationship between the variance and the mean of the data, as well as whether the presence of excess zeros might make it appropriate to use specific models for handling missing data in this scenario. Additionally, it can be seen that especially for continuous and discrete response variable when the proportion of missing information is very high (over a 20%), imputing the missing values can lead to inaccurate results regarding relative bias and extremely low coverage rates.

One of the most unexpected findings is that fitting models using only the available data in some cases produces estimator with less bias than performing imputation of the missing values ([7]), in the case of a dichotomic response variable. Although it is recognized that fitting a model with only the available data may affect the power of the study and produce imprecise estimations, it is evident that the effects on power and precision may be significant if the percentage of missings is low and the mechanism of data loss is completely random.

5 Conclusions

In several of the scenarios considered the performance of the methods analyzed differs, something which indicates that it is important to analyze dispersion and the possible presence of excess zeros before deciding on the imputation method to use. Specifically, in the scenario of equidispersion and binary response variable, when a logistic regression model is fitted imputing missing values with the MAR mechanism using the different models mentioned we observe, as expected, that the Poisson and Negative Binomial models produce estimations with low bias and acceptable coverage. Moreover, the COMPoisson model performs well as it is flexible regarding the handling of counts with characteristics of over- and under-dispersion, as well as with equidispersion ([32]). If, however, the count variable is overdispersed, as is often the case in health research, there are various alternatives for performing imputation, the Negative Binomial model being the most recommended. In our results, just as in the case of equidispersion, when the response variable is binary, estimating using available data without performing imputation produces good estimators and with low bias, however it is important to observe that in this case the imputation methods which perform best are those employing zero-inflated models, and the COMPoisson model works very well. For the case where the variable presents underdispersion we observe that imputation based on Poisson and Negative Binomial regression models perform similarly, although the Negative Binomial can present certain difficulties with convergence in this scenario. Furthermore, regard in size of confidence intervals, it is curious that in fitting these models, the higher the percentage of missings the smaller the confidence intervals (apparently an artifact) whereas the COMPoisson is able to maintain their size, and it is worth emphasising that with listwise deletion size increases, as in the other scenarios. When in addition to having missings the data has excess zeros the estimates of the β_1 parameter in the case of binary response are more precise, i.e. they have less bias and greater coverage but, as in the previous scenarios, this result is similar to when no imputation is performed and only the available data is used. For the case of continuous response variable (Table S4) estimates of the β_1 coefficient are always smaller than the value of the parameter and coverage of the confidence intervals is acceptable, coverage being even greater for listwise, although at the expense of some confidence intervals being considerably wider.

According to the results of the simulation obtained in this study, the choice of the best method of imputation for count variables depends on various factors such as the amount of missing data, the behaviour of the expected value in relation to the variance, i.e. whether there is equi-, under-, or over-dispersion and the distribution of the response variable. In particular, if data present an excess of zeros, this represents an additional factor to be taken into account when choosing the missing data imputation method. Although in practice the exact distributional form of the incomplete covariates is unknown because, precisely, of the missing information, the behaviour of many phenomena in the context of public health are well established. For instance, it is well known that the risk of suffering a new sickness leave increases with the number of previous events (see [29] for instance), which would lead to overdispersed data, and the same behavior can be observed with the number of falls suffered by long-term centers residents (as in [23]).

Acknowledgements

David Moriña acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) and from Fundación Santander Universidades.

References

[1] P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10:1100–1120, 1982.

[2] S. Chakraborty and S. H. Ong. A COM-Poisson-type generalization of the negative binomial distribution. *Communications in Statistics - Theory and Methods*, 45(14):4117–4135, jul 2016.

[3] Committee for Medicinal Products for Human Use (CHMP). Guideline on Missing Data in Confirmatory Clinical Trials. Technical report, European Medicines Agency, London, 2010.

[4] R. W. Conway and W. L. Maxwell. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136, 1962.

[5] R. Gupta and G. Jain. A generalized hermite distribution and its properties. *SIAM Journal on Applied Mathematics*, 27(2):359–363, September 1974.

[6] Panteha Hayati Rezvan, Katherine J Lee, and Julie A Simpson. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, 15(1):30, dec 2015.

[7] Gilma Hernández, David Moriña, and Albert Navarro. Imputing missing data in public health general concepts and application to dichotomous variables. *Gaceta Sanitaria*, 31(4):342–345, 2017.

[8] J. Hilbe. *Negative Binomial regression*. New York: Cambridge University Press, 2 edition, 2011.

[9] Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[10] D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 1980.

[11] Adrienne W. Kemp and C. D. Kemp. An alternative derivation of the hermite distribution. *Biometrika*, 53(3-4):627–628, December 1966.

[12] CD Kemp and Adrienne W Kemp. Some properties of the hermite distribution. *Biometrika*, 52(3-4):381–394, 1965.

[13] Michael G. Kenward and James Carpenter. Multiple imputation: current perspectives. *Statistical methods in medical research*, 16(3):199–218, jun 2007.

[14] Soeun Kim, Catherine A Sugar, and Thomas R Belin. Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in medicine*, 34(11):1876–88, may 2015.

[15] Kristian Kleinke, Roel de Jong, Martin Spiess, and Jost Reinecke. Multiple imputation of incomplete ordinary and overdispersed count data. *Bielefeld University, Faculty of Sociology and Centre for Statistics*, 1, 2011.

[16] Lawrence R. Landerman, Kenneth C. Land, and Carl F. Pieper. An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. *Sociological Methods & Research*, 26(1):3–33, aug 1997.

[17] J. D. Lewsey and W. M. Thomson. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of

cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology*, 32(3):183–189, jun 2004.

[18] Roderick JA Little. Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[19] T. Martin Lukusa, Shen-Ming Lee, and Chin-Shang Li. Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. *Metrika*, 79(4):457–483, may 2016.

[20] P (Peter) McCullagh and John A Nelder. *Generalized linear models*. Chapman and Hall, 1989.

[21] Geert Molenberghs and Els Goetghebeur. Simple fitting algorithms for incomplete categorical data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):401–414, 1997.

[22] David Moriña, Manuel Higueras, Pedro Puig, and María Oliveira. Generalized Hermite distribution modelling with the R package hermite. *R Journal*, 7(2):263 – 274, 2015.

[23] Albert Navarro and Iciar Ancizu. Analyzing the occurrence of falls and its risk factors: Some considerations. *Preventive Medicine*, 48(3):298–302, mar 2009.

[24] Bhavna T. Pahel, John S. Preisser, Sally C. Stearns, and R. Gary Rozier. Multiple imputation of dental caries data using a zero-inflated Poisson regression model. *Journal of Public Health Dentistry*, 71(1):71–78, jan 2011.

[25] Therese D Pigott. A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383, 2001.

[26] Peterson A. V. Prentice R. L., Williams B. J. On the regression analysis of multivariate failure time data. *Biometrika*, 68:373–379, 1981.

[27] Pedro Puig. Characterizing Additively Closed Discrete Models by a Property of Their Maximum Likelihood Estimators, With an Application to Generalized Hermite Distributions. *Journal of the American Statistical Association*, 98(463):687–692, 2003.

[28] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

[29] Ricardo J. Reis, Mireia Utzet, Poliana F. La Rocca, Fúlvio B. Nedel, Miguel Martín, and Albert Navarro. Previous sick leaves as predictor of subsequent ones. *International Archives of Occupational and Environmental Health*, 84(5):491–499, jun 2011.

[30] Martin S Ridout and Panagiotis Besbeas. An empirical model for underdispersed count data. *Statistical Modelling*, 4(1):77–89, 2004.

[31] D.B. Rubin. *Multiple Imputation for nonresponse in Surveys.* John Wiley & Sons, Inc., 1987.

[32] Kimberly F Sellers, Sharad Borle, and Galit Shmueli. The com-poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28(2):104–116, 2012.

[33] Galit Shmueli, Thomas P Minka, Joseph B Kadane, Sharad Borle, and Peter Boatwright. A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005.

[34] Bao Luo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric J. Tchetgen. Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable. *Statistica Sinica*, 28(4):1965, oct 2018.