An analytical approximate causal inference for integer-valued outcomes

Young Lee Marie-Abéle Bind

January 8, 2020

Abstract

The literature for count modeling provide useful tools to conduct causal inference when outcomes take non-negative integer values. Applied to the potential outcomes framework and utilizing the Bayesian approximation, we link aspects of the causal inference literature with some statistical methods for count data. Special considerations for estimating treatment effects are discussed, some generalizing certain relationships and some not hitertho encountered in the causal inference literature.

1 Introduction

Outcomes with non-negative integer values commonly occur across many fields in social science. Whilst the Bayesian approach to model-based inference for causal effects is well developed for continuous and binary outcomes, there has been comparatively little work on outcomes with non-negative and integer-valued support. These outcomes often arise cohort studies, in which participants regularly visit a medical center and the counts of health-related incidence (e.g., number of panic attacks, number of depressive episodes) between visits are recorded (Thall and Lachin, 1988; Sun and Zhao, 2016). Count outcomes are also common in randomized studies of certain treatment. For example, subjects may be queried about their daily consumption of alcohol, measured as a number of drinks over a recent period (Horton et al., 2007) where estimating differences between treatment and control groups could be of interest.

This manuscript is concerned with the estimation of causal effects when the *potential outcomes* take the form of non-negative integer values $0, 1, 2, 3, \ldots$, or *count* data. Potential outcomes are characteristics of a unit (e.g., an individual) before that unit is assigned to a treatment condition and typically depend on some covariates. For example the potential outcomes may represent the daily count of deaths by suicide after a wildfire episode and the daily count of deaths by suicide had the wildfire not occurred. However, we immediately notice a fundamental difficulty, in that, we cannot simultaneously observe both values.

One approach to statistical inference is multiple imputation (Rubin, 1987; Reiter and Raghunathan, 2007; Zhou and Reiter, 2010) where the idea is to fill in any missing values by repeatedly sampling from the predictive distributions of the missing values. With this, we can provide statistical summaries as risk differences, which are optimal for risk assessors and public understanding.

Most causal estimands derived in the literature implicitly assume continuity of the potential outcomes (Chapter 8 in Rubin (2005)). On the other hand, the causal estimand when the potential outcomes take binary data (ibid., p. 176-177) is also well studied (Gutman and Rubin, 2012, 2015). On the other spectrum, Hill (2011) proposed a nonparametric method by modulating the potential outcomes with Bayesian Additive Regression Trees (BART). A different thread of research is launched in Hollenbach et al. (2018) that uses Gaussian copulas to model the joint potential outcomes. These approaches also assume continuity of the outcomes. Very recently, a new Fisher's randomization inference methods for count data that has an excess of zeros is being proposed (Keele and Miratrix, 2019).

Various statistical models have been developed to handle count data such and among some of the contributions to count regression models, El-Sayyad (1973); Lawless (1987); Diggle et al. (1998); Winkelmann (2008); Chan and Vasconcelos (2012); Kim et al. (2013) consider the case of Poisson regression and negative Binomial regression. On the other hand, (Breslow, 1984; Agresti, 2007) study the Lognormal-Poisson regression and their estimation procedures. Leveraging on their ideas, we link the causal inference literature with statistical methods for count data. Our proposed potential outcomes framework accounts for overdispersion; hence we extend some of their analyses in the formulation and estimation of causal estimands.

The main contribution of this paper is to design an approximation framework to determine the causal effects when the potential outcomes take non-negative integer values. In a Bayesian context, we discuss the general architectural considerations for constructing the conditional distribution of the missing potential outcomes given the observed data. Conceptually, determining the causal effects in this setting involve the following several key steps: (i) posit a suitable family of potential outcomes, then the conditional distribution of the missing potential outcomes given the observed data is evaluated; (ii) compute the posterior distribution of the parameters of the potential outcomes. In this step, we extend the Poisson regression results in El-Sayyad (1973) to handle overdispersion in our potential outcomes framework. This yields an *approximate* posterior distribution due to the non-conjugacy of Poissonnian likelihood and our Normal priors. We further characterize its speed of convergence formalizing the approximation of Bartlett and Kendall (1946), which forms the results in El-Sayyad (1973). In step (iii), we evaluate the conditional distribution of the missing data given the observed data and finally (iv) the estimand of causal interest is derived.

The paper is structured in the following manner. In Section 2 we introduce our non-negative potential outcomes framework. In Section 3, we develop a

Bayesian imputation model for the missing potential outcomes of count type. Of particular interest is the characterization of approximation that is used to compute the posterior distribution in step (ii) towards the evaluation of causal estimands. Section 4 concludes.

2 An integer-valued potential outcomes framework

The concept of potential outcomes was launched in Splawa-Neyman et al. (1990) for randomization-based inference in randomized experiments and later used by other researchers including Kempthorne (1955) and Wilk (1955) for causal inference from randomized experiments. The concept was extended by Rubin (confer Rubin (1974, 1975, 1976, 1977, 1978)) to other forms of causal inference from randomized experiments to observations studies.

In the context of causal inference, we consider the causal effect of binary treatment W, with $W = 1$ indicating assignment to treatment and $W = 0$ indicating assigment to control. Following convention in the literature (confer e.g., Rubin (1978)), the causal effect for individual i, with $i = 1, \ldots, N$ is defined as a comparison of potential outcomes, $Y_i(1)$ and $Y_i(0)$, where these are the outcomes that would be observed under $W = 1$ and $W = 0$, respectively. This may be recast as $Y_i = Y_i(1) \cdot W_i + Y_i(0) \cdot (1 - W_i)$. We can never observe both $Y_i(1)$ and $Y_i(0)$ for any unit i, because it is not possible to go back in time and expose the ith unit to the other treatment. This is called the 'fundamental problem of causal inference' (Holland, 1986). Put differently, we are implicitly trying to figure out what would have happened to an individual if they had taken a different path. Therefore, unit level causal effects cannot be known and must be inferred. The framework permits prediction of unit level causal effects from either the Neymanian perspective (Splawa-Neyman et al., 1990) or from the Bayesian perspective (Rubin, 1978), although such estimations are generally imprecise relative to the estimation of poulation or subpopulation causal effects. Due to this, researchers focus on drawing inferences on the average treatment effect defined either over the sample or the population, for example the conditional average treatement effect (ATE):

$$\tau_{\text{fs}} := \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)).$$

An integer-valued potential outcomes. Let $Y_i(z)$, $z \in \{0, 1\}$ denote the potential outcome of the ith unit if exponsed to treatment z. When introducing this notation, we tacitly make the stable unit treatement value assumption, which means that the potential outcome of a particular unit depends only on the treatment combination it is assigned, rather than on the assignments of the remaining units. Also, we postulate that there are no hidden versions of the

treatments not represented by the values of z, i.e., under this assumption, the potential outcome Y_i of each unit i in an experiment only depends on whether it receives the treatment ($W = 1$) or not ($W = 0$). The potential outcomes of all N units in an experiment, can be partitioned into two vectors of N components: $\mathbf{Y}(0)$ for all outcomes under control and $\mathbf{Y}(1)$ for all outcomes under treatment. The potential outcomes can also be partitioned according to whether it is observed whereby a unit is either under control or under treatment. Therefore, half of all the potential outcomes are observed, denoted as $\mathbf{Y}^{\text{obs}}$. The other half of the potential outcomes are unobserved, denoted as $\mathbf{Y}^{\text{mis}}$.

We propose a count model that is able account for overdispersion, wherein the variance is greater than the mean:

$$Y_i(0) \mid \beta^{[c]}, \epsilon_i^{[c]} \sim \mathsf{Po}(\mu_i^{[c]}), \quad Y_i(1) \mid \beta^{[t]}, \epsilon_i^{[t]} \sim \mathsf{Po}(\mu_i^{[t]}) \tag{1}$$

with $\mu_i^{[c]} := \exp(\mathbf{x}_i^\top \beta^{[c]})\epsilon_i^{[c]}$ and similarly $\mu_i^{[t]} := \exp(\mathbf{x}_i^\top \beta^{[t]})\epsilon_i^{[t]}$ where $\epsilon_i^{[\cdot]}$ is taken to be a non-negative multiplicative random-effect term to model individual heterogeneity (Long, 1997; Winkelmann, 2008; Cameron and Trivedi, 2013). The covariates $\mathbf{x}_i$ are the $(k+1)$-dimensional features for every i and $\beta^{[c]}, \beta^{[t]} \in \mathbb{R}^{k+1}$. We further let

$$\mathbb{R}^{2(k+1)} \ni \boldsymbol{\beta} := \begin{pmatrix} \beta^{[c]} \\ \beta^{[t]} \end{pmatrix} \sim \mathsf{N}\left(\mathbf{0}_{2(k+1)}, \sigma_\beta^2 \cdot \mathbf{I}_{2(k+1)}\right) \tag{2}$$

with σ_β^2 being a fixed positive number. We allow for $\beta^{[c]} \neq \beta^{[t]}$ to account for the possiblity that the corresponding parameters are different for treatment and control groups. The Negative Binomial regression model is constructed by placing a Gamma prior on $\epsilon_i^{[\cdot]}$ (Lawless, 1987; Hilbe, 2007). On the other hand, a Lognormal-Poisson regression model can be constructed by taking a Lognormal prior on $\epsilon_i^{[\cdot]}$ (Breslow, 1984; Agresti, 2007). In contrast to the Negative Binomial model, there is no closed-form solution for the distribution of $Y_i(\cdot)$ when $\epsilon_i^{[\cdot]}$ is integrated out thus making it a less popular choice in the literature. However, as pointed out in Winkelmann (2008), this is an appealing model that may be a good fit in practice. Another possible avenue is to attribute an inverse Gaussian prior on $\epsilon_i^{[\cdot]}$ which would result in a heavy-tailed count behavior (Dean et al., 1989).

The assignment mechanism. Drawing inference for causal effects requires the specification of an assignment mechanism, which is a probabilistic model for how experimental units are allocated to the treatment combination. Let the $N-$component column vector of treatment assignments denoted by $\mathbf{W}$, with ith element $W_i \in \{0, 1\}$, with $W_i = 0$ if unit i received the control treantment and $W_i = 1$ if this unit received the active treatment. We define $N_c := \sum_{i=1}^N (1 - W_i)$ and $N_t = \sum_{i=1}^N W_i$ be the number of units assigned to the control and active treatment respectively, with $N_c + N_t = N$. We assert that the conditional distribution of $\mathbf{W}$ given the potential outcomes is independent of covariates $\mathbf{x}_i$,

so by definition, the assignment mechanism is equal to

$$P(\mathbf{W} = \mathbf{w} \,|\, \mathbf{Y}(0), \mathbf{Y}(1), \beta^{[c]}, \beta^{[t]}, \epsilon_i^{[c]}, \epsilon_i^{[t]}) = \binom{N}{N_t}^{-1} \tag{3}$$

for all $\mathbf{W}$ such that $\sum_{i=1}^{N} W_i = N_t$, and null otherwise.

3 Bayesian count models for potential outcomes

We work within the model based Bayesian framework for causal inference (Rubin, 1975, 1978) and evaluate a posterior predictive distribution for at least half of the missing potential outcomes. The fundamental idea is to initiate an imputation model for the missing potential outcomes $\mathbf{Y}^{\text{mis}}$, conditionally on the observed outcomes $\mathbf{Y}^{\text{obs}}$ as well as the observed assignment vector $\mathbf{W}$. Because the treatment effect is a function of the vector $\mathbf{Y}$ of all potential outcomes, this can be partitioned as $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}})$, from which the conditional posterior distribution of the weights $\boldsymbol{\beta}$ can be determined. To this end, we let $f(\mathbf{Y} \,|\, \boldsymbol{\beta})$ represent a suitable probabilistic model for $\mathbf{Y}$, where $\boldsymbol{\beta}$ is a vector of parameters with a suitable prior distribution placed on them. Posterior inference for the causal estimand entails the following steps:

- (i) Evaluate $\mathbf{Y}^{\text{mis}} \,|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\beta}, \mathbf{W}, \boldsymbol{\epsilon}$, the conditional posterior distribution of $\mathbf{Y}^{\text{mis}}$ given $\mathbf{Y}^{\text{obs}}, \boldsymbol{\beta}$ and $\mathbf{W}$.
- (ii) Evaluate $\boldsymbol{\beta} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon}$, the posterior distribution for the parameters $\boldsymbol{\beta}$ conditional on the observed data $\mathbf{Y}^{\text{obs}}$, the observed assignemnt mechanism $\mathbf{W}$ and $\boldsymbol{\epsilon}$.
- (iii) Proceed by obtaining the imputation model $\mathbf{Y}^{\text{mis}} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}$ by marginalizing the posterior predictive distribution in step (ii) over the parameters ϑ and $\boldsymbol{\epsilon}$.
- (iv) Finally, obtain the posterior distribution for the causal estimand.

The following expression of $(Y_i^{\text{mis}}, Y_i^{\text{obs}})$ as functions of $(Y_i(0), Y_i(1), W_i)$ holds true:

$$Y_i^{\text{mis}} = Y_i(1)\mathbb{1}_{\{W_i=0\}} + Y_i(0)\mathbb{1}_{\{W_i=1\}} = (1 - W_i) \cdot Y_i(1) + W_i \cdot Y_i(0).$$

Before we proceed to evaluate the following aforementioned steps, we discuss the Bayesian approximation which would be used in step (ii) in our framework owing to the non-conjugacy of Normal priors in equation (2) relative to Poisson likelihood due to our model specifictions in equation (1).

3.1 Divergence and convergence

We state two important results due to Bartlett and Kendall (1946) and El-Sayyad (1973) where these results would be extended and formalized to aid

the posterior approximation in step (ii). First, we note that if $\tilde{X}$ is a Gamma random variable $\mathsf{Ga}(r, s)$ with density function

$$P(\tilde{X} \in dx) = (\Gamma(r)s^r)^{-1}x^{r-1}\exp\left(-\frac{x}{s}\right), \quad x \geq 0, \tag{4}$$

then a transformation of $\tilde{Y} := \log \tilde{X}$ yields the log-Gamma distribution for $\tilde{Y}$ which has density

$$P(\tilde{Y} \in dy) = (\Gamma(r)s^r)^{-1}\exp\left(ry - \frac{e^y}{s}\right), \quad \forall y > 0. \tag{5}$$

Let $\tilde{X}$ and $\tilde{Y}$ as defined in equations (4) and (5), respectively. Then the transformation of $\tilde{Y} := \log \tilde{X}$ is *approximately* Normal distributed with mean $\log rs$ and variance $\log r^{-1}$ for large r (Bartlett and Kendall, 1946). Leveraging on this result, El-Sayyad (1973) noted the variables and parameters of the log-Gamma and Poisson distributions are related in the following sense:

$$\begin{aligned} f_{\text{Poisson}}(y \,|\, \mu \equiv e^{\xi}) = \frac{e^{-\mu}\mu^y}{y!} = \frac{1}{y} \cdot \frac{e^{-e^{\xi}}(e^{\xi})^y}{\Gamma(y)} = \frac{1}{y} \cdot f_{\text{logGamma}}(\xi \,|\, y, 1) \\ \overset{(\dagger)}{\approx} \frac{1}{y} \cdot f_{\text{Normal}}(\xi \mid \log y, y^{-1}) \end{aligned} \tag{6}$$

where the approximation $(\dagger)$ is taken in the sense of Bartlett and Kendall (1946).

We now investigate the speed of convergence between the log-Gamma and Poisson distributions in the sense of Bartlett and Kendall (1946). The Kullback-Leibler divergence between $\bar{f}$ and $\underline{g}$, denoted by $\mathbb{D}_{\text{KL}}(\bar{f} \,\|\, \underline{g})$ is given by $\int \bar{f}(x)\log(\frac{\bar{f}(x)}{\underline{g}(x)})dx$. Direct calculation shows that

$$\begin{aligned} &\mathbb{D}(f_{\text{normal}}(\xi \mid \log y, y^{-1}) \,\|\, f_{\text{log-Gamma}}(\xi \,|\, y, 1)) \\ &\qquad = \log\left(\frac{\Gamma(y)}{\sqrt{2\pi y^{-1}}}\right) - y\log y + ye^{\frac{1}{2y}} - \frac{1}{2} \end{aligned} \tag{7}$$

where the computations reduce to evaluating the mean, variance and moment generating function at point 1 of the normal distribution, respectively. Note that $\mathbb{D}(f_{\text{normal}}(\xi \mid \log y, y^{-1}) \,\|\, f_{\text{log-Gamma}}) \to 0$ when $y \to \infty$. Stirling's formula (Andrews et al., 1999) duly extends to the gamma function

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}}\left(\frac{z}{e}\right)^z\left(1 + \mathcal{O}\left(\frac{1}{z}\right)\right). \tag{8}$$

We present the following:

Lemma 1. It holds true that

$$\mathbb{D}(f_{\text{normal}}(\xi \mid \log y, y^{-1}) \,\|\, f_{\text{log-Gamma}}(\xi \,|\, y, 1)) = \frac{5}{24y} + \mathcal{O}\left(\frac{1}{y^2}\right). \tag{9}$$

Proof. The Gamma function in (7) is expanded through Stirling's series (Uhler, 1942; Arfken, 1985) for its associated formula in (8). Then using the fact that $y(e^{\frac{1}{2y}} - 1) \to 1/2$ when $y \to \infty$ yields the result.

3.2 An imputation model

Since Y_i^{mis} is a linear combination of $Y_i(1)$ and $Y_i(0)$, we have the following:

Lemma 2. The conditional distribution of Y_i^{mis} given $Y_i^{\text{obs}}, \beta^{[c]}, \beta^{[t]}, \epsilon_i^{[c]}, \epsilon_i^{[t]}$ is given by the following:

$$Y_i^{\text{mis}} \,|\, Y_i^{\text{obs}}, W_i, \vartheta = \mathsf{Po}(\mu_i^{\text{mis}})$$

where $\mu_i^{\text{mis}} := \exp(\xi_i^{\text{mis}})$, $\vartheta := (\beta^{[c]}, \beta^{[t]}, \epsilon_i^{[c]}, \epsilon_i^{[t]})$ and

$$\xi_i^{\text{mis}} := \Big((1 - W_i) \cdot \{\mathbf{x}_i^\top \beta^{[t]} + \log \epsilon_i^{[t]}\} + W_i \cdot \{\mathbf{x}_i^\top \beta^{[c]} + \log \epsilon_i^{[c]}\}\Big). \tag{10}$$

Remark 1. In our model setup, the conditional distribution of Y_i^{mis} given $(Y_i^{\text{obs}}, \vartheta)$ is simply equal to the marginal distribution Y_i^{mis} given ϑ.

Remark 2. Due to the binary nature of W_i, we have the following equivalent expressions:

$$\begin{aligned} &W_i \cdot e^{\mathbf{x}_i^\top \beta^{[c]}} \epsilon_i^{[c]} + (1 - W_i) \cdot e^{\mathbf{x}_i^\top \beta^{[t]}} \epsilon_i^{[t]} \\ =& \exp\Big((1 - W_i) \cdot \{\mathbf{x}_i^\top \beta^{[t]} + \log \epsilon_i^{[t]}\} + W_i \cdot \{\mathbf{x}_i^\top \beta^{[c]} + \log \epsilon_i^{[c]}\}\Big). \end{aligned}$$

We re-express the following: $Y_i^{\text{obs}} = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1)$. Moreover, $Y_i^{\text{obs}} \,|\, \vartheta$ is Poisson with parameter μ_i^{obs} where $\mu_i^{\text{obs}} := \exp(\xi_i^{\text{obs}})$ with $\xi_i^{\text{obs}} = (1 - W_i)\{\mathbf{x}_i^\top \beta^{[c]} + \log \epsilon_i^{[c]}\} + W_i \cdot \{\mathbf{x}_i^\top \beta^{[t]} + \log \epsilon_i^{[t]}\}$. Furthermore by defining $m_i := \begin{bmatrix}(1 - W_i) & W_i\end{bmatrix} \tilde{\epsilon}_i$, we can re-express ξ_i^{obs} succinctly as follows

$$\xi_i^{\text{obs}} = \tilde{\mathbf{x}}_i \boldsymbol{\beta} + m_i. \tag{11}$$

Due to the choice of our model setup, an analytical expression for the posterior for $\boldsymbol{\beta}$ is not available owing to the lack of conjugacy between the Normal priors and the Poissonnian likelihood. However, we now give a result which concern the approximate posterior distribution of $\boldsymbol{\beta}$ conditional on $\mathbf{Y}^{\text{obs}}$ and its assignment mechanism $\mathbf{W}$:

Lemma 3. Define $\tilde{\mathbf{x}}_i := \begin{bmatrix}(1 - W_i) & W_i\end{bmatrix} \otimes \mathbf{x}_i^\top$, $\tilde{\mathbf{X}} := (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_N)^\top$, $\tilde{\mathbf{Y}} := (\log Y_1^{\text{obs}}), \ldots, \log Y_N^{\text{obs}}))^\top$, $\tilde{\epsilon}_i := (\log \epsilon_i^{[c]}, \log \epsilon_i^{[t]})^\top$, $\tilde{\boldsymbol{\epsilon}} := (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_N)^\top$, $\mathbf{M}^{\boldsymbol{\epsilon}} = (m_1, m_2, \ldots, m_N)^\top$ and $\tilde{\mathbf{R}}^{\boldsymbol{\epsilon}} := \tilde{\mathbf{Y}} - \mathbf{M}^{\boldsymbol{\epsilon}}$. Then

$$\boldsymbol{\beta} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon} \sim \mathsf{N}(\mu_{\boldsymbol{\beta}}^{\boldsymbol{\epsilon}}, \Sigma_{\boldsymbol{\beta}}) \tag{12}$$

where $\Sigma_y^{\text{obs}} = \text{diag}\left[(Y_1^{\text{obs}})^{-1}, \ldots, (Y_N^{\text{obs}})^{-1}\right]$, $\Sigma_{\boldsymbol{\beta}}^{-1} = \tilde{\mathbf{X}}^\top (\Sigma_y^{\text{obs}})^{-1} \tilde{\mathbf{X}} + (\sigma_\beta^2)^{-1} \mathbf{I}_{2(k+1)}$ and $\mu_{\boldsymbol{\beta}}^{\boldsymbol{\epsilon}} = \Sigma_{\boldsymbol{\beta}} \tilde{\mathbf{X}} \, (\Sigma_y^{\text{obs}})^{-1} \;\; \tilde{\mathbf{R}}^{\boldsymbol{\epsilon}}$.

Remark 3. Trivial bookkeeping: $\tilde{\mathbf{x}}_i^\top \in \mathbb{R}^{2(k+1)}$ and $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times 2(k+1)}$ so that $\tilde{\mathbf{X}}\boldsymbol{\beta} \in \mathbb{R}^{N \times 1}$. Since $\tilde{\epsilon}_i \in \mathbb{R}^{2 \times 1}$, we also have $m_i \in \mathbb{R}$ and $\mathbf{M}^{\boldsymbol{\epsilon}} \in \mathbb{R}^{N \times 1}$.

Proof. We combine the prior distribution in equation (2) with the distribution of the observed data given $\boldsymbol{\beta}$ to compute the posterior distribution of $\boldsymbol{\beta}$. With the approximation in Lemma 1, computing the product of priors and likelihood yield the result. ∎

We now proceed to determine the distribution of μ_i^{mis}, conditional on $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon})$. Some additional notation is warranted. Let $\breve{\mathbf{x}}_i := \begin{bmatrix} W_i & (1-W_i) \end{bmatrix} \otimes \mathbf{x}_i^\top$ and $\breve{m}_i := \begin{bmatrix} W_i & (1-W_i) \end{bmatrix} \tilde{\epsilon}_i$. Hence we can be neatly recast the following $\xi_i^{\text{mis}} = \breve{\mathbf{x}}_i \boldsymbol{\beta} + \breve{m}_i$. Then, from Lemma 3, we establish that $\xi_i^{\text{mis}} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon}$ is also normally distributed with mean $\breve{\mathbf{x}}_i \mu_{\boldsymbol{\beta}}^{\boldsymbol{\epsilon}} + \breve{m}_i$ and variance $\breve{\mathbf{x}}_i^\top \Sigma_{\boldsymbol{\beta}} \breve{\mathbf{x}}_i$. Hence $\xi_i^{\text{mis}} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon}$ is approximately log-Gamma from Lemma 1 with parameters $(\breve{\mathbf{x}}_i^\top \Sigma_{\boldsymbol{\beta}} \breve{\mathbf{x}}_i)^{-1}$ and $\breve{\mathbf{x}}_i^\top \Sigma_{\boldsymbol{\beta}} \breve{\mathbf{x}}_i \exp(\breve{\mathbf{x}}_i \mu_{\boldsymbol{\beta}}^{\boldsymbol{\epsilon}} + \breve{m}_i)$. Hence we see that the the conditional distribution of μ_i^{mis} given $\mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon}$ is approximately Gamma $\mathsf{Ga}(\cdot, \cdot)$ with the following parameters

$$\mu_i^{\text{mis}} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon} \sim \mathsf{Ga}(h_i^-, h_i^+) \tag{13}$$

where

$$h_i^- := (\breve{\mathbf{x}}_i^\top \Sigma_{\boldsymbol{\beta}} \breve{\mathbf{x}}_i)^{-1}, \qquad h_i^+ := \breve{\mathbf{x}}_i^\top \Sigma_{\boldsymbol{\beta}} \breve{\mathbf{x}}_i \exp(\breve{\mathbf{x}}_i \mu_{\boldsymbol{\beta}}^{\boldsymbol{\epsilon}} + \breve{m}_i). \tag{14}$$

Now we proceed to compute the conditional distribution of the missing data given the observed data, but *neither* conditioning on the parameters *nor* on ϵ_i, i.e. $Y_i^{\text{mis}} \,|\mathbf{Y}^{\text{obs}}, \mathbf{W}$. This is done by amalgamating the conditional distribution of Y_i^{mis} given $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_i^{\text{mis}}, \epsilon_i)$ in Lemma 2 and using the results of Lemma 3 and equations (13) and (14) by integrating over μ_i^{mis} and ϵ_i to obtain the following main result:

Theorem 1. The conditional density of $Y_i^{\text{mis}} = y_i$ given $\mathbf{Y}^{\text{obs}}, \mathbf{W}$ can be expressed as:

$$P(Y_i^{\text{mis}} = y \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}) = \int \psi_\epsilon(y) f(\epsilon_i) d\epsilon_i$$

where

$$\psi_\epsilon(y) = \binom{h^- + y - 1}{y} \left(\frac{1}{1 + h^+(\boldsymbol{\epsilon})} \right)^{h^-} \left(1 - \frac{1}{1 + h^+(\boldsymbol{\epsilon})} \right)^{y} \tag{15}$$

whose values of h^- and $h^+(\boldsymbol{\epsilon})$ are given by are given in equation (14) and $f(\epsilon_i)$ denotes the density of the non-negative multiplicative random-effect term ϵ_i.

Proof. First note that we may write

$$P(Y_i^{\text{mis}} = y \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}) = \int \int P(Y_i^{\text{mis}} = y, \mu_i^{\text{obs}}, \epsilon_i \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}) d\mu_i^{\text{obs}} d\epsilon_i.$$

Then computing the inner integral

$$\psi_\epsilon(y) := \int P(Y_i^{\text{mis}} = y \,|\, \mu_i^{\text{obs}}, \epsilon_i, \mathbf{Y}^{\text{obs}}, \mathbf{W}) P(\mu_i^{\text{obs}} \,|\, \epsilon_i, \mathbf{Y}^{\text{obs}}, \mathbf{W}) d\mu_i^{\text{obs}} \tag{16}$$

yields equation (15). All that is left is to integrate over equation (16) with respect to ϵ_i for which the assertion follows. ∎

Corollary 1. Let the quantity ν_i^+ be defined as $\nu_i^+ := \int h_i^+(\epsilon_i)P(\epsilon_i)d\epsilon_i$. Then the conditional mean and its variance are readily computed as follows:

$$\mathbb{E}[\tau_{\text{fs}} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}] = \frac{1}{N}\sum_{i=1}^{N}\left((2W_i - 1)Y_i^{\text{obs}} + (1 - 2W_i)h_i^-\nu_i^+\right),$$

$$\mathbb{V}\text{ar}[\tau_{\text{fs}} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}] = \frac{1}{N^2}\sum_{i=1}^{N} h_i^- \frac{(1 - \nu_i^+)}{(\nu_i^+)^2}.$$

4 Concluding remarks

As noted in Rubin (1975), the Bayesian approach of imputation is a natural extension of the Fisherian approach where it extends the single-imputation strategy for inference to a strategy based on multiple imputation of missing potential outcomes. One of the advantages of the Bayesian methodology is that it permits us to easily encode prior information on $\boldsymbol{\beta}$.

Implications. It is envisaged that these approximation results may be useful in other branches of causal analysis that deals with count data. We remark that, it is possible to use the Markov chain Monte Carlo (MCMC) to produce (almost) exact samples from the posterior distribution of $\boldsymbol{\beta} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \boldsymbol{\epsilon}$ in step (ii). However, the posterior distribution would likely not be of recognizable form and would not scale well with data. The availability of other count models in the statistics literature and the approximation presented here leaves us the modeling freedom to adapt to each real-life application. For instance, it would be of interest to investigate the identification results for randomized experiments suffering from noncompliance when the potential outcomes take non-negative integer values. Most, if not all the analyses were done with binary outcomes (Zhang and Rubin, 2003; Imai, 2008; Lee, 2009; Mealli and Pacini, 2013)). Also of interest is the investigation into multiple outcomes in principal stratification when the potential outcomes are of count data. This is in contrast to the Normality assumption placed on the potential outcomes (p.2345 in Mattei et al. (2013)).

Correlation. It would be possible to introduce a model that has two contemporaneously correlated count variables (Holgate, 1964; King, 1989). However, in practice, the data generally do not contain useful information to the researcher to draw inferences on the correlation coefficient between the two potential outcomes. Hence a prudent measure is usually placed on the correlation coefficient modeling the two potential outcomes. It is possible that the researcher often wish to avoid modeling the correlation coefficient thus choosing to model the two potential outcomes distributions as conditionally independent in an approach that is prudent in another sense.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley-Blackwell.

Andrews, G. E., Askey, R. A., and Roy, R. (1999). *Special functions*. Cambridge Univ. Press.

Arfken, G. (1985). *Stirling Series, 10.3. Mathematical Methods for Physicists*. Academic Press, Inc., third edition.

Bartlett, M. S. and Kendall, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, 8(1):128–138.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Journal of the Royal Statistical Society: Series C*, 33(1):38–44.

Cameron, A. and Trivedi, P. (2013). *Regression Analysis of Count Data*. Cambridge University Press.

Chan, A. B. and Vasconcelos, N. (2012). Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21:2160–2177.

Dean, C., Lawless, J. F., and Willmot, G. E. (1989). A mixed poisson-inverse-gaussian regression model. *Canadian Journal of Statistics*, 17(2):171–181.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C*, 47(3):299–350.

El-Sayyad, G. (1973). Bayesian and classical analysis of poisson regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 445–451.

Gutman, R. and Rubin, D. (2012). Analyses that inform policy decisions. *Biometrics*, 68(3):671–675.

Gutman, R. and Rubin, D. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in Medicine*, 34.

Hilbe, J. (2007). *Negative Binomial Regression*. Cambridge University Press.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Holgate, P. (1964). Estimation for the bivariate poisson distribution. *Biometrika*, 51(1-2):241–287.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hollenbach, F. M., Bojinov, I., Minhas, S., Metternich, N. W., Ward, M. D., and Volfovsky, A. (2018). Multiple imputation using gaussian copulas. *Sociological Methods & Research*.

Horton, N., Kim, E., and Saitz, R. (2007). A cautionary note regarding count models of alcohol consumption in randomized controlled trials. *BMC medical research methodology*, 7:9.

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statistics & probability letters*, 78(2):144–149.

Keele, L. and Miratrix, L. (2019). Randomization inference for outcomes with clumping at zero. *The American Statistician*, 73(2):141–150.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967.

Kim, S., Chen, Z., Zhang, Z., Simons-Morton, B. G., and Albert, P. S. (2013). Bayesian hierarchical poisson regression models: An application to a driving study with kinematic events. *Journal of the American Statistical Association*, 108(502):494–503.

King, G. (1989). A seemingly unrelated poisson regression model. *Sociological Methods and Research*, 17:235–255.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, 76(3):1071–1102.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Sage Publications.

Mattei, A., Li, F., and Mealli, F. (2013). Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, 7(4):2336–2360.

Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108(503):1120–1131.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *ASA Proceedings of the Social Statistics Section*, pages 233–239. American Statistical Association.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–590.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate (corr: V3 p384). *J. Educ. Behav. Statist.*, 2:1–26.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*, 6:34–58.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

Sun, J. and Zhao, X. (2016). *Statistical Analysis of Panel Count Data*. Springer.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, 83(402):339–347.

Uhler, H. S. (1942). The coefficients of stirling's series for loggamma(z). *Proceedings of the National Academy of Sciences*, 28(2):59–62.

Wilk, M. B. (1955). The randomization analysis of a generalized randomized block design. *Biometrika*, 42(1/2):70–79.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhou, X. and Reiter, J. P. (2010). A note on bayesian inference after multiple imputation. *The American Statistician*, 64(2):159–163.