Estimación e imputación de datos faltantes en modelos longitudinales con variable respuesta tipo Poisson y Binomial Negativa con exceso de ceros

Danny Samuel Martínez Lobo

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2018

Estimación e imputación de datos faltantes en modelos longitudinales con variable respuesta tipo Poisson y Binomial Negativa con exceso de ceros

Danny Samuel Martínez Lobo

Tesis presentada como requisito parcial para optar al título de:
Magister en Ciencias-Estadística

Director:
Ph.D. Oscar Orlando Melo Martínez

Línea de Investigación:
Modelos Lineales
Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2018

Dedicatoria

A Angela, por su calidez y amor en esta fría ciudad.

Agradecimientos

Al profesor Oscar Melo, director de esta tesis, sin su colaboración, apoyo, concejos, correcciones, estusiasmo y mucha paciencia, nada de este trabajo hubiera sido posible.

Resumen

El objetivo de esta investigación es la estimación e imputación de datos faltantes en modelos longitudinales con variable respuesta tipo Poisson y Binomial Negativa cero inflada. Para responder al objetivo de está investigación, se propone una metodología que se basa en el uso de la máxima verosimilitud. Se supone que los datos son faltantes de forma aleatoria (FFA), en cada uno de los tiempos se hace uso del algoritmo EM: en el paso E se realiza una regresión ponderada, condicionada a los tiempo anteriores que son tomados como covariables, utilizando la propuesta de Ibrahim (1990). En el paso M se realiza la estimación e imputación de los datos faltantes utilizando la propuesta de Ayala y Melo (2007). La metodología propuesta es aplicada para el caso Poisson cero inflado en el estudio relacionado con el crecimiento del maíz presentado en Da Costa (2003). En el caso Binomial Negativa cero inflada, se aplica a un estudio del forrajeo del polen presentado en Rodríguez (2014).
Palabras clave: Modelos Longitudinales, Algoritmo EM, Poisson cero inflada, Binomial negativa cero inflada, imputación, datos faltantes

Abstract

The objective of this research is the estimation and imputation of missing data in longitudinal models with variable response type Poisson and Negative Binomial Zero Inflated. In order to answer the objective of this study, a methodology is proposed base on the use of the maximum likelihood. The data is supposed to be missing at random (MAR) and in each time the algorithm EM is used. In step E a weighted regression is carried out, conditioned to the previous time that is taken as covariables using the proposal of Ibrahim's (1990). In step M, the estimation and imputation of the missing data is carried out using the methodology of Ayala and Melo (2007). The proposed methodology is applied for the Poisson Zero Inflated Case in a study related to the growth of corn presented in Da Costa (2003). In the case Binomial Negative Zero Inflated, our strategy is applied to a study of the foraging of pollen presented in Rodríguez (2014).
Key words:Longitudinal Models, EM Algorithm, Poisson Zero Inflated, Negative Zero Binomial Inflated, imputation, missing data.

Contenido

Agradecimientos iv

Resumen v

Abstract v

Lista de tablas vii

1. Introducción 2

2. Marco teórico 5
2.1. Datos longitudinales . . . 5
2.2. Modelos lineales generalizados . . . 5
2.2.1. Distribución de probabilidad Poisson inflada por ceros . . . 7
2.2.2. Distribución de probabilidad Binomial Negativa inflada por ceros . . . 8
2.2.3. Estimación de los parámetros del modelo . . . 8
2.3. Datos faltantes en modelos longitudinales . . . 10

3. Algoritmo de estimación e imputación propuesto 12
3.1. Distribución Poisson Cero Inflada . . . 12
3.1.1. Tiempo 1 . . . 17
3.1.2. Tiempo 2 . . . 28
3.1.3. Tiempo 3 . . . 36
3.1.4. Tiempo t . . . 39
3.2. Binomial Negativa Cero Inflada . . . 43
3.2.1. Tiempo 1 . . . 43
3.2.2. Tiempo 2 . . . 55
3.2.3. Tiempo t . . . 56

4. Aplicación 57
4.1. Poisson cero inflado: Estudio mejoramiento del maíz . . . 57
4.1.1. Estimación e imputación de datos faltantes . . . 59
4.1.2. Análisis de los datos: Poisson cero inflada . . . 80
4.2. Binomial Negativa Cero Inflada: estudio del forrajeo del polen . . . 83
4.2.1. Análisis de los datos: Binomial Negativa cero inflada . . . 93

5. Conclusiones **96**

A. Anexo:Regresión Binomial Negativa **101**
A.1. Datos Completos del Forrajeo del Polen 101

B. Código R algoritmo Poisson cero inflada: Datos del Maíz **104**

C. Código R algoritmo Binomial Negativa cero inflada: Datos del Polen **129**

Lista de Tablas

4-1. Estudio del maíz. 59
4-2. Datos del maíz: 20 % de información perdida. 61
4-3. Pesos tiempo 1. 62
4-4. Imputación de datos, tiempo 1. 63
4-5. Pesos tiempo 2. 64
4-6. Imputación de datos, tiempo 2. 65
4-7. Pesos tiempo 3. 66
4-8. Imputación de datos, tiempo 3. 67
4-9. Pesos tiempo 4. 68
4-10.Imputación de datos, tiempo 4. 69
4-11.Pesos tiempo 5. 70
4-12.Imputación de datos, tiempo 5. 71
4-13.Pesos tiempo 6. 72
4-14.Imputación de datos, tiempo 6. 73
4-15.Pesos tiempo 7. 74
4-16.Imputación de datos, tiempo 7. 75
4-17.Pesos tiempo 8. 76
4-18.Imputación de datos, tiempo 8. 77
4-19.Pesos tiempo 9. 78
4-20.Imputación de datos, tiempo 9. 79
4-21.Estimadores de los parámetros para los datos completos: Modelo Poisson . . 80
4-22.Varianza de la parte aleatoria a_i, en el modelo Poisson 81
4-23.Estimadores de los parámetros para los datos completos. 82
4-24.Éxito del algoritmo: Respuesta Poisson cero inflada. 82
4-25.Test de Log-verosimilitud. 84
4-26.Pesos Respuesta Binomial Negativa, tiempo 1. 87
4-27.Pesos Respuesta Binomial Negativa, tiempo 2. 89
4-28.Pesos Respuesta Binomial, tiempo 3. 91
4-29.Parámetros estimados para los datos completos: Modelo Binomial Negativo. 93
4-30.Varianza para la parte aleatoria a_i, del modelo Binomial Negativo Cero Inflada. 94
4-31.Estimadores de los parámetros para los datos completos: Binomial Negativa. 95
4-32.Porcentaje éxito: Respuesta Binomial Negativa cero inflada 95

1. Introducción

En el análisis estadístico de datos es común tener información faltante, esto se debe a diferentes situaciones que se presentan durante el levantamiento de la información. La pérdida de información genera un gran número de problemas en la inferencia de los resultados sobre la población de intéres. Se pueden presentar problemas como estimadores sesgados en la estimación de los parámetros y baja eficiencia de los mismos, entre otros.

En estudios longitudinales es muy común que se tenga la información incompleta y esto constituye uno de los mayores desafíos para el análisis de datos longitudinales (Fitzmaurice et. al. 2009). Existen diferentes situaciones que son comunes en este tipo de estudios como: sujetos que se retiran de la investigación, individuos que se retiran por un tiempo del estudio y regresan a fases posteriores del mismo, sujetos que se incorporan al estudio después del inicio de la investigación. Además de las típicas situaciones de pérdida de información como son: errores en la digitación, falta de compromiso de los sujetos, etc.

En la literatura se encuentran gran variedad de investigaciones sobre información faltante. Se destacan las investigaciones de Dempster, Laird, y Rubin (1977) que proponen completar la información faltante vía algoritmo EM, Fitzmaurice, Laird y Lipsitz (1994) describen un método para el análisis de datos incompletos en estudios longitudinales con respuesta del tipo binaria, Twisk (2003) realiza un análisis del uso de las ecuaciones de estimación generalizadas, Ayala y Melo (2007) estiman datos faltantes en modelos de respuesta tipo binario y de Poisson vía algoritmo EM, Daniels y Hogan (2008) utilizan metodologías bayesiana para estimar datos faltantes en estudios longitudinales, Fitzmaurice et. al. (2009) muestran un resumen de las diferentes metodologías actuales para la estimación de datos faltantes en modelos longitudinales, Chan y Wan (2011) en datos de respuestas longitudinales bivariadas utilizando metodologías bayesianas para estimar datos perdidos y Lukusa, Lee y Li (2014) estiman un modelo lineal ponderado para una variable Poisson cero inflada con datos faltantes en las covariables.

Es de resaltar que la información faltante en datos longitudinales se puede presentar tanto en la variable respuesta como en las variables explicativas. En diversas investigaciones donde las variables explicativas son categóricas, la presencia de información faltante no representa una preocupación muy alta debido a que de antemano se conoce, por el diseño del experimento, la categoría en que se encuentra cada una de las unidades experimentales. Por tanto, en

datos longitudinales donde las variables explicativas son categóricas es de mayor interés el tratamiento de la información faltante en la variable respuesta.

Por otro lado, en los estudios de datos longitudinales es muy común que la variable respuesta no se distribuya de forma normal, para esto existen métodos de estimación como las ecuaciones de estimación generalizadas propuestas por Liang y Zeger (1986) y los modelos lineales generalizados presentados en McCullagh y Nelder (1989).

Los procedimientos anteriores permiten modelar la variable respuesta en los casos que esta sea del tipo discreta. Existen gran cantidad de investigaciones donde se muestra la aplicabilidad de estos procedimientos en diferentes tipos de contextos. Un problema muy frecuente durante las investigaciones con distribuciones de conteo es el exceso de ceros, para ello existen propuestas como la distribución de Poisson y Binomial Negativa inflada por ceros que buscan resolver este inconveniente.

La distribución de Poisson inflada de ceros propuesta por Lambert (1992) y la Binomial Negativa inflada de ceros propuesta por Greene (1994), son distribuciones que permiten tener en cuenta la existencia de excesos de ceros. Las investigaciones con las anteriores distribuciones son muy recientes y la mayor parte son aplicaciones a diferentes contextos que permiten el adecuado ajuste de los estadísticos. Las anteriores distribuciones se hacen presentes en situaciones de diferentes disciplinas tales como: agricultura (Hall, 2000), sociología (Famoye y Singh, 2006), odontología (Mwalili y Declerck, 2007), psicometría (Karazsia y Van Dulmen, 2008), accidentes de tránsito (Sharma y Landge, 2013) y ciencias de la salud (Yao y Liu, 2013), entre otras.

Por consiguiente, las distribuciones infladas por ceros son una muy buena alternativa de ajuste en diferentes investigaciones y los modelos longitudinales con variable respuesta inflada por ceros con información faltante no son ajenos a las posibilidades de modelamiento. Lukusa, Lee y Li (2014) muestran una revisión teórica de la bibliografía existente que relaciona datos cero inflados y datos perdidos con diferentes métodos de estimación propuestos. Es de resaltar, que no se encontraron durante la revisión bibliográfica estimación de datos faltantes en modelos longitudinales que tengan en cuenta las anteriores distribuciones en la variable respuesta. La presente investigación propone una metodología para la estimación e imputación de la información faltante en la variable respuesta de modelos longitudinales con distribuciones Poisson o Binomial Negativa inflada por ceros.

El presente documento se organiza en seis partes que se describen brevemente a continuación.

En el segundo capítulo se presenta el marco teórico, las teorías que se requieren para la investigación como: datos longitudinales, modelos lineales generalizados, métodos de estimación y datos faltantes. En el tercer capítulo la metodología propuesta muestra los pasos necesarios para la estimación e imputación de los datos faltantes, tanto para variable respuesta Poisson cero inflada y Binomial Negativa cero inflada. En el cuarto capítulo se presentan y analizan los resultados obtenidos de la aplicación de la metodología propuesta para el caso Poisson cero inflado a partir de unos datos tomados de Da Costa (2003), relacionado con un estudio del maíz. Para el caso Binomial Negativo cero inflado se utilizan unos datos tomados de Rodríguez (2014). En el quinto capítulo las conclusiones de los resultados más importantes de la investigación llevada a cabo. Finalmente, se presenta la bibliografía consultada para la realización de este trabajo.

2. Marco teórico

Para la elaboración del trabajo se requiere el manejo de ciertas temáticas como: datos longitudinales, modelos lineales generalizados, distribuciones Poisson y Binomial Negativa inflada de ceros, los métodos de estimación y la teoría concerniente a datos faltantes.

2.1. Datos longitudinales

Los datos longitudinales son mediciones que se realizan en el tiempo en un mismo sujeto. En los estudios longitudinales, el intéres radica en modelar el cambio de la variable respuesta a través del tiempo (Fitzmaurice et. al. 2009), además que la inclusión de las diferentes covariables, las cuales pueden variar en el tiempo, complica hacer un buen ajuste.

Un problema que es común en la estadística es la información faltante y los estudios longitudinales son particularmente propensos a problemas de datos faltantes (Fitzmaurice et. al., 2009). Además, que por su propia naturaleza, los datos longitudinales tienen una estructura compleja aleatoria de error que debe tenerse en cuenta para el análisis.

Los modelos lineales predominan en el análisis de los datos longitudinales, en los casos que la variable respuesta es continua y de estructura aleatoria se debe recurrir a los modelos lineales mixtos. En el caso que la variable respuesta es discreta se utilizan los modelos lineales generalizados, y con estructura aleatoria existen los modelos lineales generalizados mixtos y las ecuaciones de estimación generalizadas. Por otro lado, si los coeficientes del modelo de regresión no son lineales, existen alternativas de estimación como los modelos no-parámetricos y semi-parámetricos. Dado que en la presente investigación la variable respuesta es discreta se presentá una introducción a los modelos lineales generalizados.

2.2. Modelos lineales generalizados

Los modelos lineales generalizados se consideran una extensión de los modelos lineales clásicos y se usan en los casos que la distribución de la variable respuesta pertenezca a la familia exponencial. Fueron propuestos inicialmente por Nelder y Wedderburn (1972) y McCullagh y Nelder (1989) unificaron los modelos y las propuestas de modelamiento.

Sea Y una variable aleatoria, está pertenece a la familia exponencial si su función de densidad se puede escribir como:

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

donde $a(\phi)$, $b(\theta)$ y $c(y,\phi)$ son funciones particulares. Si ϕ es conocido entonces la variable aleatoria pertenece a la familia exponencial con θ el parámetro de la distribución. Ahora, si ϕ no es conocido, la variable aleatoria puede pertenecer a la familia exponencial de dos parámetros o no. Si una variable Y pertenece a la familia exponencial entonces se puede escribir como:

$$\ln \ell(\theta,\phi,Y) = \ln f_Y(y,\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

A partir de la función anterior se puede determinar la media y la varianza de una variable aleatoria Y cuya distribución pertenece a la familia exponencial. Las expresiones están dadas por

$$E\left[\frac{d\ell}{d\theta}\right] = 0$$
$$V\left[\frac{d\ell}{d\theta}\right] = -E\left[\frac{d^2\ell}{d\theta^2}\right]$$

de donde

$$E\left[\frac{d\ell}{d\theta}\right] = E\left[\frac{y - b'(\theta)}{a(\phi)}\right] = 0$$
$$\frac{\mu - b'(\theta)}{a(\phi)} = 0$$

Despejando, se tiene que

$$\mu = b'(\theta) = E[Y]$$

y la varianza está dada por

$$V\left[\frac{d\ell}{d\theta}\right] = a(\phi)b''(\theta)$$

Los modelos lineales generalizados se encuentran conformados por tres partes fundamentales:

1. Variables explicativas que pueden ser: continuas, discretas o categóricas.

2. *Función de enlace* o función *link* g que permite relacionar el predictor lineal η con el valor esperado μ de Y. Es decir, describe la relación entre la media μ de la variable respuesta y y las variables explicativas,

$$g(\mu) = \sum_{i=1}^{p} \beta_i x_{ij}$$

3. Variable respuesta donde la distribución de probabilidad pertenece a la familia exponencial.

La adecuada selección de la función de enlace depende de la distribución de la variable respuesta. Las funciones de enlace para las distribuciones Poisson y Binomial negativa inflada por ceros se muestran a continuación.

2.2.1. Distribución de probabilidad Poisson inflada por ceros

Sea $Y = (Y_1, ..., Y_n)'$ el vector de la variable respuesta. Los valores para Y_i son independientes (Lambert, 1992) y se definen como:

$$Y_i = \begin{cases} 0 & \text{con probabilidad} \quad \pi_i, \\ \sim Poisson(\lambda_i) & \text{con probabilidad} \quad 1 - \pi_i, \end{cases}$$

donde $0 \leq \pi_i \leq 1$ y $\lambda_i > 0$. Por tanto, la distribución de probabilidad de la variable aleatoria Y_i es:

$$P[Y_i = y_i] = \begin{cases} \pi_i + (1 - \pi_i) \exp(-\lambda_i) & \text{si} \quad y_i = 0, \\ \dfrac{(1 - \pi_i)\lambda_i^{y_i} . \exp(-\lambda_i)}{y_i!} & \text{si} \quad y_i > 0, \end{cases}$$

Se tiene que $E[Y_i] = (1 - \pi_i)\lambda_i$ y la varianza $Var[Y_i] = (1 - \pi_i)\lambda_i(1 + \pi_i\lambda_i)$. Para modelar la relación entre las covariables y la variable independiente se utiliza la función de enlace ln para λ_i y *logit* para π_i (Ridout et. al. 1998) de la siguiente forma:

$$\ln(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta}$$
$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{z}_i'\boldsymbol{\gamma}$$

donde $\mathbf{x}_i'$ y $\mathbf{z}_i'$ son los vectores de variables aleatorias explicativas, y $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son los vectores de los parámetros de la regresión (Fang, 2013).

2.2.2. Distribución de probabilidad Binomial Negativa inflada por ceros

Sea $Y = (Y_1, ..., Y_n)'$ el vector de la variable respuesta. Los valores para Y_i se definen (Greene, 1994) como:

$$Y_i = \begin{cases} 0 & \text{con probabilidad } \pi_i, \\ \sim BinomialNegativa(\lambda_i, k) & \text{con probabilidad } 1-\pi_i, \end{cases}$$

donde λ_i es la media de la distribución binomial negativa y k el parámetro de sobredispersión. Se tiene que la distribución de probabilidad (Fang, 2013) de la variable aleatoria Y_i es:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)(1+k\lambda_i)^{-1/k} & \text{si } \ y_i = 0, \\ (1-\pi_i)\dfrac{\Gamma(y_i+1/k)(ky_i)^{y_i}}{\Gamma(y_i+1)\Gamma(1/k)(1+k\lambda_i)^{y_i+1/k}} & \text{si } \ y_i > 0, \end{cases}$$

Además, se puede comprobar que $E[Y_i] = (1-\pi_i)\lambda_i$ y la varianza $Var[Y_i] = (1-\pi_i)\lambda_i(1+\pi_i\lambda_i + k\lambda_i)$. Para modelar se utiliza la función de enlace ln para λ_i y *logit* para π_i (Hall y Shen, 2009) de la siguiente forma:

$$\ln(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta}$$
$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i'\boldsymbol{\gamma}$$

donde $\mathbf{x}_i'$ y $\mathbf{z}_i'$ son los vectores de variables aleatorias explicativas, y $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son los vectores de los parámetros de la regresión.

2.2.3. Estimación de los parámetros del modelo

El procedimiento por máxima verosimilitud es el más utilizado para la estimación de los parámetros en los modelos lineales generalizados. A continuación se dan los detalles del algoritmo.

Utilizando el logaritmo de la función de máxima verosimilitud, se deriva e iguala a cero para encontrar el máximo de la función:

$$\frac{\partial \ln \ell}{\partial \boldsymbol{\beta}} = 0 \tag{2-1}$$

haciendo uso de la regla de la cadena en (3-1), se tiene que

$$\frac{\partial \ln \ell}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \tag{2-2}$$

para encontrar las soluciones a (3-2) se hace uso de diferentes algoritmos numéricos de estimación

1. Algoritmo de Newton-Raphson.

   El método usa el vector de los parámetros y la matriz de segundas derivadas de manera iterativa para encontrar los valores estimados de los betas (Davis, 2002). El método se centra en la aproximación mediante la fórmula de Taylor que se encuentra definida por:

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} - \left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}\right]^{-1}_{\beta=b^{(m-1)}} \boldsymbol{U}^{(m-1)}$$

   donde $\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}\right]$ es la matriz de segundas derivadas de la función de verosimilitud evaluados en $\boldsymbol{\beta} = \boldsymbol{b}^{(m-1)}$ y $\mathbf{U}^{(m-1)}$ corresponde al vector de primeras derivadas (Ayala, 2006).

2. Algoritmo de Fisher-Scoring

   Es un algoritmo que se basa en el algoritmo de Newton-Rhapson reemplazando la matriz de segundas derivadas por la matriz de información esperada, este procedimiento muestra estimadores más robustos (Davis, 2002), aunque el algoritmo se considera más lento que Newton-Raphson.

3. Algoritmo EM.

   El algoritmo fue propuesto inicialmente por Dempster et al. (1977). El procedimiento es una técnica iterativa que fue propuesto para calcular y computar estimadores de máxima verosimilitud en datos incompletos.

   Seleccionando un criterio de convergencia, se realizan los siguientes pasos:

   a) Se debe encontrar la esperanza de la log-verosimilitud de los datos faltantes condicionada a los datos observados, teniendo en cuenta a $\boldsymbol{\theta}^{(m)}$, la m-ésima iteracción.

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = E[l(\boldsymbol{\theta}; \mathbf{Y}_{falt})|\mathbf{Y}_{obs}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}]$$

   donde $\ell(\boldsymbol{\theta}; \mathbf{Y}_{falt}|\mathbf{Y}_{obs}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(m)})$ es la log-verosimilitud asociada a los $\mathbf{Y}_{falt}$ que son los valores faltantes de la variable respuesta los $\mathbf{Y}_{obs}$ son los valores observados de la variable respuesta y $\boldsymbol{\theta}$ son los parámetros del modelo de regresión a estimar.

   b) Se debe maximizar encontrando el valor de $\boldsymbol{\theta}$ que máximiza la log-verosimilitud de los resultados del paso anterior, este valor ahora $\boldsymbol{\theta}^{(m+1)}$ se sustituye en el valor de $\boldsymbol{\theta}^{(m)}$ y se vuelve a iterar hasta que converja.

2.3. Datos faltantes en modelos longitudinales

En los estudios longitudinales los problemas de datos faltantes son más probables que en los estudios transversales, generalmente este tipo de investigaciones sufren el problema del desgaste en el tiempo, por lo cual muchos sujetos que hacen parte del estudio deciden abandonar el estudio por un período de tiempo o se retiran definitivamente de la investigación.

Es claro, que tener información faltante tiene implicaciones en el análisis de los datos. Pero tener datos incompletos, no implica que se deba tener una reducción en la precisión de la estimación del modelo, si la información faltante es tratada de forma correcta se pueden tener buenas estimaciones del modelo. Ahora, si no es tratada con cuidado puede introducir sesgo en las estimaciones del modelo.

La precisión en la estimación depende de la cantidad de datos faltantes y la manera que estos se presentan, influyen directamente en el método de análisis. El problema radica en la manera que se presenta la información perdida, es decir el patrón de los datos faltantes.

El inconveniente radica en que el patrón de la información faltante no se encuentra generalmente bajo el control de la investigación. Luego, se hacen suposiciones sobre el mecanismo de la información faltante y la validez de los resultados depende si la suposición sobre el patrón de los datos faltantes es correcta. Es de resaltar, que la investigación se centra en los casos que la información perdida se encuentra presente en la variable respuesta y no en las covariables.

La notación y terminología es: se tienen T medidas repetidas en el tiempo de los n individuos que hagan parte del estudio. El vector de respuestas del i-ésimo individuo en el tiempo t-ésimo está dado por $Y_i = (y_{i1}, \ldots, y_{iT})'$ de tamaño $n \times 1$. Asociada a $\mathbf{y}_i$ se tiene un vector que se define como $\mathbf{x}_i'$ de tamaño $1 \times p$ con p el número de covariables. Ahora, dado que existe información faltante en las respuestas del sujeto se tiene una parte faltante y otra parte observada, el vector de respuestas se puede escribir de la forma $\mathbf{y}_i = (\mathbf{y}_{i(falt)}, \mathbf{y}_{i(obs)})$, donde $\mathbf{y}_{i(obs)}$ el vector de respuestas observadas de tamaño $n_i \times 1$, donde n_i hace referencia a los datos completamente observados en el i-ésimo individuo y $\mathbf{y}_{i(falt)}$ denota el vector de respuestas faltante de tamaño $(T - n_i) \times 1$. Se define como $\mathbf{R}_i$ un vector indicador $\mathbf{R}_i = (R_{i1}, \ldots, R_{iT})'$, con $R_i = 1$ si y_{it} es una respuesta observada y $R_i = 0$ si y_{it} es un dato faltante. Los diferentes patrones de respuesta fueron caracterizados por Rubin (1976) y son:

1. **Datos perdidos completamente aleatorios**.

   Estos ocurren cuando los datos faltan por motivos ajenos a la variable respuesta o a las covariables medidas, y son independientes tanto de los valores observados como no observados de la variable respuesta. Es decir, $\mathbf{R}_i$ es independiente de $\mathbf{y}_{i(falt)}$ y $\mathbf{y}_{i(obs)}$ o que es igual a:

$$P(\mathbf{R}_i|\mathbf{y}_{i(falt)}, \mathbf{y}_{i(obs)}, \mathbf{x}'_{i\times p}) = P(\mathbf{R}_i)$$

En el caso que los datos perdidos sean independientes de los valores de $\mathbf{y}_i$ pero dependan de $\mathbf{x}'_{i\times p}$, se tiene que:

$$P(\mathbf{R}_i|\mathbf{y}_{i(falt)}, \mathbf{y}_{i(obs)}, \mathbf{x}'_{i\times p}) = P(\mathbf{R}_i|\mathbf{x}'_{i\times p})$$

Esto implica que los datos perdidos pueden ser explicados por las covariables que se encuentran en el modelo completo. Bajo el anterior esquema se tiene que la inferencia resultante se debe realizar sobre la información del conjunto de datos completos (Daniels y Hogan, 2008). Es de resaltar, que para los estudios longitudinales este tipo de estructura es muy poco común, ya que perder información de esta manera es muy inusual.

2. **Datos perdidos aleatorios**.

   La estructura consiste en que los datos faltantes son independientes de las respuestas perdidas y se encuentran condicionadas a las respuestas observadas y al modelo de las covariables. Es decir los datos faltantes, dados los observados, son condicionalmente independientes de los datos no observados. Esto implica que los datos faltantes dependen de los observados (Ayala, 2006). Lo anterior, se puede escribir como:

$$P(\mathbf{R}_i|\mathbf{y}_{i(falt)}, \mathbf{y}_{i(obs)}, \mathbf{x}'_{i\times p}) = P(\mathbf{R}_i|\mathbf{y}_{i(obs)}, \mathbf{x}'_{i\times p})$$

   La implicación de la anterior estructura es que los datos observados no pueden ser vistos como una muestra aleatoria de los datos completos, al contrario de la primera estructura. Esta estructura debería ser el supuesto por defecto para el análisis de información faltante en datos longitudinales a menos que exista una razón fuerte y convincente para apoyar otra suposición (Fitzmaurice et. al. 2009).

3. **Datos perdidos no aleatorios**.

   Esta estructura se presenta en los casos que la probabilidad de un dato perdido depende del valor de la respuesta perdida o de otros valores no observables.

$$P(\mathbf{R}_i|\mathbf{y}_{i(falt)}, \mathbf{y}_{i(obs)}, \mathbf{x}'_{i\times p}) = P(\mathbf{R}_i|\mathbf{y}_{i(falt)}, \mathbf{x}'_{i\times p})$$

   Las implicaciones de este tipo de estructura es que los métodos estándar de análisis de datos longitudinales son inválidos.

3. Algoritmo de estimación e imputación propuesto

3.1. Distribución Poisson Cero Inflada

Sean $Y_1, Y_2, \ldots, Y_n$ un conjunto de variables aleatorias independientes con distribución de probabilidad Poisson Cero Inflada, la función de probabilidad se define como:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\lambda_i) & \text{si } y_i = 0 \\ \dfrac{(1-\pi_i)\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!} & \text{si } y_i > 0 \end{cases}$$

donde $E[Y_i] = (1-\pi_i)\lambda_i$ y $Var[Y_i] = (1-\pi_i)[\lambda_i + \pi_i\lambda_i^2]$. Se puede modelar π_i usando un modelo *logit* y λ_i utilizando un modelo ln; los modelos están dados, respectivamente por:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z_i'}\boldsymbol{\gamma}$$
$$\ln(\lambda_i) = \mathbf{x_i'}\boldsymbol{\beta}$$

donde $\mathbf{z}_i'$ es el vector de las covariables asociadas a los datos faltantes del modelo cero de tamaño $1 \times p$, $\boldsymbol{\gamma}$ es el vector de parámetros asociados de tamaño $p \times 1$, $\mathbf{x}_i'$ es el vector de covariables asociadas al modelo de Poisson de tamaño $1 \times p$ y $\boldsymbol{\beta}$ es el vector de parámetros asociados a las covariables de $\mathbf{x}_i$. Despejando en cada una de las ecuaciones, se tiene que:

$$\begin{aligned} \pi_i &= \frac{\exp(\mathbf{z_i'}\boldsymbol{\gamma})}{1+\exp(\mathbf{z_i'}\boldsymbol{\gamma})} \\ \lambda_i &= \exp(\mathbf{x_i'}\boldsymbol{\beta}) \end{aligned} \tag{3-1}$$

En los modelos de ceros inflados, los parámetros asociados a cada uno de los modelos pueden estar relacionados; por ejemplo, el parámetro π_i puede estar relacionado con λ_i. En esta investigación se supone que los modelos que generan los ceros y el modelo Poisson son independientes, por tanto se asume que π_i no está relacionado con λ_i y que $Y_1, Y_2, \ldots, Y_n$ son independientes.

La log-verosimilitud se define como:

$$\begin{aligned}
\ln \ell &= \ln \prod_{i=1}^{N} P(Y_i = y_i) \\
&= \ln \left\{ \prod_{y_i=0} [\pi_i + (1-\pi_i)e^{-\lambda_i}] \cdot \prod_{y_i \neq 0} \left[\frac{(1-\pi_i)e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \right] \right\} \\
&= \sum_{y_i=0} \ln[\pi_i + (1-\pi_i)\exp(-\lambda_i)] \\
&\quad + \sum_{y_i \neq 0} \{\ln(1-\pi_i) - \lambda_i + y_i \ln(\lambda_i) - \ln(y_i!)\} \qquad (3\text{-}2)
\end{aligned}$$

Reemplazando (3-1) en (3-2), se tiene que:

$$\begin{aligned}
\ln \ell &= \sum_{y_i=0} \ln \left[\frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})} + \frac{\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}{1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})} \right] \\
&\quad + \sum_{y_i \neq 0} \left\{ \ln \left[\frac{1}{1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})} \right] - \exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i \mathbf{x}_i'\boldsymbol{\beta} - \ln[y_i!] \right\} \\
&= \sum_{y_i=0} \{-\ln[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})] + \ln[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))]\} \\
&\quad + \sum_{y_i \neq 0} \{-\ln[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})] - \exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i \mathbf{x}_i'\boldsymbol{\beta} - \ln[y_i!]\} \\
&= \sum_{y_i=0} \ln[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))] + \sum_{i=1}^{n} \{-\ln[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})]\} \\
&\quad + \sum_{y_i \neq 0} \{y_i \mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \ln(y_i!)\} \qquad (3\text{-}3)
\end{aligned}$$

Se define una función indicadora D_i como:

$$D_i = \begin{cases} 1 & \text{si } Y_i = 0 \\ 0 & \text{si } Y_i > 0 \end{cases}$$

que permite expresar la verosimilitud como una suma de los datos completos.

Por tanto, la log-verosimilitud queda definida como:

$$\begin{aligned}\ln \ell &= \sum_{i=1}^{n} D_i \ln\left[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\right] - \sum_{i=1}^{n} D_i \ln[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})] \\ &\quad + \sum_{i=1}^{n}(1-D_i)\left[y_i\mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \ln(y_i!)\right] \\ &= \sum_{i=1}^{n}\{(-D_i \ln[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})] + D_i \ln[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))]) \\ &\quad +(1-D_i)\left[-\ln[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})] - \exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln[y_i!]\right]\} \end{aligned} \tag{3-4}$$

La función de verosimilitud se maximiza para estimar los parámetros del modelo. La derivada parcial con respecto a $\boldsymbol{\beta}$ es:

$$\begin{aligned}\frac{\partial \ln \ell}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\left\{D_i\left[\frac{-\exp(\mathbf{x}_i'\boldsymbol{\beta})\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i'}{\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}\right] + (1-D_i)\left[-\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i' + y_i\mathbf{x}_i'\right]\right\} \\ &= \sum_{i=1}^{n}\mathbf{x}_i'\left\{D_i\left[\frac{-\exp(\mathbf{x}_i'\boldsymbol{\beta})\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}{\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}\right] + (1-D_i)\left[-\exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i\right]\right\} \end{aligned} \tag{3-5}$$

Ahora realizando los pasos correspondientes, se tiene que la segunda derivada parcial de (3-5) con respecto a $\boldsymbol{\beta}$ es:

$$\begin{aligned}\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= \sum_{i=1}^{n}\{\mathbf{x}_i'\left[-D_i \exp(\mathbf{x_i}'\boldsymbol{\beta})\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\cdot\right. \\ &\quad \left.\left(\frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta})) - \exp(\mathbf{z}_i'\boldsymbol{\gamma})\exp(\mathbf{x}_i'\boldsymbol{\beta})}{[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))]^2}\right) + (1-D_i)[-\exp(\mathbf{x}_i'\boldsymbol{\beta})]\right]\mathbf{x}_i\Big\} \end{aligned} \tag{3-6}$$

Para facilitar la notación se define:

$$w_i = \exp(\mathbf{z_i}'\boldsymbol{\gamma}) \tag{3-7}$$

Se reemplaza (3-7) en (3-6) y (3-5) y se obtiene:

$$\frac{\partial \ln \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left\{\mathbf{x}_i'\left[\frac{-D_i\lambda_i\exp(-\lambda_i)}{w_i + \exp(-\lambda_i)} + (1-D_i)(y_i - \lambda_i)\right]\right\}$$

$$\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \sum_{i=1}^{n}\left\{\mathbf{x}_i'\left[-D_i\lambda_i\exp(-\lambda_i)\left(\frac{w_i + \exp(-\lambda_i) - w_i\lambda_i}{[w_i + \exp(-\lambda_i)]^2}\right) + (1-D_i)[-\lambda_i]\right]\mathbf{x}_i\right\}$$

Por otro lado, la derivada en (3-4) con respecto a $\boldsymbol{\gamma}$ es:

$$\begin{aligned}\frac{\partial \ln \ell}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n} \left\{ -D_i \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i'}{1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})} + D_i \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i'}{\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))} \right. \\ &\quad \left. + (1-D_i)(-1)\frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i'}{1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})} \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{-\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\mathbf{z}_i' + D_i \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i'}{\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\beta))} \right\} \\ &= \sum_{i=1}^{n} \mathbf{z}_i' \left\{ \frac{-\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})} + D_i \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\beta))} \right\} \end{aligned} \tag{3-8}$$

Reemplazando (3-1) y (3-7) en (3-8) se tiene que:

$$\begin{aligned}\frac{\partial \ln \ell}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n} \mathbf{z}_i' \left\{ \frac{-w_i}{1+w_i} + D_i \frac{w_i}{w_i + \exp(-\lambda_i)} \right\} \\ &= \sum_{i=1}^{n} \mathbf{z}_i' \left\{ -\pi_i + D_i \frac{w_i}{w_i + \exp(-\lambda_i)} \right\} \end{aligned} \tag{3-9}$$

La segunda derivada parcial de (3-9) con respecto a $\boldsymbol{\gamma}$ es:

$$\begin{aligned}\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} &= \sum_{i=1}^{n} \left\{ \left(-\mathbf{z}_i' \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i(1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})) - (\exp(\mathbf{z}_i'\boldsymbol{\gamma}))^2\mathbf{z}_i}{[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})]^2} \right) \right. \\ &\quad \left. + D_i\mathbf{z}_i' \left(\frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i \exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta})) - (\exp(\mathbf{z}_i'\gamma))^2\mathbf{z}_i}{[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))]^2} \right) \right\} \\ &= \sum_{i=1}^{n} \left\{ -\mathbf{z}_i' \left[\frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{[1+\exp(\mathbf{z}_i'\boldsymbol{\gamma})]^2} \right] \mathbf{z}_i \right. \\ &\quad \left. + D_i\mathbf{z}_i' \left[\frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}{[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))]^2} \right] \mathbf{z}_i \right\} \end{aligned}$$

Reemplazando w_i y λ_i se tiene que la segunda derivada con respecto a $\boldsymbol{\gamma}$ es:

$$\begin{aligned}\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} &= \sum_{i=1}^{n} \left\{ -\mathbf{z}_i' \left[\pi_i(1-\pi_i) \right] \mathbf{z}_i + D_i\mathbf{z}_i' \left[\frac{w_i \exp(-\lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \right] \mathbf{z}_i \right\} \\ &= \sum_{i=1}^{n} \mathbf{z}_i' \left[-\pi_i(1-\pi_i) + \frac{D_i w_i \exp(-\lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \right] \mathbf{z}_i \end{aligned}$$

Finalmente derivando parcialmente (3-9) con respecto a $\boldsymbol{\beta}$, se obtiene

$$\begin{aligned}\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}'} &= \sum_{i=1}^{n} -D_i \mathbf{z}_i' \left[\frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma}) \exp(\mathbf{x}_i'\boldsymbol{\beta}) \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}{(\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta})))^2} \right] \mathbf{x}_i \\ &= \sum_{i=1}^{n} -D_i \mathbf{z}_i' \left[\frac{w_i \lambda_i \exp(-\lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \right] \mathbf{x}_i\end{aligned}$$

De igual manera, al derivar parcialmente (3-5) con respecto a $\boldsymbol{\gamma}$ se obtiene:

$$\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} -\mathbf{x}_i' D_i \left[\frac{\lambda_i w_i \exp(-\lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \right] \mathbf{z}_i$$

Luego, la matriz de información de Fisher es:

$$\mathfrak{S} = \begin{pmatrix} \Im_{11} & \Im_{12} \\ \Im_{21} & \Im_{22} \end{pmatrix}$$

donde los elementos $\Im_{ij}$ están dados por las expresiones:

$$\begin{aligned}\Im_{11} &= E\left[-\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right] = \sum_{i=1}^{n} \mathbf{z}_i' \left[\pi_i(1-\pi_i) - \frac{D_i w_i \exp(-\lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \right] \mathbf{z}_i \\ \Im_{22} &= E\left[-\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right] = \sum_{i=1}^{n} \{ \mathbf{x}_i' [(1-D_i)(\lambda_i) \\ &\qquad + \frac{D_i \lambda_i \exp(-\lambda_i)(w_i + \exp(-\lambda_i) - w_i \lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \Big] \mathbf{x}_i \Big\} \\ \Im_{12} &= E\left[-\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}'} \right] = \sum_{i=1}^{n} \mathbf{z}_i' \left[\frac{D_i w_i \lambda \exp(-\lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \right] \mathbf{x}_i \\ \Im_{21} &= E\left[-\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}'} \right] = \sum_{i=1}^{n} \mathbf{x}_i' \left[\frac{D_i w_i \lambda_i \exp(-\lambda_i)}{[w_i + \exp(-\lambda_i)]^2} \right] \mathbf{z}_i\end{aligned}$$

La estimación de los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ del modelo no pueden ser estimados de manera sencilla, dado que la log-verosimilitud queda expresada en términos de los dos parámetros (Hall y Shen, 2009). Luego, una manera sencilla de maximizar la función de log-verosimilitud es haciendo uso del algoritmo EM como lo propone Mouatassim y Ezzahid (2012). Por tanto, a continuación se muestra la descripción detallada del algoritmo propuesto para la estimación e imputación de los datos faltantes, en el caso que la variable respuesta siga una distribución Poisson Cero Inflada. El modelo se esquematiza, para cada tiempo t, en dos diferentes pasos: se generan los k posibles patrones del valor faltante y con estos se realiza la estimación del vector de parámetros con el uso del algoritmo EM y los pesos correspondientes a cada patrón. En el segundo paso, se realiza la imputación teniendo en cuenta los pesos estimados en el paso anterior y se vuelven a estimar los parámetros del modelo. Así, sucesivamente para cada tiempo t.

3.1.1. Tiempo 1

A partir de la función de log-verosimilitud expresada en (3-3):

$$\ln \ell = \sum_{y_{i1}=0} \ln \left[\exp(\mathbf{z}'_{i1}\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}))\right] - \sum_{i=1}^{n} \ln[1 + \exp(\mathbf{z}'_{i1}\boldsymbol{\gamma})] + \sum_{y_{i1}\neq 0} \{y_{i1}\mathbf{x}'_{i1}\boldsymbol{\beta} - \exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) - \ln(y_{i1}!)\} \tag{3-10}$$

donde y_{i1} es el vector respuesta de tamaño $n \times 1$, $\mathbf{x}_{i1}$ es el vector de las covariables asociadas al modelo Poisson y $\mathbf{z}_{i1}$ las covariables del modelo de ceros para la observación i-ésima en el tiempo 1 vectores de tamaño $1 \times p$, donde p es el número de covariables. Se define, para la estimación de los parámetros del modelo en el primer tiempo, una variable indicadora como:

$$D_{i1} = (D_{11}, ..., D_{n1})' \tag{3-11}$$

donde:

$$D_{i1} = \begin{cases} 1 & \text{si } y_{i1} = 0 \\ 0 & \text{si } y_{i1} > 0 \end{cases}$$

tal como lo propone Hall y Shen (2009). Reemplazando (3-11) en (3-10) y se tiene que:

$$\ln \ell = \sum_{i=1}^{n} D_{i1} \left(\ln[\exp(\mathbf{z}'_{i1}\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}))] - \ln[1 + \exp(\mathbf{z}'_{i1}\boldsymbol{\gamma})]\right) + \sum_{i=1}^{n} (1 - D_{i1}) \left(y_{i1}\mathbf{x}'_{i1}\boldsymbol{\beta} - \exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) - \ln[y_{i1}!]\right) \tag{3-12}$$

La maximización de la función log-verosimilitud anterior es complicada por el término:

$$\ln[\exp(\mathbf{z}'_{i1}\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}))]$$

porque involucra los dos parámetros que se buscan estimar. Ahora, si la función indicadora es $D_{i1} = 0$ se tiene que el término:

$$D_{i1} \ln[\exp(\mathbf{z}'_{i1}\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}))] = 0$$

La función indicadora es $D_{i1} = 1$ cuando $y_{i1} = 0$, luego el modelo busca estimar la parte del modelo que es igual a cero o estado perfecto según Lambert (1992), y por tanto, las covariables asociadas al modelo Poisson no se tienen en cuenta en la estimación. Luego, el término que involucra los dos parámetros que se buscan estimar queda reducido a:

$$D_{i1} \ln[\exp(\mathbf{z}'_{i1}\boldsymbol{\gamma})] = D_{i1}\mathbf{z}'_{i1}\boldsymbol{\gamma}$$

Entonces, la expresión (3-10) de la función de log-verosimilitud se reduce a:

$$\ln \ell = \sum_{i=1}^{n} \{D_{i1}\mathbf{z}'_{i1}\boldsymbol{\gamma} - \ln[1 + \exp(\mathbf{z}'_{i1}\boldsymbol{\gamma})]\} + \sum_{i=1}^{n}(1 - D_{i1})\left(y_{i1}\mathbf{x}'_{i1}\boldsymbol{\beta} - \exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) - \ln[y_{i1}!]\right) \quad (3\text{-}13)$$

que puede ser maximizada fácilmente porque:

$$\begin{aligned}\ln \ell &= \sum_{i=1}^{n} \{D_{i1}\mathbf{z}'_{i1}\boldsymbol{\gamma} - \ln[1 + \exp(\mathbf{z}'_{i1}\boldsymbol{\gamma})]\} + \sum_{i=1}^{n}(1 - D_{i1})\left(y_{i1}\mathbf{x}'_{i1}\boldsymbol{\beta} - \exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) - \ln[y_{i1}!]\right) \\ &= \ell(D_{i1}, \boldsymbol{\gamma}, y_{i1}) + \ell(D_{i1}, \boldsymbol{\beta}, y_{i1})\end{aligned}$$

donde:

$$\begin{aligned}\ell(D_{i1}, \boldsymbol{\gamma}, y_{i1}) &= \sum_{i=1}^{n} \{D_{i1}\mathbf{z_{i1}}'\boldsymbol{\gamma} - \ln[1 + \exp(\mathbf{z}'_{i1}\boldsymbol{\gamma})]\} \\ \ell(D_{i1}, \boldsymbol{\beta}, y_{i1}) &= \sum_{i=1}^{n}(1 - D_{i1})\left(y_{i1}\mathbf{x}'_{i1}\boldsymbol{\beta} - \exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) - \ln[y_{i1}!]\right)\end{aligned} \quad (3\text{-}14)$$

Teniendo en cuenta que $y_{i1} = (y_{i1(obs)}, y_{i1(falt)})$ es el vector que hace referencia a los valores de la variable respuesta observados y faltantes. Las verosimilitudes de $\ell(D_{i1}, \boldsymbol{\gamma}, y_{i1})$ y $\ell(D_{i1}, \boldsymbol{\beta}, y_{i1})$ pueden ser maximizadas de manera separada, haciendo uso del algoritmo EM de manera iterativa alternando entre la estimación inicial de $D_{i1}^{(0)}$ dada la esperanza condicional bajo unos parámetros iniciales $(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)})$. Luego, con $D_{i1}^{(0)}$ estimado se maximiza para $\boldsymbol{\beta}^{(1)}$ y $\boldsymbol{\gamma}^{(1)}$ y se procede de manera iterativa hasta la m-ésima iteración para asegurar la convergencia.

A partir de $\ell(D_{i1}, \boldsymbol{\gamma}, y_{i1})$, la verosimilitud para el paso de esperanza, se escribe como:

$$\begin{aligned} Q_{i1}(D_{i1}, \boldsymbol{\gamma}|, D_{i1}^{(m)}, \boldsymbol{\gamma}^{(m)}) &= E[\ell(D_{i1}, \boldsymbol{\gamma}, y_{i1})|x_{i1}, D_{i1} = D_{i1}^{(m)}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(m)}] \\ &= \sum_{i=1}^{n} \ell(D_{i1}, \boldsymbol{\gamma}, y_{i1}) P(y_{i1(falt)(k)}|x_{i1}, D_{i1}^{(m)}, \boldsymbol{\gamma}^{(m)}) \end{aligned} \tag{3-15}$$

Teniendo en cuenta (3-15) se define:

$$w_{i1(k)}^{(m)} = P(y_{i1(falt)(k)}|\mathbf{x}_{i1}, D_{i1}^{(m)}, \boldsymbol{\gamma}^{(m)}) \tag{3-16}$$

como la probabilidad de que un dato $y_{i1(falt)}$ sea igual al valor indicado por k. Luego, reemplazando (3-16) en (3-15) se tiene que:

$$Q_{i1}(D_{i1}, \boldsymbol{\gamma}|D_{i1}^{(m)}, \boldsymbol{\gamma}^{(m)}) = \sum_{i=1}^{n} \ell(D_{i1}, \boldsymbol{\gamma}, y_{i1}) w_{i1(k)}^{(m)} \tag{3-17}$$

De la misma manera a partir de $\ell(D_{i1}, \boldsymbol{\beta}, y_{i1})$ se llega a:

$$Q_{i1}(D_{i1}, \boldsymbol{\beta}|D_{i1}^{(m)}, \boldsymbol{\beta}^{(m)}) = \sum_{i=1}^{n} \ell(D_{i1}, \boldsymbol{\beta}, y_{i1}) w_{i1(k)}^{(m)} \tag{3-18}$$

Las ecuaciones (3-17) y (3-18) corresponden a las verosimilitudes para los parámetros $\boldsymbol{\gamma}$ y $\boldsymbol{\beta}$, respectivamente, de datos completos ponderados que tiene en cuenta el nuevo conjunto de información. Para cada valor faltante, se toman k valores, estos patrones se generan teniendo en cuenta la naturaleza de la variable respuesta. Es decir, como la variable aleatoria se compone de un modelo Poisson y un modelo de exceso de ceros, el valor mínimo que puede tomar es $k = 0$. El modelo Poisson puede generar infinitos número enteros positivos, luego para definir una cota superior se tiene en cuenta el teorema del límite central que muestra que una variable aleatoria $X \sim P(\lambda)$ se aproxima a la distribución normal mediante $Y = \dfrac{X - \lambda}{\sqrt{\lambda}}$ y como $Y \sim N(0, 1)$ el 99 % de los valores de la distribución se encuentran en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$ luego se define la cota superior como $E[y_{i1(obs)}] + 3\sqrt{E[y_{i1(obs)}]}$, recordando que la esperanza es igual a la varianza en una variable aleatoria Poisson. Por ejemplo, si la variable aleatoria tiene una esperanza de 4, la cota superior es de $4 + 3 * \sqrt{4} = 10$, luego los valores o patrones k que puede tomar $y_{i1(falt)}$ son $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ y 10. Las ponderaciones o pesos para el modelo Poisson Cero Inflado se especifican a continuación:

Ponderaciones para el modelo Poisson Cero Inflado

Los pesos para los datos faltantes se definen teniendo en cuenta la propuesta de Ayala y Melo (2007). Sean k los posibles patrones de la variable respuesta, se especifican los valores de la siguiente forma:

Si $k = 1$, se tiene que $y_{i1} = 0$, luego

$$\begin{aligned} w_{i1(1)}^{(m)} &= P[y_{i1(falt)}, k = 1 | \mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}] \\ &= P[y_{i1(falt)} = 0 | \mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}] \\ &= \pi_{i1}^{(m)} + (1 - \pi_{i1}^{(m)}) \exp(-\lambda_{i1}^{(m)}) \end{aligned}$$

Si $k = 2$, se tiene que $y_{i1} = 1$, luego

$$\begin{aligned} w_{i1(2)}^{(m)} &= P[y_{i1(falt)}, k = 2 | \mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}] \\ &= P[y_{i1(falt)} = 1 | \mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}] \\ &= \frac{(1 - \pi_{i1}^{(m)}) \lambda_{i1}^{(m)} \exp(-\lambda_{i1}^{(m)})}{1!} \end{aligned}$$

Siguiendo de la misma manera se tiene que, si $k = k$, se tiene que $y_{i1} = k - 1$, luego

$$\begin{aligned} w_{i1(k)}^{(m)} &= P[y_{i1(falt)}, k = k | \mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}] \\ &= \frac{(1 - \pi_{i1}^{(m)}) (\lambda_{i1}^{(m)})^{k-1} \exp(-\lambda_{i1}^{(m)})}{(k-1)!} \end{aligned}$$

Finalmente, si el dato es observado se tiene que $k = 0$ y se define $w_{i1(0)}^{(m)} = 1$. Por tanto, las ponderaciones para el modelo son:

$$w_{i1(k)}^{(m)} = \begin{cases} 1 & \text{si } k = 0 \\ \pi_{i1}^{(m)} + (1 - \pi_{i1}^{(m)}) \exp(-\lambda_{i1}^{(m)}) & \text{si } k = 1 \\ \dfrac{(1 - \pi_{i1}^{(m)}) \exp(-\lambda_{i1}^{(m)}) (\lambda_{i1}^{(m)})^{k-1}}{(k-1)!} & \text{si } k > 1 \end{cases}$$

Paso 1

Se encuentran los valores estimados para $\boldsymbol{\gamma}$ y $\boldsymbol{\beta}$ haciendo uso del algoritmo EM para los datos del tiempo 1.

Paso E. De la definición del modelo Poisson Cero Inflado se tiene que:

$$\begin{aligned}(Y_{i1}|D_{i1}=1) &= 0\\ (Y_{i1}|D_{i1}=0) &\sim P(\lambda_i)\end{aligned}$$

Luego el valor esperado de D_{i1} bajo los valores iniciales para $\boldsymbol{\beta}^{(m)}$ y $\boldsymbol{\gamma}^{(m)}$, se escribe como:

$$\begin{aligned}E[D_{i1}^{(m)}|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] &= \sum_{i=0}^{1} D_{i1}^{(m)}\cdot P[D_{i1}^{(m)}|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}]\\ &= 1\cdot P[D_{i1}^{(m)}=1|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}]+0\cdot P[D_{i1}^{(m)}=0|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}]\\ &= P[D_{i1}^{(m)}=1|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}]\end{aligned}$$

Usando el teorema de Bayes, se tiene que:

$$P\left[D_{i1}^{(m)}=1|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}\right] = \frac{P[D_{i1}^{(m)}=1]\cdot P[y_{i1}=0|D_{i1}^{(m)}=1,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}]}{P[D_{i1}^{(m)}=1]\cdot P[y_{i1}=0|D_{i1}^{(m)}=1,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}]+P[D_{i1}^{(m)}=0]\cdot P[y_{i1}\neq 0|D_{i1}^{(m)}=0,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}]}$$

Por la definición de la función indicadora se tiene que $P[D_{i1}^{(m)}=1]=\pi_{i1}$ y $P[D_{i1}^{(m)}=0]=1-\pi_{i1}$ se tiene que:

$$\begin{aligned}P[D_{i1}^{(m)}=1|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] &= \begin{cases}\dfrac{1\cdot\pi_{i1}^{(m)}}{1\cdot\pi_{i1}^{(m)}+\exp(-\lambda_{i1}^{(m)})\cdot(1-\pi_{i1}^{(m)})} & \text{si } y_{i1}=0\\ \dfrac{0\cdot\pi_{i1}^{(m)}}{0\cdot\pi_{i1}^{(m)}+(1-\pi_{i1}^{(m)})\cdot\exp(-\lambda_{i1}^{(m)})\cdot\lambda_{i1}^{y_{i1}}/y_{i1}!} & \text{si } y_{i1}>0\end{cases}\\ &= \begin{cases}\dfrac{1}{1+\exp(-\lambda_{i1}^{(m)})\cdot(1-\pi_{i1}^{(m)})/\pi_{i1}^{(m)}} & \text{si } y_{i1}=0\\ 0 & \text{si } y_{i1}>0\end{cases}\\ &= \begin{cases}\dfrac{1}{1+\exp(-\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}^{(m)}))\cdot\exp(-\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})} & \text{si } y_{i1}=0\\ 0 & \text{si } y_{i1}>0\end{cases}\\ &= \begin{cases}\dfrac{1}{1+\exp(-\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}^{(m)})-\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})} & \text{si } y_{i1}=0\\ 0 & \text{si } y_{i1}>0\end{cases}\end{aligned}$$

Por tanto:

$$E[D_{i1}^{(m)}|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] = \begin{cases} \dfrac{1}{1+\exp[-\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}^{(m)}) - \mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}]} & \text{si } y_{i1}=0 \\ 0 & \text{si } y_{i1}>0 \end{cases}$$

Paso M. Se ajusta $\boldsymbol{\gamma}^{(m)}$ por la maximización de $\ell(D_{i1}^{(m)},\boldsymbol{\gamma}^{(m)},\mathbf{y}_{i1})$. A partir del primer término de (3-17):

$$\ell(D_{i1}^{(m)},\boldsymbol{\gamma}^{(m)},\mathbf{y}_{i1}) = \sum_{i=1}^{n}\left\{D_{i1}^{(m)}\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)} - \ln[1+\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})]\right\} w_{i1(k)}^{(m)} \tag{3-19}$$

donde $w_{i1(k)}^{(m)}$ es el peso correspondiente para la $i-$ésima observación en la $m-$ésima iteración del algoritmo en el tiempo 1 en el $k-$ésimo patrón de respuesta faltante. Al realizar la maximización de la log-verosimilitud, según la propuesta original de Lambert (1992), se debe introducir una serie de cambios estructurales en la estimación del modelo que no garantizan la convergencia. Por tanto, se usa la propuesta de Da Costa (2003) que utiliza los valores estimados en el paso E de la variable indicadora como variable respuesta en la maximización de la verosimilitud para la estimación del parámetro $\boldsymbol{\gamma}$. La derivada parcial de (3-19) con respecto a $\boldsymbol{\gamma}^{(m)}$ es:

$$\begin{aligned}\frac{\partial\ell(D_{i1}^{(m)},\boldsymbol{\gamma}^{(m)},\mathbf{y}_{i1})}{\partial\boldsymbol{\gamma}} &= \sum_{i=1}^{n}\left\{D_{i1}^{(m)}\mathbf{z}_{i1}' - \frac{\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})\mathbf{z}_{i1}'}{1+\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})}\right\} w_{i1(k)}^{(m)} \\ &= \sum_{i=1}^{n}\mathbf{z}_{i1}'\left[D_{i1}^{(m)} - \frac{\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})}{1+\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})}\right] w_{i1(k)}^{(m)} \\ &= \sum_{i=1}^{n}\mathbf{z}_{i1}'\left[D_{i1}^{(m)} - \pi_{i1}^{(m)}\right] w_{i1(k)}^{(m)}\end{aligned}$$

La segunda derivada es:

$$\begin{aligned}\frac{\partial^2\ell(D_{i1}^{(m)},\boldsymbol{\gamma}^{(m)},\mathbf{y}_{i1})}{\partial\boldsymbol{\gamma}'\partial\boldsymbol{\gamma}} &= \sum_{i=1}^{n} -\mathbf{z}_{i1}'\left[\frac{\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})}{[1+\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})]^2}\right] w_{i1(k)}^{(m)}\mathbf{z}_i \\ &= \sum_{i=1}^{n} -\mathbf{z}_{i1}'\left[\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})\right] w_{i1(k)}^{(m)}\mathbf{z}_i \end{aligned} \tag{3-20}$$

Luego, la ecuación a resolver es:

$$\frac{\partial^2\ell(D_{i1}^{(m)},\boldsymbol{\gamma}^{(m)},\mathbf{y}_{i1})}{\partial\boldsymbol{\gamma}'\partial\boldsymbol{\gamma}} = \sum_{i=1}^{n} -\mathbf{z}_{i1}'\left[\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})w_{i1(k)}^{(m)}\right]\mathbf{z}_{i1} \tag{3-21}$$

Esta deducción de los pesos a partir de la función de verosimilitud, para la estimación de los datos faltantes, coincide con la propuesta de Hall y Shen (2009), quienes asignaron pesos para realizar una regresión cero inflada más robusta. La propuesta consiste en reemplazar las funciones de estimación para el paso M del algoritmo EM, con funciones que asignan pesos a los datos faltantes. Los pesos son asignados dependiendo del tipo de dato perdido y la estimación se realiza conforme la propuesta de Ayala y Melo (2007), donde se tiene en cuenta la distribución de la variable respuesta para la asignación de los pesos correspondientes.

Para la estimación de los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ se hace uso del método Fisher-Scoring, de donde se tiene que:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + [\mathfrak{S}_z^{(m)}]^{-1}\boldsymbol{U}_z^{(m)} \tag{3-22}$$

donde

$$\begin{aligned}\mathfrak{S}_z^{(m)} &= E\left[-\frac{\partial^2 \ln \ell}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right] \\ &= \sum_{i=1}^{n} \mathbf{z}'_{i1}\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})w_{i1(k)}^{(m)}\mathbf{z}_{i1}\end{aligned}$$

De esta forma, la matriz de información de Fisher queda definida en forma matricial como:

$$\mathfrak{S}_z^{(m)} = \boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1 \tag{3-23}$$

donde $\boldsymbol{W}_1^{(m)} = diag(w_{i1(k)}^{(m)})$ y $\boldsymbol{M}_{1z}^{(m)} = diag(\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)}))$ son matrices de tamaño $n \times n$ y $\boldsymbol{Z}'_1$ es la matriz de covariables asociadas al modelo cero inflado en el tiempo 1 de tamaño $n \times p$. Ahora,

$$\boldsymbol{U}_z^{(m)} = \sum_{i=1}^{n} \mathbf{z}'_{i1}\left[D_{i1}^{(m)} - \pi_{i1}^{(m)}\right] w_{i1(k)}^{(m)}$$

Se multiplica el mismo término en el numerador y denominador, y se obtiene:

$$\boldsymbol{U}_z^{(m)} = \sum_{i=1}^{n}\left\{\mathbf{z}'_{i1}\left[D_{i1}^{(m)} - \pi_{i1}^{(m)}\right] w_{i1(k)}^{(m)}\frac{\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})}{\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})}\right\} \tag{3-24}$$

Se define:

$$\boldsymbol{v}_{z1}^{(m)} = \left(\ldots, \left[D_{i1}^{(m)} - \pi_{i1}^{(m)}\right]\frac{1}{\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})}, \ldots\right)' \tag{3-25}$$

vector de tamaño $n \times 1$. Al reemplazar (3-25) en (3-24), $\boldsymbol{U}_z^{(m)}$ en forma matricial queda expresada como:

$$\boldsymbol{U}_z^{(m)} = \boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{v}_{z1}^{(m)} \tag{3-26}$$

Reemplazando (3-23) y (3-26) en (3-22) se tiene que:

$$\begin{aligned}
\boldsymbol{\gamma}^{(m+1)} &= \boldsymbol{\gamma}^{(m)} + \left[\boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1\right]^{-1}\cdot\boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{v}_{z1}^{(m)} \\
\boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1\boldsymbol{\gamma}^{(m+1)} &= \boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1\boldsymbol{\gamma}^{(m)} + \boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{v}_{z1}^{(m)} \\
\boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1\boldsymbol{\gamma}^{(m+1)} &= \boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\left(\boldsymbol{Z}_1\boldsymbol{\gamma}^{(m)} + \boldsymbol{v}_{z1}^{(m)}\right) \\
\boldsymbol{\gamma}^{(m+1)} &= \left[\boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1\right]^{-1}\cdot\boldsymbol{Z}'_1\boldsymbol{M}_{1z}^{(m)}\boldsymbol{W}_1^{(m)}\left(\boldsymbol{Z}_1\boldsymbol{\gamma}^{(m)} + \boldsymbol{v}_{z1}^{(m)}\right)
\end{aligned}$$

De igual manera, se ajusta $\boldsymbol{\beta}^{(m)}$ por la maximización de $\ell(D_{i1}, \boldsymbol{\beta}^{(m)}, y_{i1})$. A partir del primer término de (3-18) se tiene que:

$$\ell(D_{i1}^{(m)}, \boldsymbol{\beta}^{(m)}, y_{i1}) = \sum_{i=1}^{n}(1 - D_{i1}^{(m)})\left[y_{i1}\mathbf{x}'_{i1}\boldsymbol{\beta}^{(m)} - \exp\left(\mathbf{x}'_{i1}\boldsymbol{\beta}^{(m)}\right)\right]w_{i1(k)}^{(m)}$$

La derivada con respecto a $\boldsymbol{\beta}$ es:

$$\begin{aligned}
\frac{\partial\ell(D_{i1}^{(m)}, \boldsymbol{\beta}^{(m)}, y_{i1})}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}(1 - D_{i1}^{(m)})\left[y_{i1}\mathbf{x}'_{i1} - \exp\left(\mathbf{x}'_{i1}\boldsymbol{\beta}^{(m)}\right)\mathbf{x}'_{i1}\right]w_{i1(k)}^{(m)} \\
&= \sum_{i=1}^{n}\mathbf{x}'_{i1}(1 - D_{i1}^{(m)})\left[y_{i1} - \exp\left(\mathbf{x}'_{i1}\boldsymbol{\beta}^{(m)}\right)\right]w_{i1(k)}^{(m)} \\
&= \sum_{i=1}^{n}\mathbf{x}'_{i1}(1 - D_{i1}^{(m)})\left[y_{i1} - \lambda_{i1}^{(m)}\right]w_{i1(k)}^{(m)} \qquad (3\text{-}27)
\end{aligned}$$

La segunda derivada con respecto a $\boldsymbol{\beta}$ es:

$$\frac{\partial^2\ell(D_{i1}^{(m)}, \boldsymbol{\beta}^{(m)}, y_{i1})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \sum_{i=1}^{n} -\mathbf{x}'_{i1}(1 - D_{i1}^{(m)})\left[\exp\left(\mathbf{x}'_{i1}\boldsymbol{\beta}^{(m)}\right)\right]w_{i1(k)}^{(m)}\mathbf{x}_{i1}$$

De igual manera, se tiene que:

$$\begin{aligned}
\mathfrak{F}_x^{(m)} &= E\left[-\frac{\partial^2\ln\ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right] \\
&= \sum_{i=1}^{n}\mathbf{x}'_{i1}(1 - D_{i1}^{(m)})\lambda_{i1}^{(m)}w_{i1(k)}^{(m)}\mathbf{x}_{i1}
\end{aligned}$$

La expresión de la matriz de información de Fisher en forma matricial es:

$$\mathfrak{F}_x^{(m)} = \mathbf{X_1}'\mathbf{W}_1^{(m)}\mathbf{M}_{1x}^{(m)}\mathbf{X}_1$$

donde $\mathbf{M}_{1x}^{(m)} = diag(\lambda_{i1})$ matriz de tamaño $n \times n$ y $\mathbf{X}_1$ es la matriz de covariables asociadas al modelo Poisson en el tiempo 1 de tamaño $n \times p$.

De esta forma,

$$\boldsymbol{U}_x^{(m)} = \sum_{i=1}^{n} w_{i1(k)}^{(m)}(1 - D_{i1}^{(m)})\mathbf{x}_{i1}' \left[y_{i1} - \lambda_{i1}^{(m)}\right]$$

Al multiplicar por $\lambda_{i1}^{(m)}$ en el numerador y denominador, se obtiene:

$$\boldsymbol{U}_x^{(m)} = \sum_{i=1}^{n} w_{i1(k)}^{(m)}(1 - D_{i1}^{(m)})\mathbf{x}_{i1}' \left[y_{i1} - \lambda_{i1}^{(m)}\right] \frac{\lambda_{i1}^{(m)}}{\lambda_{i1}^{(m)}}$$

Se define:

$$\boldsymbol{v}_{x1}^{(m)} = \left(\ldots, (1 - D_{i1}^{(m)}) \left[y_{i1} - \lambda_{i1}^{(m)}\right] \frac{1}{\lambda_{i1}^{(m)}}, \ldots\right)'$$

vector de tamaño $n \times 1$. Luego, se tiene que

$$\boldsymbol{U}_x^{(m)} = \boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{v}_{x1}^{(m)}$$

De manera similar que en la estimación de $\boldsymbol{\gamma}^{(m+1)}$, se debe recurrir al método Fisher-Scoring para la estimación de $\boldsymbol{\beta}$. La deducción es:

$$\begin{aligned}
\boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} + [\boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{1x}^{(m)} \boldsymbol{X}_1]^{-1} \cdot \boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{x1}^{(m)} \boldsymbol{v}_{x1}^{(m)} \\
\boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{1x}^{(m)} \boldsymbol{X}_1 \boldsymbol{\beta}^{(m+1)} &= \boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{1x}^{(m)} \boldsymbol{X}_1 \boldsymbol{\beta}^{(m)} + \boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{x1}^{(m)} \boldsymbol{v}_{x1}^{(m)} \\
\boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{1x}^{(m)} \boldsymbol{X}_1 \boldsymbol{\beta}^{(m+1)} &= \boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{1x}^{(m)} \left(\boldsymbol{X}_1 \boldsymbol{\beta}^{(m)} + \boldsymbol{v}_{x1}^{(m)}\right) \\
\boldsymbol{\beta}^{(m+1)} &= [\boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{1x}^{(m)} \boldsymbol{X}_1]^{-1} \cdot \boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{1x}^{(m)} \left(\boldsymbol{X}_1 \boldsymbol{\beta}^{(m)} + \boldsymbol{v}_{x1}^{(m)}\right)
\end{aligned}$$

Paso 2

Estimación e imputación de los datos faltantes en el tiempo 1. Se realiza el algoritmo EM para los datos en el primer tiempo teniendo en cuenta como estimador inicial los parámetros obtenidos en el paso 1.

Paso E. El paso E imputa los valores para los datos faltantes, utilizando el modelo propuesto por Bartlett (1937):

$$E[g(\mathbf{y}_1)|\mathbf{X}_1, \mathbf{Z}_1] = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}_1 \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{0} \\ \mathbf{B}_1 \end{pmatrix} \boldsymbol{\tau} \tag{3-28}$$

donde $\mathbf{A}_1$ y $\mathbf{B}_1$ son matrices de tamaños $n \times (n - n_1)$ donde n es el total de datos, n_1 indica el total de datos observados y $(n - n_1)$ indican las covariables de valor faltante del primer tiempo con las matrices $\mathbf{A}_1$ y $\mathbf{B}_1$ que corresponden a las covariables del modelo Poisson y

cero inflado, respectivamente, de los valores faltantes del primer tiempo. $\boldsymbol{\alpha}$ y $\boldsymbol{\tau}$ corresponden a los vectores de los $(n-n_1)$ coeficientes de regresión para las covariables de valor faltante y $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son los coeficientes estimados en el primer tiempo. Las matrices $\mathbf{X}_1$ y $\mathbf{Z}_1$ son las matrices de covariables asociadas al primer tiempo del modelo Poisson y cero inflado respectivamente.

Se particionan las matrices $\mathbf{A}_1$ y $\mathbf{B}_1$ entre valores observados y esperados y se obtiene:

$$\begin{pmatrix} \mathbf{ln}(\boldsymbol{\lambda}_{\mathbf{1}(\boldsymbol{onc})}) \\ \mathbf{ln}(\boldsymbol{\lambda}_{\mathbf{1}(\boldsymbol{fnc})}) \\ \mathbf{ln}\left(\dfrac{\boldsymbol{\pi}_{\mathbf{1}(\boldsymbol{oc})}}{\mathbf{1}-\boldsymbol{\pi}_{\mathbf{1}(\boldsymbol{oc})}}\right) \\ \mathbf{ln}\left(\dfrac{\boldsymbol{\pi}_{\mathbf{1}(\boldsymbol{fc})}}{\mathbf{1}-\boldsymbol{\pi}_{\mathbf{1}(\boldsymbol{fc})}}\right) \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{1(obs)} \\ \mathbf{X}_{1(falt)} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{Z}_{1(o)} \\ \mathbf{Z}_{1(f)} \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} \mathbf{0}_{1(obs)} \\ -\boldsymbol{I}_{1(falt)} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0}_{1(obs)} \\ -\boldsymbol{I}_{1(falt)} \end{pmatrix} \boldsymbol{\tau} \tag{3-29}$$

donde $\mathbf{X}_{\mathbf{1(obs)}}$ y $\mathbf{Z}_{\mathbf{1(obs)}}$ corresponden a matrices de covariables de los valores observados $\mathbf{X}_{\mathbf{1(falt)}}$ y $\mathbf{Z}_{\mathbf{1(falt)}}$ corresponden a matrices de covariables de valores faltantes en el tiempo 1, tanto del modelo Poisson y el modelo Cero Inflado respectivamente y $\mathbf{0}_{1(obs)}$ son matrices relacionadas con los datos observados e $-\mathbf{I}_{1(falt)}$ son matrices identidad negativa relacionadas con los faltantes en el tiempo 1. $\boldsymbol{\lambda}_{1(onc)}$ es el vector relacionado con los valores observados diferentes de cero en el tiempo 1, $\boldsymbol{\lambda}_{1(fnc)}$ es el vector relacionado con los valores faltantes diferentes de cero, $\boldsymbol{\lambda}_{1(oc)}$ es el vector relacionado de los valores observados que son cero y $\boldsymbol{\lambda}_{1(fc)}$ es el vector de los valores faltantes que son cero.
Con base en lo anterior, para el primer tiempo la log-verosimilitud esperada dados los datos observados se escribe como:

$$Q_{i1}[\boldsymbol{\theta}|\boldsymbol{\theta}^m] = E[L(\boldsymbol{\theta}; g(y_{i1}))|\mathbf{x}_{i1}, \mathbf{z}_{i1}, \mathbf{y}_{i1(obs)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}] \tag{3-30}$$

donde $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\tau})$. De (3-30) se tiene que:

$$\begin{aligned} E[Y_{1(obs)}] &= \mathbf{X}_{1(obs)}\boldsymbol{\beta} + \mathbf{Z}_{1(obs)}\boldsymbol{\gamma} \\ E[Y_{1(falt)}] &= \mathbf{X}_{1(falt)}\boldsymbol{\beta} - \boldsymbol{\alpha} + \mathbf{Z}_{1(falt)}\boldsymbol{\gamma} - \boldsymbol{\tau} \end{aligned} \tag{3-31}$$

Suponiendo un vector inicial para $Y_{1(falt)}$, $\hat{Y}_{1(falt)} = 0$ de acuerdo a la descripción original del método de Bartlett (1937) y al considerar que el modelo que concierne a la parte Poisson es independiente del modelo de ceros, la suma de los cuadrados de los errores al ser minimizados sobre $\boldsymbol{\theta}^{(m)}$ es:

$$\begin{aligned} SCE(\boldsymbol{\theta}^{(m)}) = &\sum_{i=1}^{n_{oc}} (\boldsymbol{y}_{i1onc} - \mathbf{x}'_{i1(obs)}\boldsymbol{\beta})^2 + \sum_{i=n_{oc}+1}^{n_{fc}} (\widehat{\boldsymbol{y}}_{i1fnc} - \mathbf{x}'_{i1(falt)}\boldsymbol{\beta} + \boldsymbol{\alpha})^2 \\ &+ \sum_{i=n_{fc}+1}^{n_{fnc}} (\boldsymbol{y}_{i1oc} - \mathbf{z}'_{i1(obs)}\boldsymbol{\gamma})^2 + \sum_{i=n_{fnc}+1}^{n_f} (\widehat{\boldsymbol{y}}_{i1fc} - \mathbf{z}'_{i1(falt)}\boldsymbol{\gamma} + \boldsymbol{\tau})^2 \end{aligned} \tag{3-32}$$

donde n_{oc} hace referencia a los valores observados que son cero, n_{fc} valores faltantes que son cero, n_{fnc} valores faltantes que no son cero y n_{fc} valores faltantes que son cero.

Teniendo en consideración la suposición inicial acerca de $\widehat{\boldsymbol{y}}_{i1(falt)} = 0$, el estimador de mínimos cuadrados para $\boldsymbol{\alpha}$ y $\boldsymbol{\tau}$, se obtiene a partir de (3-32) :

$$\frac{\partial SCE(\boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\alpha}} = 2 \sum_{i=n_{oc}+1}^{nfc} (\widehat{y}_{i1fnc} - \mathbf{x}'_{i1}\boldsymbol{\beta} + \boldsymbol{\alpha}) = 0$$
$$\frac{\partial SCE(\boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\tau}} = 2 \sum_{i=n_{fnc}+1}^{nf} (\widehat{y}_{i1fc} - \mathbf{z}'_{i1}\boldsymbol{\gamma} + \boldsymbol{\tau}) = 0$$

Al despejar en las ecuaciones y escribiendo en forma matricial se tiene que:

$$\widehat{\boldsymbol{\alpha}} = \mathbf{X}_{1(falt)}\widehat{\boldsymbol{\beta}}$$
$$\widehat{\boldsymbol{\tau}} = \mathbf{Z}_{1(falt)}\widehat{\boldsymbol{\gamma}}$$

Por tanto, el estimador de mínimos cuadrados para la información faltante corresponde al valor esperado de la respuesta. Luego, de manera más específica se tiene que en la iteracción m-ésima los parámetros estimados son:

$$\widehat{\alpha}_i^{(m)} = \mathbf{x}'_{i1(falt)}\widehat{\boldsymbol{\beta}}^{(m)} = \ln[\widehat{\lambda}^{(m)}_{i1(falt)}] \qquad \text{si } y_{i1} > 0,\ i = n_{oc}+1, \ldots, n_{fc}$$
$$\widehat{\tau}_i^{(m)} = \mathbf{z}'_{i1(falt)}\widehat{\boldsymbol{\gamma}}^{(m)} = \ln\left[\frac{\widehat{\pi}^{(m)}_{i1(falt)}}{1 - \widehat{\pi}^{(m)}_{i1(falt)}}\right] \qquad \text{si } y_{i1} = 0,\ i = n_{fnc}+1, \ldots, n_f$$

Por lo tanto,

$$\widehat{\lambda}^{(m)}_{i1(falt)} = \exp\left(\mathbf{x}'_{i1(falt)}\boldsymbol{\beta}^{(m)}\right) \qquad \text{si } y_{i1} > 0$$
$$\widehat{\pi}^{(m)}_{i1(falt)} = \frac{\exp\left(\mathbf{z}'_{i1(falt)}\boldsymbol{\gamma}^{(m)}\right)}{1 + \exp\left(\mathbf{z}'_{i1(falt)}\boldsymbol{\gamma}^{(m)}\right)} \qquad \text{si } y_{i1} = 0$$

donde $\hat{\lambda}^{(m)}_{i1}$ y $\hat{\pi}^{(m)}_{i1}$ hacen referencia al valor de la media correspondiente al individuo i en el tiempo 1, si su respuesta es diferente de cero o igual a cero respectivamente. Dada la m-ésima iteración el criterio de selección es:

$$\widehat{y}_{i1(falt)} = \begin{cases} 0 & \text{si } \widehat{\pi}^{(m)}_{i1(falt)} > p_o \\ \widehat{\lambda}^{(m)}_{i1(falt)} & \text{si } \widehat{\pi}^{(m)}_{i1(falt)} \leq p_o \end{cases}$$

El valor p_0 es un valor de referencia para imputar el dato como cero o diferente de cero. En las variables cero infladas, generalmente según Da Costa (2003), el porcentaje de ceros es mínimo del 40 % de las observaciones. Por tanto, 0.4 puede ser un valor mínimo si no se tiene conocimiento profundo de la variable que se trabaja.

Paso M. Imputadas las observaciones se procede a la maximización partiendo del conjunto de datos completos, teniendo en cuenta el método Fisher-Scoring, con el estimador inicial dado en el paso 1. Se realiza el paso E y M hasta lograr la convergencia. La variable indicadora que permite estimar el exceso de ceros y el modelo de Poisson como independiente es un vector de tamaño n que toma valores según:

$$E\left[D_{i1}^{(m)}|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}\right]=\begin{cases}\dfrac{1}{1+\exp\left(-\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}-\exp\left(\mathbf{x}_{i1}'\boldsymbol{\beta}^{(m)}\right)\right)} & \text{si } y_{i1}=0\\ 0 & \text{si } y_{i1}>0\end{cases}$$

Las expresiones para la estimación de los parámetros son de la forma:

$$\begin{aligned}\boldsymbol{\gamma}^{(m+1)}&=\left[\boldsymbol{Z}_1'\boldsymbol{M}_{1z}^{(m)}\boldsymbol{Z}_1\right]^{-1}\boldsymbol{Z}_1'\boldsymbol{M}_{1z}^{(m)}\left(\boldsymbol{Z}_1\boldsymbol{\gamma}^{(m)}+\boldsymbol{v}_{z1}^{(m)}\right)\\ \boldsymbol{\beta}^{(m+1)}&=[\boldsymbol{X}_1'\boldsymbol{M}_{1x}^{(m)}\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'\boldsymbol{M}_{1x}^{(m)}\left(\boldsymbol{X}_1\boldsymbol{\beta}^{(m)}+\boldsymbol{v}_{x1}^{(m)}\right)\end{aligned}$$

con $\boldsymbol{Z}_1$ y $\boldsymbol{X}_1$ de tamaños $n\times p$. $\boldsymbol{M}_{1z}^{(m)}=diag(\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)}))$ y $\boldsymbol{M}_{1x}^{(m)}=diag(\lambda_{i1}^{(m)})$ de tamaños $n\times n$ y los vectores siguientes de dimensión $n\times 1$:

$$\begin{aligned}\boldsymbol{v}_{z1}^{(m)}&=\left(\ldots,\left[D_{i1}^{(m)}-\pi_{i1}^{(m)}\right]\frac{1}{\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})},\ldots\right)'\\ \boldsymbol{v}_{x1}^{(m)}&=\left(\ldots,(1-D_{i1}^{(m)})\left[y_{i1}-\lambda_{i1}^{(m)}\right]\frac{1}{\lambda_{i1}^{(m)}},\ldots\right)'\end{aligned}$$

3.1.2. Tiempo 2

En la estimación de los parámetros del modelo en el segundo tiempo, se emplea la variable respuesta del primer tiempo y_{i1} como covariable asociada en el modelo utilizando el algoritmo EM con ponderaciones. Durante la imputación de los valores faltantes del primer tiempo, puede suceder que las entradas de la variable sean solamente cero, está situación se presentó en algunas ocasiones durante las simulaciones llevadas a cabo en el desarrollo de la presente investigación. Al tener una variable con todas las entradas iguales, no mejora la explicación de la variabilidad de la variable respuesta. Luego, si sucede el caso, esta variable no se debe considerar y las funciones de enlace para los parámetros λ_{i2} y π_{i2} se deben definir como:

$$\begin{aligned}\ln[\lambda_{i2}]&=\boldsymbol{x}_{i2}'\boldsymbol{\beta}\text{ si }y_{i2}>0\\ \ln\left[\frac{\pi_{i2}}{1-\pi_{i2}}\right]&=\boldsymbol{z}_{i2}'\boldsymbol{\gamma}\text{ si }y_{i2}=0\end{aligned}\tag{3-33}$$

La estimación de los parámetros del modelo y la imputación de los datos se lleva a cabo de manera similar conforme a lo propuesto para el primer tiempo.

Ahora, si las entradas de y_{i1} no son todas cero, situación más usual, se debe llevar a cabo la estimación del vector de parámetros máximo verosímiles en la segunda ocasión $t = 2$ igual que en el paso 1, con el empleo del algoritmo EM teniendo en cuenta y_{i1} como covariable asociada en el modelo. Las funciones de enlace para los parámetros son:

$$\begin{aligned} \ln[\lambda_{i2}] &= \mathbf{x}'_{i2}\boldsymbol{\beta} \text{ si } y_{i2} > 0 \\ \ln\left[\frac{\pi_{i2}}{1-\pi_{i2}}\right] &= \mathbf{z}'_{i2}\boldsymbol{\gamma} \text{ si } y_{i2} = 0 \end{aligned} \tag{3-34}$$

El modelo en forma matricial está dado por:

$$E[g(\mathbf{y}_2)] = \begin{pmatrix} \mathbf{X}_2 & \mathbf{Z}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} + e$$

donde $\mathbf{X}_2$ y $\mathbf{Z}_2$ son las matrices de covariables relacionadas con $\mathbf{y}_2$, el vector de respuesta en el segundo tiempo. Las matrices $\mathbf{X}_2$ y $\mathbf{Z}_2$ son de tamaño $n \times (p+1)$, donde p columnas con las variables independientes que explican el comportamiento de la variable dependiente y la última columna es el vector de las respuestas en el tiempo 1.
La log-verosimilitud se define como:

$$\begin{aligned} \ln \ell = &\sum_{y_{i2}=1} \ln[\pi_{i2} + (1-\pi_{i2})\exp(-\lambda_{i2})] \\ &+ \sum_{y_{i2}\neq 0} \{\ln(1-\pi_{i2}) - \lambda_{i2} + y_{i2}\ln(\lambda_{i2}) - \ln(y_{i2}!)\} \end{aligned} \tag{3-35}$$

Sustituyendo (3-34) en (3-35) se tiene que:

$$\begin{aligned} \ln \ell = &\sum_{y_{i2}\neq 0} \{-\ln[1+\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma})] - \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})\} + \sum_{y_{i2}\neq 0} \{y_{i2}(\mathbf{x}'_{i2}\boldsymbol{\beta}) - \ln(y_{i2}!)\} \\ &+ \sum_{y_{i2}=0} \ln\left[\frac{\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma})}{1+\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma})} + \frac{\exp(-\exp(\mathbf{x}'_{i2}\boldsymbol{\beta}))}{1+\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma})}\right] \\ = &\sum_{y_{i2}\neq 0} \{-\ln[1+\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma})] - \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})\} + \sum_{y_{i2}\neq 0} \{y_{i2}(\mathbf{x}'_{i2}\boldsymbol{\beta}) - \ln(y_{i2}!)\} \\ &+ \sum_{y_{i2}=0} \ln[\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}'_{i2}\boldsymbol{\beta}))] - \sum_{i=1}^{n} \ln[1+\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma})] \\ = &\sum_{y_{i2}=0} \ln[\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}'_{i2}\boldsymbol{\beta}))] + \sum_{y_{i2}\neq 0} \{y_{i2}(\mathbf{x}'_{i2}\boldsymbol{\beta}) - \exp(\mathbf{x}'_{i2}\boldsymbol{\beta}) - \ln(y_{i2}!)\} \\ &- \sum_{i=1}^{n} \ln[1+\exp(\mathbf{z}'_{i2}\boldsymbol{\gamma})] \end{aligned}$$

Nuevamente se define una función indicadora D_{i2} como:

$$D_{i2} = \begin{cases} 1 & \text{si } y_{i2} = 0 \\ 0 & \text{si } y_{i2} > 0 \end{cases}$$

que permite expresar la verosimilitud como una suma de datos completos. Por tanto, la log-verosimilitud queda del modelo queda definida como:

$$\begin{aligned} \ln \ell = & \sum_{i=1}^{n} D_{i2} \ln \left[\exp \left(\mathbf{z}_{i2}' \boldsymbol{\gamma} \right) + \exp \left(- \exp \left(\mathbf{x}_{i2}' \boldsymbol{\beta} \right) \right) \right] \\ & + \sum_{i=1}^{n} (1 - D_{i2}) \left[y_{i2} \left(\mathbf{x}_{i2}' \boldsymbol{\beta} \right) - \exp \left(\mathbf{x}_{i2}' \boldsymbol{\beta} \right) - \ln(y_{i2}!) \right] \\ & - \sum_{i=1}^{n} \ln[1 + \exp \left(\mathbf{z}_{i2}' \boldsymbol{\gamma} \right)] \end{aligned}$$

Si la función indicadora es $D_{i2} = 0$ se tiene que el término:

$$D_{i2} \ln \left[\exp \left(\mathbf{z}_{i2}' \boldsymbol{\gamma} \right) + \exp \left(- \exp \left(\mathbf{x}_{i2}' \boldsymbol{\beta} \right) \right) \right] = 0$$

De otro lado, si la función indicadora es $D_{i2} = 1$ entonces el modelo busca estimar el estado cero o estado perfecto (Lambert, 1992) y por tanto, las covariables asociadas al modelo Poisson no se tienen en cuenta. Luego,

$$D_{i2} \ln[\exp \left(\mathbf{z}_{i2}' \boldsymbol{\gamma} \right)] = D_{i2} (\mathbf{z}_{i2}' \boldsymbol{\gamma})$$

Entonces, la log-verosimilitud queda expresada como:

$$\begin{aligned} \ln \ell = & \sum_{i=1}^{n} \left\{ D_{i2} (\mathbf{z}_{i2}' \boldsymbol{\gamma}) - \ln[1 + \exp \left(\mathbf{z}_{i2}' \boldsymbol{\gamma} \right)] \right\} \\ & + \sum_{i=1}^{n} (1 - D_{i2}) \left[y_{i2} (\mathbf{x}_{i2}' \boldsymbol{\beta}) - \exp \left(\mathbf{x}_{i2}' \boldsymbol{\beta} \right) - \ln(y_{i2}!) \right] \end{aligned}$$

donde

$$\begin{aligned} \ell(D_{i2}, \boldsymbol{\gamma}, y_{i2}) &= \sum_{i=1}^{n} \left\{ D_{i2} (\mathbf{z}_{i2}' \boldsymbol{\gamma}) - \ln[1 + \exp \left(\mathbf{z}_{i2}' \boldsymbol{\gamma} \right)] \right\} \\ \ell(D_{i2}, \boldsymbol{\beta}, y_{i2}) &= \sum_{i=1}^{n} (1 - D_{i2}) \left[y_{i2} (\mathbf{x}_{i2}' \boldsymbol{\beta}) - \exp \left(\mathbf{x}_{i2}' \boldsymbol{\beta} \right) - \ln(y_{i2}!) \right] \end{aligned}$$

La verosimilitud de $\ell(D_{i2}, \boldsymbol{\gamma}, y_{i2})$ y $\ell(D_{i2}, \boldsymbol{\beta}, y_{i2})$ pueden ser maximizadas de manera separada utilizando el algoritmo EM de manera iterativa, alternando entre la estimación de $D_{i2}^{(0)}$ dada la esperanza condicional bajo unos parámetros iniciales $\boldsymbol{\gamma}^{(0)}$ y $\boldsymbol{\beta}^{(0)}$ estimados inicialmente. Luego con $D_{i2}^{(0)}$ estimado se maximiza para $\boldsymbol{\gamma}^{(1)}$ y $\boldsymbol{\beta}^{(1)}$, y se procede de manera iterativa hasta la m-ésima iteración para lograr la convergencia.

Paso 1

Se encuentran los valores estimados para $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ haciendo uso del algoritmo EM para los datos del tiempo 2.

Paso E. Se estima D_{i2} dada la media condicional $D_{i2}^{(m)}$ bajo las estimaciones iniciales de $\boldsymbol{\beta}^{(m)}$ y $\boldsymbol{\gamma}^{(m)}$, esto es:

$$\begin{aligned} E[D_{i2}^{(m)}|y_{i2},\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}] &= \sum_{i=1}^{1} D_{i2}^{(m)} \cdot P[D_{i2}^{(m)}|y_{i2},\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}] \\ &= 1 \cdot P[D_{i2}^{(m)} = 1|y_{i2},\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}] + 0 \cdot P[D_{i2}^{(m)} = 0|y_{i2},\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}] \\ &= P[D_{i2}^{(m)} = 1|y_{i2},\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}] \end{aligned}$$

Luego, por el teorema de bayes se tiene que:

$$P\left[D_{i2}^{(m)} = 1|y_{i2},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}\right] = \frac{P\left[y_{i2} = 0,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}|D_{i2}^{(m)} = 1\right] \cdot P\left[D_{i2}^{(m)} = 1\right]}{P\left[y_{i2} = 0,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}|D_{i2}^{(m)} = 1\right] \cdot P\left[D_{i2}^{(m)} = 1\right] + P\left[y_{i2} \neq 0,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}|D_{i2}^{(m)} = 0\right] \cdot P\left[D_{i2}^{(m)} = 0\right]}$$

Por la definición de la función indicadora se tiene que $P[D_{i2}^{(m)} = 1] = \pi_{i2}^{(m)}$ y de manera similiar $P[D_{i2}^{(m)} = 0] = 1 - \pi_{i2}^{(m)}$ luego

$$\begin{aligned} E\left[D_{i2}^{(m)}|y_{i2},\boldsymbol{\gamma^{(m)}},\boldsymbol{\beta^{(m)}}\right] &= \begin{cases} \dfrac{1 \cdot \pi_{i2}^{(m)}}{1 \cdot \pi_{i2}^{(m)} + \exp\left(-\lambda_{i2}^{(m)}\right) \cdot (1 - \pi_{i2}^{(m)})} & \text{si } y_{i2} = 0 \\ \dfrac{0 \cdot \pi_{i2}^{(m)}}{0 \cdot \pi_{i2}^{(m)} + (1 - \pi_{i2}^{(m)}) \cdot \exp\left(-\lambda_{i2}^{(m)}\right) \cdot \lambda_{i2}^{y_{i2}}/y_{i2}!} & \text{si } y_{i2} > 0 \end{cases} \\ &= \begin{cases} \dfrac{1}{1 + \exp\left(-\lambda_{i2}^{(m)}\right) \cdot (1 - \pi_{i2}^{(m)})/\pi_{i2}^{(m)}} & \text{si } y_{i2} = 0 \\ 0 & \text{si } y_{i2} > 0 \end{cases} \\ &= \begin{cases} \dfrac{1}{1 + \exp\left(-\exp\left(\mathbf{x}_{i2}'\boldsymbol{\beta^{(m)}}\right)\right) \cdot \exp\left(-\mathbf{z}_{i2}'\boldsymbol{\gamma^{(m)}}\right)} & \text{si } y_{i2} = 0 \\ 0 & \text{si } y_{i2} > 0 \end{cases} \\ &= \begin{cases} \dfrac{1}{1 + \exp\left[-\exp\left(\mathbf{x}_{i2}'\boldsymbol{\beta^{(m)}}\right) - \mathbf{z}_{i2}'\boldsymbol{\gamma^{(m)}}\right]} & \text{si } y_{i2} = 0 \\ 0 & \text{si } y_{i2} > 0 \end{cases} \end{aligned}$$

Paso M. Se ajusta $\boldsymbol{\gamma}^{(m)}$ por la maximización de $\ell(\boldsymbol{\gamma}^{(m)}, y_{i2}, D_{i2}^{(m)})$ se tiene que la primera derivada es:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\gamma}^{(m)}|y_{i2}, D_{i2}^{(m)})}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n}\left\{D_{i2}^{(m)}\mathbf{z}_{i2}' - \frac{\exp\left(\mathbf{z}_{i2}'\boldsymbol{\gamma}^{(m)}\right)}{1+\exp\left(\mathbf{z}_{i2}'\boldsymbol{\gamma}^{(m)}\right)}\mathbf{z}_{i2}'\right\}\\ &= \sum_{i=1}^{n}\mathbf{z}_{i2}'\left(D_{i2}^{(m)} - \frac{\exp\left(\mathbf{z}_{i2}'\boldsymbol{\gamma}^{(m)}\right)}{1+\exp\left(\mathbf{z}_{i2}'\boldsymbol{\gamma}^{(m)}\right)}\right)\\ &= \sum_{i=1}^{n}\mathbf{z}_{i2}'\left(D_{i2}^{(m)} - \pi_{i2}^{(m)}\right)\end{aligned}$$

La segunda derivada es:

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\gamma}^{(m)}|y_{i2}, D_{i2}^{(m)})}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'} &= \sum_{i=1}^{n} -\mathbf{z}_{i2}'\left[\frac{\exp\left(\mathbf{z}_{i2}'\boldsymbol{\gamma}^{(m)}\right)}{[1+\exp\left(\mathbf{z}_{i2}'\boldsymbol{\gamma}^{(m)}\right)]^2}\right]\mathbf{z}_{i2}\\ &= \sum_{i=1}^{n} -\mathbf{z}_{i2}'\left[\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)})\right]\mathbf{z}_{i2}\end{aligned}$$

Al igual que en el tiempo 1, se puede deducir el peso correspondiente a cada dato faltante a partir de la log-verosimilitud, utilizando la propuesta de Ayala y Melo (2007) o reemplazar las funciones de estimación, conforme a Hall y Shen (2009), para el paso M con funciones que asignan pesos a los datos faltantes que conllevan a iguales resultados. Para facilitar la notación se utiliza la propuesta de Hall y Shen (2009). Luego, la ecuación a resolver es:

$$\frac{\partial^2 \ell(\boldsymbol{\gamma}^{(m)}|y_{i2}, D_{i2}^{(m)})}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'} = \sum_{i=1}^{n} -\mathbf{z}_{i2}'\left[\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)})\right] w_{i2(k)}^{(m)}\mathbf{z}_{i2}$$

donde $w_{i2(k)}^{(m)}$ es el peso correspondiente a la i-ésima observación en la m-ésima iteracción en el k-ésimo patrón de respuesta faltante en el tiempo 2, estos valores se especifican de la misma manera que en el tiempo 1, es decir:

$$w_{i2(k)}^{(m)} = \begin{cases} 1 & \text{si } k=0\\ \pi_{i2}^{(m)} + (1-\pi_{i2}^{(m)})\exp\left(-\lambda_{i2}^{(m)}\right) & \text{si } k=1\\ \dfrac{(1-\pi_{i2}^{(m)})\exp\left(-\lambda_{i2}^{(m)}\right)(\lambda_{i2}^{(m)})^{k-1}}{(k-1)!} & \text{si } k>1\end{cases}$$

Para la estimación de $\boldsymbol{\gamma}$ se hace uso del método Scoring-Fisher. De donde se tiene que:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + [\Im_z^{(m)}]^{-1}\boldsymbol{U}_z^{(m)} \tag{3-36}$$

donde

$$\Im_z^{(m)} = E\left[-\frac{\partial^2 \ell(\boldsymbol{\gamma}^{(m)}|y_{i2}, D_{i2}^{(m)})}{\partial \boldsymbol{\gamma}'\partial \boldsymbol{\gamma}}\right] = \sum_{i=1}^{n}\mathbf{z}_{i2}'\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)})w_{i2(k)}^{(m)}\mathbf{z}_{i2}$$

De esta forma, la matriz de información de Fisher queda definida en forma matricial como:

$$\mathfrak{S}_z^{(m)} = \mathbf{Z}_2' \boldsymbol{M}_{2z}^{(m)} \boldsymbol{W}_2^{(m)} \mathbf{Z}_2 \tag{3-37}$$

donde $\boldsymbol{M}_{2z}^{(m)} = diag(\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)}))$ y $\boldsymbol{W}_2^{(m)} = diag(w_{i2(k)}^{(m)})$ son matrices de tamaño $n \times n$ y Z_2 es la matriz de covariables asociadas al modelo cero inflado en el tiempo 2 de tamaño $n \times (p+1)$. Además, se tiene que:

$$\boldsymbol{U}_z^{(m)} = \sum_{i=1}^{n} \mathbf{z}_{i2}' \left(D_{i2}^{(m)} - \pi_{i2}^{(m)} \right) \boldsymbol{w}_{i2(k)}^{(m)}$$

Al multiplicar el numerador y denominador por la misma expresión, se obtiene:

$$\boldsymbol{U}_z^{(m)} = \sum_{i=1}^{n} \mathbf{z}_{i2}' \left(D_{i2}^{(m)} - \pi_{i2}^{(m)} \right) \cdot \boldsymbol{w}_{i2(k)}^{(m)} \frac{\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)})}{\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)})} \tag{3-38}$$

Definiendo:

$$\boldsymbol{v}_{z2}^{(m)} = \left(\ldots, \left(D_{i2}^{(m)} - \pi_{i2}^{(m)} \right) \frac{1}{\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)})}, \ldots \right)'$$

un vector de tamaño $n \times 1$. Las definiciones anteriores se sustituyen en (3-38) y expresando de manera matricial se tiene que:

$$\boldsymbol{U}_z^{(m)} = \mathbf{Z}_2' \boldsymbol{M}_{2z}^{(m)} \boldsymbol{W}_2^{(m)} \boldsymbol{v}_{z2}^{(m)} \tag{3-39}$$

Reemplazando (3-39) y (3-37) en (3-36) se llega a:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + \left[\mathbf{Z}_2' \boldsymbol{M}_{2z}^{(m)} \boldsymbol{W}_2^{(m)} \mathbf{Z}_2 \right]^{-1} \cdot \mathbf{Z}_2' \boldsymbol{M}_{2z}^{(m)} \boldsymbol{W}_2^{(m)} \boldsymbol{v}_{z2}^{(m)}$$

que permite la estimación de $\boldsymbol{\gamma}$.

De manera similar, se deduce a partir de la log-verosimilitud para $\boldsymbol{\beta}$, luego se tiene que $\boldsymbol{M}_{x2}^{(m)} = diag(\lambda_{i2}^{(m)})$ es una matriz de tamaño $n \times n$ y

$$\boldsymbol{v}_{x2}^{(m)} = \left(\ldots, (y_{i2} - \lambda_{i2}^{(m)}) \frac{1}{\lambda_{i2}^{(m)}}, \ldots \right)'$$

es un vector de tamaño $n \times 1$. Por el algoritmo Fisher-Scoring en forma matricial se tiene que:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}_2' \boldsymbol{M}_{x2}^{(m)} \boldsymbol{W}_2^{(m)} \mathbf{X}_2 \right]^{-1} \cdot \mathbf{X}_2' \boldsymbol{M}_{x2}^{(m)} \boldsymbol{W}_2^{(m)} \boldsymbol{v}_{x2}^{(m)}$$

es la expresión para la estimación de $\boldsymbol{\beta}$, donde $\mathbf{X}_2$ es la matriz de covariables de tamaño $n \times p$.

Paso 2

Estimación e imputación de datos faltantes en el tiempo 2. Se realiza el algoritmo EM para los datos en el segundo tiempo teniendo en cuenta como estimador inicial de los parámetros los obtenidos en el paso 2 del tiempo 1.

Paso E. Se realiza el mismo procedimiento dado en el paso 2 del tiempo 1, para obtener las estimaciones de los datos faltantes en el tiempo 2. El modelo a tener en cuenta es:

$$E[g(\mathbf{y}_2)|\mathbf{X}_2, \mathbf{Z}_2] = \begin{pmatrix} \mathbf{X}_2 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}_2 \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} \mathbf{A}_2 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{0} \\ \mathbf{B}_2 \end{pmatrix} \boldsymbol{\tau}$$

Los parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\tau}$ son los coeficientes estimados para las covariables de los valores faltantes y $\mathbf{A}_2$ y $\mathbf{B}_2$ corresponden a las matrices de las covariables de valor faltante del segundo tiempo para el modelo Poisson y cero inflado respectivamente. Al realizar la misma deducción que el paso 2 del tiempo 1 se llega a:

$$\widehat{\boldsymbol{\alpha}} = \mathbf{X}_{2(falt)}\widehat{\beta}$$
$$\widehat{\boldsymbol{\tau}} = \mathbf{Z}_{2(falt)}\widehat{\gamma}$$

Las demás características se deducen al igual que en el paso 2 del tiempo 1. Es decir, teniendo en cuenta que la distribución de la variable respuesta es Poisson con exceso de ceros se tiene que los modelos a estimar son:

$$\widehat{\alpha}_i^{(m)} = \mathbf{x}'_{i2(falt)}\widehat{\boldsymbol{\beta}}^{(m)} = \ln[\widehat{\lambda}_{i2(falt)}^{(m)}] \qquad \text{si } y_{i2} > 0$$
$$\widehat{\tau}_i^{(m)} = \mathbf{z}'_{i2(falt)}\widehat{\boldsymbol{\gamma}}^{(m)} = \ln\left[\frac{\widehat{\pi}_{i2(falt)}^{(m)}}{1 - \widehat{\pi}_{i2(falt)}^{(m)}}\right] \qquad \text{si } y_{i2} = 0$$

Por lo tanto,

$$\widehat{\lambda}_{i2(falt)}^{(m)} = \exp\left(\mathbf{x}'_{i2(falt)}\widehat{\boldsymbol{\beta}}^{(m)}\right) \qquad \text{si } y_{i2} > 0$$
$$\widehat{\pi}_{i2(falt)}^{(m)} = \frac{\exp\left(\mathbf{x}'_{i2(falt)}\widehat{\boldsymbol{\gamma}}^{(m)}\right)}{1 + \exp\left(\mathbf{z}'_{i2(falt)}\widehat{\boldsymbol{\gamma}}^{(m)}\right)} \qquad \text{si } y_{i2} = 0$$

donde $\hat{\lambda}_{i2}^{(m)}$ y $\hat{\pi}_{i2}^{(m)}$ hacen referencia al valor de la media correspondiente al individuo i en el tiempo 2.

El criterio de selección es igual al del paso 2 del tiempo 1. Es decir, dada la m-ésima iteración se tiene que:

$$\widehat{y}_{i2(falt)} = \begin{cases} 0 & \text{si } \widehat{\pi}^{(m)}_{i2(falt)} > p_o \\ \widehat{\lambda}^{(m)}_{i2(falt)} & \text{si } \widehat{\pi}^{(m)}_{i2(falt)} \leq p_o \end{cases}$$

Paso M. Imputadas las observaciones se procede a la maximización partiendo del conjunto de datos completos, teniendo en cuenta el método Fisher-Scoring, con el estimador inicial dado en el paso 1. Se itera entre el paso E y M hasta lograr la convergencia. Las expresiones son:

$$E\left[D^{(m)}_{i2}|y_{i2}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\beta}^{(m)}\right] = \begin{cases} \dfrac{1}{1+\exp\left(-\mathbf{z}'_{i2}\boldsymbol{\gamma}^{(m)} - \exp\left(\mathbf{x}'_{i1}\boldsymbol{\beta}^{(m)}\right)\right)} & \text{si } y_{i2} = 0 \\ 0 & \text{si } y_{i2} > 0 \end{cases}$$

Definiendo:

$$\boldsymbol{v}^{(m)}_{z2} = \left(\ldots, \left(D^{(m)}_{i2} - \pi^{(m)}_{i2}\right)\frac{1}{\pi^{(m)}_{i2}(1-\pi^{(m)}_{i2})}, \ldots\right)'$$
$$\boldsymbol{M}^{(m)}_{2z} = diag(\pi^{(m)}_{i2}(1-\pi^{(m)}_{i2}))$$

donde $\boldsymbol{M}_{2z}$, es una matriz de tamaño de $n \times n$ y $\boldsymbol{v}_{z2}$ es un vector de $n \times 1$. El algoritmo Fisher-Scoring para $\boldsymbol{\gamma}$ es:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + \left[\mathbf{Z}'_2\boldsymbol{M}^{(m)}_{2z}\mathbf{Z}_2\right]^{-1} \cdot \left(\mathbf{Z}'_2\boldsymbol{M}^{(m)}_{2z}\boldsymbol{v}^{(m)}_{z2}\right)$$

donde $\mathbf{Z}_2$ una matriz de covariables de $n \times (p+1)$. De manera similar, para $\boldsymbol{\beta}$ las expresiones son:

$$\boldsymbol{M}^{(m)}_{x2} = diag(\lambda^{(m)}_{i2})$$
$$\boldsymbol{v}^{(m)}_{x2} = \left(\ldots, (y_{i2} - \lambda^{(m)}_{i2})\frac{1}{\lambda^{(m)}_{i2}}, \ldots\right)'$$

donde $\boldsymbol{M}_{x2}$ es de tamaño $n \times n$ y $\boldsymbol{v}_{x2}$ de $n \times 1$. Por el algoritmo Fisher-Scoring y expresando en matrices se tiene que:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}'_2\boldsymbol{M}^{(m)}_{x2}\mathbf{X}_2\right]^{-1} \cdot \mathbf{X}'_2\boldsymbol{M}^{(m)}_{x2}\boldsymbol{v}^{(m)}_{x2}$$

donde $\mathbf{X}_2$ es una matriz de covariables de $n \times (p+1)$.

3.1.3. Tiempo 3

Estimación del vector de parámetros máximo verosímiles en la ocasión 3 por medio del algoritmo EM por ponderaciones.
Los vectores y_{i1} y y_{i2} contienen las respuestas en el tiempo 1 y 2 para evitar problemas de multicolinealidad, por la medición longitudinal, se emplea el método de componentes principales, con el fin de generar un nuevo conjunto de covariables ortogonales representativas. La matriz de covariables está dada por:

$$C_2 = [y_{i1}, y_{i2}]A'$$

donde $A = [a_1, a_2]$ es la matriz de vectores propios correspondientes a los valores propios de norma 1. El vector propio a_1 se obtiene de solucionar:

$$|S - \Delta_1| \, I a_1 = 0$$

donde S es la matriz de varianzas-covarianzas de $[y_{i1}, y_{i2}]$ y Δ_1 es el valor propio más grande de S y I es una matriz identidad. La extracción de la segundo componente principal se realiza de igual manera.

Paso 1

Se lleva a cabo la estimación del vector de parámetros máximo verosímiles en la tercera ocasión $t = 3$ igual que en el tiempo 1, con el empleo del algoritmo EM.
Las funciones de enlace para los parámetros son:

$$\begin{aligned} \ln[\lambda_{i3}] &= \mathbf{x}'_{i3}\boldsymbol{\beta} \text{ si } y_{i3} > 0 \\ \ln\left[\frac{\pi_{i3}}{1-\pi_{i3}}\right] &= \mathbf{z}'_{i3}\boldsymbol{\gamma} \text{ si } y_{i3} = 0 \end{aligned} \tag{3-40}$$

donde $\mathbf{X}_3$ y $\mathbf{Z}_3$ son matrices de tamaño $n \times (p+1)$, donde p columnas hace referencias a las covariables que explican el comportamiento de la variable respuesta y la última columna es el primer vector propio del análisis de componentes principales realizado con los valores de y_{i1} y y_{i2}. Se define, para la estimación de los parámetros en el tiempo 3, una variable indicadora D_{i3} al igual que el tiempo 1, que permite particionar la log-verosimilitud en una parte para el modelo Poisson $\ell(D_{i3}, \boldsymbol{\gamma}, \boldsymbol{\delta}, y_{i3})$ y otra para el modelo de exceso de ceros $\ell(D_{i3}, \boldsymbol{\beta}, \boldsymbol{\phi}, y_{i3})$ para ser maximizadas de manera separada. Se define $y_{i3} = (y_{i3(obs)}, y_{i3(falt)})$ como el vector que hace referencia a valores observados y faltantes.

Se puede deducir a partir de la propuesta de Ayala y Melo (2007) los pesos para el tiempo 3 que quedan definidos como:

$$w_{i3(k)}^{(m)} = \begin{cases} 1 & \text{si } k = 0 \\ \pi_{i3}^{(m)} + (1 - \pi_{i3}^{(m)}) \exp\left(-\lambda_{i3}^{(m)}\right) & \text{si } k = 1 \\ \dfrac{(1 - \pi_{i3}^{(m)}) \exp\left(-\lambda_{i3}^{(m)}\right)(\lambda_{i3}^{(m)})^{k-1}}{(k-1)!} & \text{si } k > 1 \end{cases}$$

Las matrices para la estimación de los parámetros del modelo en el tiempo 3 se deducen exactamente igual que en el paso 1 del tiempo 1 y se expresan como:

$$\boldsymbol{M}_{3z}^{(m)} = diag(\pi_{i3}^{(m)}(1 - \pi_{i3}^{(m)}))$$
$$\boldsymbol{W}_{3}^{(m)} = diag(w_{i3(k)}^{(m)})$$

de tamaños $n \times n$. El vector

$$\boldsymbol{v}_{z3}^{(m)} = \left(\ldots, \left(D_{i3}^{(m)} - \pi_{i3}^{(m)}\right) \frac{1}{\pi_{i3}^{(m)}(1 - \pi_{i3}^{(m)})}, \ldots \right)'$$

un vector de $n \times 1$. Reemplazando en el algoritmo Fisher-Scoring se tiene que:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + \left[\mathbf{Z}_3' \boldsymbol{M}_{3z}^{(m)} \boldsymbol{W}_{3}^{(m)} \mathbf{Z}_3\right]^{-1} \cdot \mathbf{Z}_3' \boldsymbol{M}_{3z}^{(m)} \boldsymbol{W}_{3}^{(m)} \boldsymbol{v}_{z3}^{(m)}$$

con $\mathbf{Z}_3$ la matriz de covariables de tamaño $n \times (p+1)$. De manera similar, se deduce a partir de la log-verosimilitud para $\boldsymbol{\beta}$ que $\boldsymbol{M}_{x3}^{(m)} = diag(\lambda_{i3}^{(m)})$ una matriz de tamaño $n \times n$ y el vector:

$$\boldsymbol{v}_{x3}^{(m)} = \left(\ldots, (y_{i3} - \lambda_{i3}^{(m)}) \frac{1}{\lambda_{i3}^{(m)}}, \ldots \right)'$$

de tamaño $n \times 1$. Por el algoritmo Fisher-Scoring y expresando en matrices se tiene que:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}_3' \boldsymbol{M}_{x3}^{(m)} \boldsymbol{W}_{3}^{(m)} \mathbf{X}_3\right]^{-1} \cdot \mathbf{X}_3' \boldsymbol{M}_{x3}^{(m)} \boldsymbol{W}_{3}^{(m)} \boldsymbol{v}_{x3}^{(m)}$$

con $\mathbf{X}_3$ la matriz de covariables de tamaño $n \times (p+1)$.

Paso 2

Estimación e imputación de datos faltantes en el tiempo 3. Se realiza el algoritmo EM para los datos en el segundo tiempo teniendo tomando como estimadores iniciales de los parámetros los obtenidos en el paso 2 del tiempo 3.

Paso E. Se realiza el procedimiento mostrado en el paso 2 del tiempo 1 para obtener las estimaciones de los datos faltantes en el tiempo 3. El modelo a tener en cuenta es:

$$E[g(\mathbf{y}_3)|\mathbf{X}_3, \mathbf{Z}_3] = \begin{pmatrix} \mathbf{X}_3 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}_3 \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} \mathbf{A}_3 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{0} \\ \mathbf{B}_3 \end{pmatrix} \boldsymbol{\tau}$$

donde $\boldsymbol{X}_3$ y $\boldsymbol{Z}_3$ son matrices de covariables en el tiempo 3. Los coeficientes: $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son los parámetros asociados a la regresión de las covariables. Los parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\tau}$ son los coeficientes estimados para las covariables de los valores faltantes y $\mathbf{A}_3$ y $\mathbf{B}_3$ son las matrices de las covariables de valor faltante del tercer tiempo para el modelo Poisson y cero inflado respectivamente. Al realizar la misma deducción que en el paso 2 del tiempo 1 se llega a:

$$\widehat{\boldsymbol{\alpha}} = \mathbf{X}_{3(falt)}\widehat{\beta}$$
$$\widehat{\boldsymbol{\tau}} = \mathbf{Z}_{3(falt)}\widehat{\gamma}$$

Las demás características se deducen al igual que en el paso 2 del tiempo 1. Es decir, teniendo en cuenta que la distribución de la variable respuesta es Poisson con exceso de ceros se tiene que los modelos a estimar son:

$$\widehat{\alpha}_i^{(m)} = \mathbf{x}'_{i3(falt)}\widehat{\boldsymbol{\beta}}^{(m)} = \ln[\widehat{\lambda}_{i3(falt)}^{(m)}] \qquad \text{si } y_{i3} > 0$$
$$\widehat{\tau}_i^{(m)} = \mathbf{z}'_{i3(falt)}\widehat{\boldsymbol{\gamma}}^{(m)} = \ln\left[\frac{\widehat{\pi}_{i3(falt)}^{(m)}}{1-\widehat{\pi}_{i3(falt)}^{(m)}}\right] \qquad \text{si } y_{i3} = 0$$

Por lo tanto,

$$\widehat{\lambda}_{i3(falt)}^{(m)} = \exp\left(\mathbf{x}'_{i3(falt)}\widehat{\boldsymbol{\beta}}^{(m)}\right) \qquad \text{si } y_{i3} > 0$$
$$\widehat{\pi}_{i3(falt)}^{(m)} = \frac{\exp\left(\mathbf{z}'_{i3(falt)}\widehat{\boldsymbol{\gamma}}^{(m)}\right)}{1+\exp\left(\mathbf{z}'_{i3(falt)}\widehat{\boldsymbol{\gamma}}^{(m)}\right)} \qquad \text{si } y_{i3} = 0$$

donde $\hat{\lambda}_{i3}^{(m)}$ y $\hat{\pi}_{i3}^{(m)}$ hacen referencia al valor de la media correspondiente al individuo i en el tiempo 3. El criterio de selección es igual al del paso 2 del tiempo 1. Es decir, dada la m-ésima iteración se tiene que:

$$\widehat{y}_{i3(falt)} = \begin{cases} 0 & \text{si } \widehat{\pi}_{i3(falt)}^{(m)} > p_o \\ \widehat{\lambda}_{i3(falt)}^{(m)} & \text{si } \widehat{\pi}_{i3(falt)}^{(m)} \leq p_o \end{cases}$$

Paso M. Imputadas las observaciones se procede a la maximización partiendo del conjunto de datos completos, teniendo en cuenta el método Fisher-Scoring, con el estimador inicial dado en el paso 1. Se itera entre el paso E y M hasta lograr la convergencia. Las expresiones son:

$$E\left[D_{i3}^{(m)}|y_{i3},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}\right]=\begin{cases}\dfrac{1}{1+\exp\left(-\mathbf{z}_{i3}'\boldsymbol{\gamma}^{(m)}-\exp\left(\mathbf{x}_{i3}'\boldsymbol{\beta}^{(m)}\right)\right)} & \text{si } y_{i3}=0\\ 0 & \text{si } y_{i3}>0\end{cases}$$

Definiendo $\boldsymbol{M}_{3z}^{(m)}=diag(\pi_{i3}^{(m)}(1-\pi_{i3}^{(m)}))$ una matriz de tamaño $n\times n$ y

$$\boldsymbol{v}_{z3}^{(m)}=\left(\ldots,\left(D_{i3}^{(m)}-\pi_{i3}^{(m)}\right)\frac{1}{\pi_{i3}^{(m)}(1-\pi_{i3}^{(m)})},\ldots\right)'$$

un vector de $n\times 1$. El algoritmo Fisher-Scoring para $\boldsymbol{\gamma}$ es:

$$\boldsymbol{\gamma}^{(m+1)}=\boldsymbol{\gamma}^{(m)}+\left[\mathbf{Z}_3'\boldsymbol{M}_{3z}^{(m)}\mathbf{Z}_3\right]^{-1}\cdot\mathbf{Z}_3'\boldsymbol{M}_{3z}^{(m)}\boldsymbol{v}_{z3}^{(m)}$$

donde $\mathbf{Z}_3$ es una matriz de covariables de tamaño $n\times(p+1)$. De manera similar, para $\boldsymbol{\beta}$ se tiene que $\boldsymbol{M}_{x3}^{(m)}=diag(\lambda_{i3}^{(m)})$ matriz de $n\times n$ y

$$\boldsymbol{v}_{x3}^{(m)}=\left(\ldots,(y_{i3}-\lambda_{i3}^{(m)})\frac{1}{\lambda_{i3}^{(m)}},\ldots\right)'$$

vector de tamaño $n\times 1$. Por el algoritmo Fisher-Scoring y expresando en matrices se tiene que:

$$\boldsymbol{\beta}^{(m+1)}=\boldsymbol{\beta}^{(m)}+\left[\mathbf{X}_3'\boldsymbol{M}_{x3}^{(m)}\mathbf{X}_3\right]^{-1}\cdot\mathbf{X}_3'\boldsymbol{M}_{x3}^{(m)}\boldsymbol{v}_{x3}^{(m)}$$

donde $\mathbf{X}_3$ matriz de covariables de tamaño $n\times(p+1)$.

3.1.4. Tiempo t

Estimación del vector de parámetros máximo verosímiles en el tiempo t por medio del algoritmo EM por ponderaciones.
Se tiene que $y_{i1},y_{i2},\ldots,y_{i(t-1)}$ son los vectores respuesta del tiempo 1 hasta el tiempo t y para evitar problemas de multicolinealidad, por la medición longitudinal, se emplea el método de componentes principales, con el fin de generar un nuevo conjunto de covariables ortogonales representativas. La matriz de covariables está dada por:

$$C_{t-1}=[y_{i1},y_{i2},\ldots,y_{i(t-1)}]A'$$

donde $A=[a_1,a_2,\ldots,a_{t-1}]$ es la matriz de vectores propios correspondientes a los valores propios de norma 1.

Paso 1

Se define C_{t-1} como el vector propio asociado a las respuestas en los tiempos $t-1$. La estimación del vector de parámetros máximo verosímiles en el tiempo t se lleva a cabo al igual que en el tiempo 1 haciendo uso del algoritmo EM. Las funciones de enlace para los parámetros son:

$$\begin{aligned} \ln[\lambda_{it}] &= \mathbf{x}'_{it}\boldsymbol{\beta} \text{ si } y_{it} > 0 \\ \ln\left[\frac{\pi_{it}}{1-\pi_{it}}\right] &= \mathbf{z}'_{it}\boldsymbol{\gamma} \text{ si } y_{it} = 0 \end{aligned} \tag{3-41}$$

donde $\mathbf{X}_t$ y $\mathbf{Z}_t$ son matrices de tamaño $n \times (p+1)$, donde p columnas hacen referencias a las covariables que explican el comportamiento de la variable en el tiempo t y la última columna es el primer vector propio del análisis de componentes principales realizado con los valores de $y_{i1}, y_{i2}, \ldots, y_{i(t-1)}$. Se define, para la estimación de los parámetros en el tiempo t, una variable indicadora D_{it} al igual que el tiempo 1, que permite particionar la log-verosimilitud en una parte para el modelo Poisson $\ell(D_{it}, \boldsymbol{\beta}, y_{it})$ y otra para el modelo de exceso de ceros $\ell(D_{it}, \boldsymbol{\gamma}, y_{it})$ para ser maximizadas de manera separada. Se define $y_{it} = (y_{it(obs)}, y_{it(falt)})$ como el vector que hace referencia a valores observados y faltantes. Se puede deducir a partir de la propuesta de Ayala y Melo (2007) los pesos para el tiempo t que quedan definidos como:

$$w_{it(k)}^{(m)} = \begin{cases} 1 & \text{si } k = 0 \\ \pi_{it}^{(m)} + (1-\pi_{it}^{(m)})\exp\left(-\lambda_{it}^{(m)}\right) & \text{si } k = 1 \\ \dfrac{(1-\pi_{it}^{(m)})\exp\left(-\lambda_{it}^{(m)}\right)(\lambda_{it}^{(m)})^{k-1}}{(k-1)!} & \text{si } k > 1 \end{cases}$$

Las matrices para la estimación de los parámetros del modelo en el tiempo t se deducen exactamente igual que en el paso 1 del tiempo 1. Es decir, $\boldsymbol{M}_{tz}^{(m)} = diag(\pi_{it}^{(m)}(1-\pi_{it}^{(m)})$ y $W_t^{(m)} = diag(w_{it(k)}^{(m)})$ de tamaños $n \times n$ y el vector

$$\boldsymbol{v}_{zt}^{(m)} = \left(\ldots, \left(D_{it}^{(m)} - \pi_{it}^{(m)}\right)\frac{1}{\pi_{it}^{(m)}(1-\pi_{it}^{(m)})}, \ldots\right)'$$

de tamaño $n \times 1$. Reemplazando en algoritmo Fisher-Scoring se tiene que:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + \left[\mathbf{Z}'_t\boldsymbol{M}_{tz}^{(m)}\boldsymbol{W}_t^{(m)}\mathbf{Z}_t\right]^{-1} \cdot \mathbf{Z}'_t\boldsymbol{M}_{tz}^{(m)}\boldsymbol{W}_t^{(m)}\boldsymbol{v}_{zt}^{(m)}$$

con $\mathbf{Z}_t$ de tamaño $n \times (p+1)$. De manera similar, se deduce a partir de la log-verosimilitud para $\boldsymbol{\beta}$ la matriz $\boldsymbol{M}_{xt}^{(m)} = diag(\lambda_{it}^{(m)})$ de tamaño $n \times n$ y el vector

$$\boldsymbol{v}_{xt}^{(m)} = \left(\ldots, (y_{it} - \lambda_{it}^{(m)})\frac{1}{\lambda_{it}^{(m)}}, \ldots\right)'$$

Reemplazando en el algoritmo Fisher-Scoring y expresando en matrices se tiene que:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}_t' \boldsymbol{M}_{xt}^{(m)} \boldsymbol{W}_t^{(m)} \mathbf{X}_t\right]^{-1} \cdot \mathbf{X}_t' \boldsymbol{M}_{xt}^{(m)} \boldsymbol{W}_t^{(m)} \boldsymbol{v}_{xt}^{(m)}$$

donde $\mathbf{X}_t$ es una matriz de covariables de tamaño $n \times (p+1)$.

Paso 2

Estimación e imputación de datos faltantes en el tiempo t. Se realiza el algoritmo EM para los datos en el tiempo t teniendo en cuenta como estimador de los parámetros los obtenidos en el paso 1 del tiempo $t-1$.

Paso E. Se realiza el mismo procedimiento dado en el paso 2 del tiempo 1. Los valores iniciales para los parámetros son los obtenidos en el paso 1 del tiempo $t-1$, para obtener las estimaciones de los datos faltantes en el tiempo 3 se tiene en cuenta el modelo:

$$E[g(\mathbf{y}_t)|\mathbf{X}_t, \mathbf{Z}_t] = \begin{pmatrix} \mathbf{X}_t \\ \mathbf{0} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}_t \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} \mathbf{A}_t \\ \mathbf{0} \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{0} \\ \mathbf{B}_t \end{pmatrix} \boldsymbol{\tau}$$

donde $\boldsymbol{x}_t$ y $\boldsymbol{z}_t$ son matrices de covariables en el tiempo t de tamaño $n \times (p+1)$. Los coeficientes: $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son los parámetros asociados a la regresión de las covariables. Los parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\tau}$ son los coeficientes estimados en para las covariables de los valores faltantes y $\mathbf{A}_t$ y $\mathbf{B}_t$ corresponden a las matrices que de las covariables de valor faltante del tercer tiempo para el modelo Poisson y cero inflado respectivamente. Al realizar la misma deducción que el paso 2 del tiempo 1 se llega a:

$$\widehat{\boldsymbol{\alpha}} = \mathbf{X}_{t(falt)} \widehat{\beta}$$
$$\widehat{\boldsymbol{\tau}} = \mathbf{Z}_{t(falt)} \widehat{\gamma}$$

Las demás características se deducen al igual que en el paso 2 del tiempo 1. Es decir, teniendo en cuenta que la distribución de la variable respuesta es Poisson con exceso de ceros se tiene que los modelos a estimar son:

$$\widehat{\alpha}_i^{(m)} = \mathbf{x}_{it(falt)}' \widehat{\boldsymbol{\beta}}^{(m)} = \ln[\widehat{\lambda}_{it(falt)}^{(m)}] \qquad \text{si } y_{it} > 0$$
$$\widehat{\tau}_i^{(m)} = \mathbf{z}_{it(falt)}' \widehat{\boldsymbol{\gamma}}^{(m)} = \ln\left[\frac{\widehat{\pi}_{it(falt)}^{(m)}}{1 - \widehat{\pi}_{it(falt)}^{(m)}}\right] \qquad \text{si } y_{it} = 0$$

Por lo tanto,

$$\widehat{\lambda}_{it(falt)}^{(m)} = \exp\left(\mathbf{x}_{it(falt)}' \widehat{\boldsymbol{\beta}}^{(m)}\right) \qquad \text{si } y_{it} > 0$$
$$\widehat{\pi}_{it(falt)}^{(m)} = \frac{\exp\left(\mathbf{z}_{it(falt)}' \widehat{\boldsymbol{\gamma}}^{(m)}\right)}{1 + \exp\left(\mathbf{z}_{it(falt)}' \widehat{\boldsymbol{\gamma}}^{(m)}\right)} \qquad \text{si } y_{it} = 0$$

donde $\hat{\lambda}_{it}^{(m)}$ y $\hat{\pi}_{it}^{(m)}$ hacen referencia al valor de la media correspondiente al individuo i en el tiempo t. El criterio de selección es igual al del paso 2 del tiempo 1. Es decir, dada la m-ésima iteración se tiene que:

$$\widehat{y}_{it(falt)} = \begin{cases} 0 & \text{si } \widehat{\pi}_{it(falt)}^{(m)} > p_o \\ \widehat{\lambda}_{it(falt)}^{(m)} & \text{si } \widehat{\pi}_{it(falt)}^{(m)} \leq p_o \end{cases}$$

Paso M. Imputadas las observaciones se procede a la maximización partiendo del conjunto de datos completos, teniendo en cuenta el método Fisher-Scoring, con el estimador inicial dado en el paso 1. Se repite el paso E y M hasta lograr la convergencia. Las expresiones son:

$$E\left[D_{it}^{(m)}|y_{it}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}\right] = \begin{cases} \dfrac{1}{1+\exp\left(-\mathbf{z}_{it}'\boldsymbol{\gamma}^{(m)} - \exp\left(\mathbf{x}_{it}'\boldsymbol{\beta}^{(m)}\right)\right)} & \text{si } y_{it} = 0 \\ 0 & \text{si } y_{it} > 0 \end{cases}$$

De manera similar que en el paso 2 del tiempo 1 se deduce la matriz $\boldsymbol{M}_{tz}^{(m)} = diag(\pi_{it}^{(m)}(1 - \pi_{it}^{(m)}))$ de tamaño $n \times n$ y el vector

$$\boldsymbol{v}_{zt}^{(m)} = \left(\ldots, \left(D_{it}^{(m)} - \pi_{it}^{(m)}\right)\frac{1}{\pi_{it}^{(m)}(1-\pi_{it}^{(m)})}, \ldots\right)'$$

de tamaño $n \times 1$. El algoritmo Fisher-Scoring para la estimación de $\boldsymbol{\gamma}$ es:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + \left[\mathbf{Z}_t'\boldsymbol{M}_{tz}^{(m)}\mathbf{Z}_t\right]^{-1} \cdot \mathbf{Z}_t'\boldsymbol{M}_{tz}^{(m)}\boldsymbol{v}_{zt}^{(m)}$$

con $\mathbf{Z}_t$ la matriz de convariables de tamaño $n \times (p+1)$. De manera similar, para $\boldsymbol{\beta}$ se tiene que $\boldsymbol{M}_{xt}^{(m)} = diag(\lambda_{it}^{(m)})$ de tamaño $n \times n$ y el vector

$$\boldsymbol{v}_{xt}^{(m)} = \left(\ldots, (y_{it} - \lambda_{it}^{(m)})\frac{1}{\lambda_{it}^{(m)}}, \ldots\right)'$$

de tamaño $n \times 1$. Por el algoritmo Fisher-Scoring y expresando en matrices se tiene que para estimar $\boldsymbol{\beta}$ las ecuaciones son:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}_t'\boldsymbol{M}_{xt}^{(m)}\mathbf{X}_t\right]^{-1} \cdot \mathbf{X}_t'\boldsymbol{M}_{xt}^{(m)}\boldsymbol{v}_{xt}^{(m)}$$

con $\mathbf{X}_t$ de tamaño $n \times (p+1)$.

3.2. Binomial Negativa Cero Inflada

Sean $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n$ un conjunto de variables aleatorias independientes con distribución de probabilidad Binomial negativa cero inflada. Una variable aleatoria Binomial Negativa Cero Inflada se define como:

$$y_i = \begin{cases} 0 & \text{con } \pi_i \\ \sim BinomialNegativa(\lambda_i, \alpha) & \text{con } (1-\pi_i) \end{cases}$$

La distribución de probabilidad de la variable aleatoria es:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\dfrac{\alpha}{\alpha+\lambda_i}\right)^{\alpha} & \text{si } y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma\left[y_i+\alpha\right]}{\Gamma(y_i+1)\Gamma\left[\alpha\right]}\left(\dfrac{\alpha}{\alpha+\lambda_i}\right)^{\alpha}\left(\dfrac{\lambda_i}{\alpha+\lambda_i}\right)^{y_i} & \text{si } y_i > 0 \end{cases}$$

donde $E[Y_i] = (1-\pi_i)\lambda_i$ y $Var[Y_i] = (1-\pi_i)\lambda_i(1+\pi_i\lambda_i+\alpha\lambda_i)$. π_i y λ_i se pueden modelar usando un modelo *logit* y un modelo ln, respectivamente. Esto es:

$$\begin{aligned} \ln[\lambda_i] &= \mathbf{x}_i'\boldsymbol{\beta} \\ \ln\left[\frac{\pi_i}{1-\pi_i}\right] &= \mathbf{z}_i'\boldsymbol{\gamma} \end{aligned}$$

Despejando en cada una de las ecuaciones se tiene que:

$$\begin{aligned} \pi_i &= \frac{\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)}{1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)} \\ \lambda_i &= \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \end{aligned} \tag{3-42}$$

Al igual que en el modelo Poisson Cero Inflado se asume que π_i y λ_i son independientes.

3.2.1. Tiempo 1

Estimación del vector de parámetros en la primera ocasión con el algoritmo EM con ponderaciones. Inicialmente se tiene que la función de log-verosimilitud se define como:

$$\begin{aligned} \ln \ell &= \ln \prod_{i=1}^{n} P[Y_{i1} = y_{i1}] \\ &= \ln\left[\prod_{Y_{i1}=0}\left(\pi_{i1} + (1-\pi_{i1})\left(\frac{\alpha}{\alpha+\lambda_{i1}}\right)^{\alpha}\right)\right. \\ &\left.\times \prod_{Y_{i1}\neq 0}\left((1-\pi_{i1})\frac{\Gamma\left(y_{i1}+\alpha\right)}{\Gamma(y_{i1}+1)\Gamma\left(\alpha\right)}\left(\frac{\alpha}{\alpha+\lambda_{i1}}\right)^{\alpha}\left[\frac{\lambda_{i1}}{\alpha+\lambda_{i1}}\right]^{y_{ii}}\right)\right] \end{aligned}$$

$$
\begin{aligned}
\ln \ell &= \ln \left[\prod_{Y_{i1}=0} \left(\pi_{i1} + (1-\pi_{i1}) \left(\frac{\alpha}{\alpha+\lambda_{i1}} \right)^{\alpha} \right) \right] \\
&+ \ln \left[\prod_{Y_{i1}\neq 0} \left((1-\pi_i) \frac{\Gamma\left(y_{i1}+\alpha\right)}{\Gamma(y_{i1}+1)\Gamma\left(\alpha\right)} \left(\frac{\alpha}{\alpha+\lambda_{i1}} \right)^{\alpha} \left[\frac{\lambda_{i1}}{\alpha+\lambda_{i1}} \right]^{y_{i1}} \right) \right] \\
&= \sum_{Y_{i1}=0} \ln \left[\pi_{i1} + (1-\pi_{i1}) \left(\frac{\alpha}{\alpha+\lambda_{i1}} \right)^{\alpha} \right] \\
&+ \sum_{Y_{i1}\neq 0} \ln \left[(1-\pi_{i1}) \frac{\Gamma\left(y_{i1}+\alpha\right)}{\Gamma(y_{i1}+1)\Gamma\left(\alpha\right)} \left(\frac{\alpha}{\alpha+\lambda_{i1}} \right)^{\alpha} \left[\frac{\lambda_{i1}}{\alpha+\lambda_{i1}} \right]^{y_{i1}} \right]
\end{aligned} \tag{3-43}
$$

Ahora, se define una función indicadora que permite particionar la verosimilitud en la parte Binomial Negativa y el modelo de ceros como:

$$
D_{i1} = \begin{cases} 1 & \text{si } y_{i1} = 0 \\ 0 & \text{si } y_{i1} > 0 \end{cases}
$$

Luego, la log-verosimilitud queda definida como:

$$
\begin{aligned}
\ln \ell &= \sum_{i=1}^{n} D_{i1} \ln \left[\pi_{i1} + (1-\pi_{i1}) \left(\frac{\alpha}{\alpha+\lambda_{i1}} \right)^{\alpha} \right] \\
&+ \sum_{i=1}^{n} (1-D_{i1}) \ln \left[(1-\pi_{i1}) \frac{\Gamma\left(y_{i1}+\alpha\right)}{\Gamma(y_{i1}+1)\Gamma\left(\alpha\right)} \left(\frac{\alpha}{\alpha+\lambda_{i1}} \right)^{\alpha} \left[\frac{\lambda_{i1}}{\alpha+\lambda_{i1}} \right]^{y_{i1}} \right] \\
\ln \ell &= \sum_{i=1}^{n} D_{i1} \ln \left[\frac{\exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right)}{1+\exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right)} + \frac{1}{1+\exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right)} \left(\frac{\alpha}{\alpha+\exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)} \right)^{\alpha} \right] \\
&+ \sum_{i=1}^{n} (1-D_{i1}) \left\{ \ln \left(\frac{1}{1+\exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right)} \right) + \ln \left[\frac{\Gamma\left(y_{i1}+\alpha\right)}{\Gamma(y_{i1}+1)\Gamma\left(\alpha\right)} \right] \right\} \\
&+ \sum_{i=1}^{n} (1-D_{i1}) \left[\alpha \ln \left(\frac{\alpha}{\alpha+\exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)} \right) + y_{i1} \ln \left[\frac{\exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)}{\alpha+\exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)} \right] \right]
\end{aligned}
$$

$$\begin{aligned}\ln \ell = &\sum_{i=1}^{n} D_{i1} \left\{ \ln \left[\exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right) + \left(\frac{\alpha}{\alpha + \exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)} \right)^{\alpha} \right] - \ln\left[1 + \exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right)\right] \right\} \\ &+ \sum_{i=1}^{n} (1 - D_{i1}) \left\{ \ln[\Gamma(y_{i1} + \alpha)] - \ln[\Gamma(\alpha)] - \ln[\Gamma(y_{i1} + 1)] \right\} \\ &+ \sum_{i=1}^{n} (1 - D_{i1}) \left\{ \alpha \ln[\alpha] + y_{i1}\boldsymbol{x}_{i1}'\boldsymbol{\beta} - (\alpha + y_{i1}) \ln[\alpha + \exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)] \right\}\end{aligned} \tag{3-44}$$

La maximización de la log-verosimilitud es complicada por el término:

$$\ln \left[\exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right) + \left(\frac{\alpha}{\alpha + \exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)} \right)^{\alpha} \right]$$

Esto debido a que involucra los parámetros $\boldsymbol{\gamma}$ y $\boldsymbol{\beta}$ que se buscan estimar. Ahora, teniendo en cuenta la función indicadora $D_{i1} = 0$ entonces:

$$D_{i1} \ln \left[\exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right) + \left(\frac{\alpha}{\alpha + \exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)} \right)^{\alpha} \right] = 0$$

La función indicadora es $D_{i1} = 1$ cuando $y_{i1} = 0$, luego el modelo busca estimar la parte del modelo que es igual a cero o estado perfecto, y por tanto, las covariables asociadas al modelo Binomial Negativo no se tienen en cuenta en la estimación. El término que involucra los dos parámetros que se buscan estimar queda reducido a:

$$D_{i1} \ln[\exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}\right)] = D_{i1}\mathbf{z}_{i1}'\boldsymbol{\gamma} \tag{3-45}$$

Entonces, al sustituir (3-45) en (3-44) la log-verosimilitud queda reducida a:

$$\begin{aligned}\ln \ell = &\sum_{i=1}^{n} \left(D_{i1}\boldsymbol{z}_{i1}'\boldsymbol{\gamma} - \ln\left[1 + \exp\left(\boldsymbol{z}_{i1}'\boldsymbol{\gamma}\right)\right] \right) \\ &+ \sum_{i=1}^{n} (1 - D_{i1}) \left\{ \ln[\Gamma(y_{i1} + \alpha)] - \ln[\Gamma(\alpha)] - \ln[\Gamma(y_{i1} + 1)] \right\} \\ &+ \sum_{i=1}^{n} (1 - D_{i1}) \left\{ \alpha \ln[\alpha] + y_{i1}\boldsymbol{x}_{i1}'\boldsymbol{\beta} - (\alpha + y_{i1}) \ln[\alpha + \exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}\right)] \right\}\end{aligned} \tag{3-46}$$

que puede ser maximizada de manera sencilla por que:

$$\ln \ell = \ell(D_{i1}, \boldsymbol{\gamma}, y_{i1}) + \ell(D_{i1}, \boldsymbol{\beta}, \alpha, y_{i1}) \tag{3-47}$$

donde $\ell(D_{i1}, \boldsymbol{\gamma}, y_{i1})$ puede ser maximizada de manera sencilla haciendo uso del algoritmo Fisher-Scoring y $\ell(D_{i1}, \boldsymbol{\beta}, y_{i1})$ corresponde a la log-verosimilitud de una regresión binomial negativa con una ponderación $(1 - D_{i1})$. Estás dos expresiones pueden ser maximizadas de manera separada, haciendo uso del algoritmo EM de manera iterativa alternando entre la estimación inicial de $D_{i1}^{(0)}$ dada la esperanza condicional bajo unos parámetros iniciales $(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)}, \alpha^{(0)})$. Luego, con $D_{i1}^{(0)}$ estimado se maximiza para $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\gamma}^{(1)}$ y $\alpha^{(1)}$ y se procede de manera iterativa hasta la m-ésima iteracción para llega a la convergencia.

Paso 1

Se encuentran los valores estimador para $\boldsymbol{\gamma}$ y $\boldsymbol{\beta}$ haciendo uso del algoritmo EM para los datos del tiempo 1.

Paso E. Se estima D_{i1} dada la media condicional $D_{i1}^{(m)}$ bajo unos valores iniciales para $\boldsymbol{\beta}^{(m)}$ y $\boldsymbol{\gamma}^{(m)}$, esto se expresa como:

$$\begin{aligned} E[D_{i1}^{(m)}|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] &= \sum_{i=1}^{1} D_{i1}^{(m)} \cdot P[D_{i1}^{(m)}|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] \\ &= 1 \cdot P[D_{i1}^{(m)}=1|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] + 0 \cdot P[D_{i1}^{(m)}=0|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] \\ &= P[D_{i1}^{(m)}=1|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] \end{aligned}$$

Por el teorema de bayes se tiene que:

$$P[D_{i1}^{(m)}=1|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] = \frac{P[y_{i1}=0|D_{i1}^{(m)}=1,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] \cdot P[D_{i1}^{(m)}=1]}{P[y_{i1}=0|D_{i1}^{(m)}=1,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] \cdot P[D_{i1}^{(m)}=1] + P[y_{i1}\neq 0|D_{i1}^{(m)}=0,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] \cdot P[D_{i1}^{(m)}=0]}$$

Por la función indicadora se tiene que $P[D_{i1}^{(m)}=1]=\pi_{i1}^{(m)}$ y $P[D_{i1}^{(m)}=0]=1-\pi_{i1}^{(m)}$, al reemplazar en la ecuación anterior se tiene que:

$$\begin{aligned} E[D_{i1}^{(m)}|y_{i1},\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)}] &= \begin{cases} \dfrac{1 \cdot \pi_{i1}^{(m)}}{1 \cdot \pi_{i1}^{(m)} + \cdot(1-\pi_{i1}^{(m)}) \left(\dfrac{\alpha}{\alpha+\lambda_{i1}^{(m)}}\right)^{\alpha}} & \text{si } y_{i1}=0 \\ \dfrac{0 \cdot \pi_{i1}^{(m)}}{0 \cdot \pi_{i1}^{(m)} + (1-\pi_{i1}^{(m)}) \cdot Binom(\lambda_{i1}^{(m)},\alpha)} & \text{si } y_{i1}>0 \end{cases} \\ &= \begin{cases} \dfrac{1}{1+\left(\dfrac{\alpha}{\alpha+\lambda_{i1}^{(m)}}\right)^{\alpha} \cdot (1-\pi_{i1}^{(m)})/\pi_{i1}^{(m)}} & \text{si } y_{i1}=0 \\ 0 & \text{si } y_{i1}>0 \end{cases} \\ &= \begin{cases} \dfrac{1}{1+\left(\dfrac{\alpha}{\alpha+\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}^{(m)})}\right)^{\alpha} \cdot \dfrac{1}{1+\exp(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)})}} & \text{si } y_{i1}=0 \\ 0 & \text{si } y_{i1}>0 \end{cases} \end{aligned}$$

Paso M. Se ajusta $\boldsymbol{\gamma}^{(m)}$ por la maximización de $\ell(\boldsymbol{\gamma}^{(m)}, y_{i1}, D_{i1}^{(m)})$. A partir de (3-47):

$$\ell(\boldsymbol{\gamma}^{(m)}, y_{i1}, D_{i1}^{(m)}) = \sum_{i=1}^{n} \left\{ D_{i1}\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)} - \ln[1 + \exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)] \right\}$$

La derivada con respecto a $\boldsymbol{\gamma}^{(m)}$ es:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\gamma}^{(m)}, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n} \left\{ D_{i1}^{(m)}\mathbf{z}_{i1}' - \frac{\exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)}{1 + \exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)}\mathbf{z}_{i1}' \right\} \\
&= \sum_{i=1}^{n} \mathbf{z}_{i1}' \left(D_{i1}^{(m)} - \frac{\exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)}{1 + \exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)} \right) \\
&= \sum_{i=1}^{n} \mathbf{z}_{i1}' \left(D_{i1}^{(m)} - \pi_{i1}^{(m)} \right)
\end{aligned}$$

La segunda derivada es:

$$\frac{\partial^2 \ell(\boldsymbol{\gamma}^{(m)}, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = \sum_{i=1}^{n} -\mathbf{z}_{i1}' \left(\frac{\exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)}{[1 + \exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)]^2} \right) \mathbf{z}_{i1}$$

Al igual que el modelo Poisson Cero Inflado se reemplazan las funciones de estimación para el paso M del algoritmo EM con funciones que asignan pesos a los datos faltantes. La ecuación a resolver es:

$$\frac{\partial^2 \ell(\boldsymbol{\gamma}^{(m)}, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = \sum_{i=1}^{n} -\mathbf{z}_{i1}' \left(\frac{\exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)}{[1 + \exp\left(\mathbf{z}_{i1}'\boldsymbol{\gamma}^{(m)}\right)]^2} w_{i1(k)}^{(m)} \right) \mathbf{z}_{i1} \tag{3-48}$$

donde $w_{i1(k)}^{(m)}$ es el peso correspondiente para la i-ésima observación en la m-ésima iteración del algoritmo en el tiempo 1, en el k-ésimo patrón de respuesta faltante. Los pesos para los datos faltantes se definen de igual manera que en el modelo Poisson Cero Inflado siguiendo la propuesta de Ayala y Melo (2007). Sean k los posibles patrones de la variable respuesta, se especifican los valores de la siguiente forma: Sea $k = 1$ el primer patrón de respuesta este primer valor de respuesta es cuando $y_i = 0$, luego se define el peso para el primer patrón de dato faltante como:
Si $k = 1$, se tiene que $y_{i1} = 0$, luego

$$\begin{aligned}
w_{i1(1)}^{(m)} &= P[Y_{i1(falt)}, k = 1 | \mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \alpha, \boldsymbol{\gamma}^{(m)}] \\
&= P[Y_{i1(falt)} = 0 | \mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \alpha, \boldsymbol{\gamma}^{(m)}] \\
&= \pi_{i1}^{(m)} + (1 - \pi_{i1}^{(m)}) \left[\frac{\alpha}{\alpha + \lambda_{i1}^{(m)}} \right]^{\alpha}
\end{aligned}$$

Si $k = 2$, se tiene que $y_{i1} = 1$, luego

$$\begin{aligned} w_{i1(2)}^{(m)} &= P[Y_{i1(falt)}, k = 2|\mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \alpha, \boldsymbol{\gamma}^{(m)}] \\ &= P[Y_{i1(falt)} = 1|\mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}] \\ &= (1 - \pi_{i1}^{(m)}) \frac{\Gamma(\alpha + 1)}{\Gamma(1 + 1)\Gamma(\alpha)} \left[\frac{\alpha}{\alpha + \lambda_{i1}^{(m)}}\right]^{\alpha} \left(\frac{\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}}\right)^{1} \end{aligned}$$

Siguiendo de la misma manera se tiene que, si $k = k$, se tiene que $y_{i1} = k - 1$, luego

$$\begin{aligned} w_{i1(k)}^{(m)} &= P[Y_{i1(falt)}, k = k|\mathbf{x}_{i1(falt)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}, \boldsymbol{\gamma}^{(m)}] \\ &= (1 - \pi_{i1}^{(m)}) \frac{\Gamma(\alpha + k - 1)}{\Gamma(1 + k - 1)\Gamma(\alpha)} \left[\frac{\alpha}{\alpha + \lambda_{i1}^{(m)}}\right]^{\alpha} \left(\frac{\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}}\right)^{k-1} \end{aligned}$$

Finalmente, si el dato es observado se tiene que $k = 0$ y se define $w_{i1(0)}^{(m)} = 1$. Por tanto, las ponderaciones para el modelo son:

$$w_{i1(k)}^{(m)} = \begin{cases} 1 & \text{si } k = 0 \\ \pi_{i1}^{(m)} + (1 - \pi_{i1}^{(m)}) \left[\dfrac{\alpha}{\alpha + \lambda_{i1}^{(m)}}\right]^{\alpha} & \text{si } k = 1 \\ (1 - \pi_{i1}^{(m)}) \dfrac{\Gamma(\alpha + k - 1)}{\Gamma(1 + k - 1)\Gamma(\alpha)} \left[\dfrac{\alpha}{\alpha + \lambda_{i1}^{(m)}}\right]^{\alpha} \left(\dfrac{\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}}\right)^{k-1} & \text{si } k > 1 \end{cases}$$

Ahora, al igual que en el modelo Poisson cero inflado para la estimación de los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ se hace uso del método Fisher-Scoring. De (3-48) donde se tiene que:

$$\Im_z^{(m)} = E\left[-\frac{\partial^2 \ell(\boldsymbol{\gamma}^{(m)}, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}\right] = \sum_{i=1}^{n} \left\{\mathbf{z}_{i1}' \pi_{i1}^{(m)} (1 - \pi_{i1}^{(m)}) w_{i1}^{(m)} \mathbf{z}_{i1}\right\} \tag{3-49}$$

y

$$\begin{aligned} \boldsymbol{U}_z^{(m)} &= \sum_{i=1}^{n} \mathbf{z}_{i1}' \left(D_{i1}^{(m)} - \frac{\exp(\mathbf{z}_{i1}' \boldsymbol{\gamma}^{(m)})}{1 + \exp(\mathbf{z}_{i1}' \boldsymbol{\gamma}^{(m)})}\right) w_{i1(k)}^{(m)} \\ &= \sum_{i=1}^{n} \mathbf{z}_{i1}' \left(D_{i1}^{(m)} - \pi_{i1}^{(m)}\right) w_{i1(k)}^{(m)} \end{aligned}$$

Se multiplica en el numerador y el denominador por la misma expresión:

$$\boldsymbol{U}_z^{(m)} = \sum_{i=1}^{n} \mathbf{z}_{i1}' \left(D_{i1}^{(m)} - \pi_{i1}^{(m)}\right) \boldsymbol{w}_{i1(k)}^{(m)} \frac{\pi_{i1}^{(m)}(1 - \pi_{i1}^{(m)})}{\pi_{i1}^{(m)}(1 - \pi_{i1}^{(m)})}$$

Se tiene que:

$$\boldsymbol{U}_z^{(m)} = \sum_{i=1}^{n} \mathbf{z}'_{i1}\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})w_{i1(k)}^{(m)}\left(\frac{D_{i1}^{(m)}-\pi_{i1}^{(m)}}{\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})}\right)$$

Ahora, al igual que el modelo Poisson cero inflado se definen $\boldsymbol{M}_{z1}^{(m)} = diag\left(\pi_{i1}^{(m)}(1-\pi_{i1}^{(m)})\right)$ y $\boldsymbol{W}_1^{(m)} = diag\left(w_{i1(k)}^{(m)}\right)$ matrices de tamaño $n \times n$ y el vector

$$\boldsymbol{v}_{z1}^{(m)} = \left(\ldots, D_{i1}^{(m)} - \pi_{i1}^{(m)}, \ldots\right)'$$

de tamaño $n \times 1$. El score del algoritmo en forma matricial se puede expresar como:

$$\boldsymbol{U}_z^{(m)} = \mathbf{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{v}_{z1}^{(m)} \tag{3-50}$$

donde $\mathbf{Z}_1$ es la matriz de covariables de tamaño $n \times p$. Por el algoritmo Fisher-Scoring y (3-49) y (3-50) se tiene que:

$$\begin{aligned}
\boldsymbol{\gamma}^{(m+1)} &= \boldsymbol{\gamma}^{(m)} + \left[\boldsymbol{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1'\right]^{-1}\boldsymbol{Z}_{i1}'\boldsymbol{M}_{i1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{v}_{z1}^{(m)} \\
\boldsymbol{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1'\boldsymbol{\gamma}^{(m+1)} &= \boldsymbol{Z}_1'\boldsymbol{M}_{i1}^{(m)}\boldsymbol{W}_{i1}^{(m)}\boldsymbol{z}_{i1}'\boldsymbol{\gamma}^{(m)} + \boldsymbol{z}_{i1}'\boldsymbol{M}_{i1}^{(m)}\boldsymbol{W}_{i1}^{(m)}\boldsymbol{v}_{i1}^{(m)} \\
\boldsymbol{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1'\boldsymbol{\gamma}^{(m+1)} &= \boldsymbol{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\left(\boldsymbol{Z}_1'\boldsymbol{\gamma}^{(m)} + \boldsymbol{v}_{z1}^{(m)}\right) \\
\boldsymbol{\gamma}^{(m+1)} &= \left[\boldsymbol{Z}_{i1}'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1'\right]^{-1}\boldsymbol{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\left(\boldsymbol{Z}_1'\boldsymbol{\gamma}^{(m)} + \boldsymbol{v}_{z1}^{(m)}\right)
\end{aligned}$$

que son las ecuaciones normales para la estimación del parámetro $\boldsymbol{\gamma}$. Ahora, para la estimación de $\ell(\boldsymbol{\beta}^{(m)}, \alpha, y_{i1}, D_{i1}^{(m)})$ se debe realizar una estimación de una binomial negativa ponderada.

Por (3-47) se tiene que:

$$\ell(\boldsymbol{\beta}^{(m)}, \alpha, y_{i1}, D_{i1}^{(m)}) \quad = \quad \sum_{i=1}^{n}(1 \; - \; D_{i1})\left\{\alpha\ln[\alpha] + y_i\boldsymbol{x}_{i1}'\boldsymbol{\beta} - (\alpha + y_{i1})\ln[\alpha + \exp\left(x_{i1}'\boldsymbol{\beta}\right)]\right\}$$

La derivada con respecto a $\boldsymbol{\beta}$ es:

$$\begin{aligned}
\frac{\partial\ell(\boldsymbol{\beta}^{(m)}, \alpha, y_{i1}, D_{i1}^{(m)})}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}(1 - D_{i1}^{(m)})\left[y_{i1}\boldsymbol{x}_{i1}' - \frac{(\alpha + y_{i1})\exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}^{(m)}\right)}{\alpha + \exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}^{(m)}\right)}\mathbf{x}_{i1}'\right] \\
&= \sum_{i=1}^{n}(1 - D_{i1}^{(m)})\boldsymbol{x}_{i1}'\left[y_{i1} - \frac{(\alpha + y_{i1})\exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}^{(m)}\right)}{\alpha + \exp\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}^{(m)}\right)}\right]
\end{aligned}$$

Al reemplazar (3-42) en la expresión anterior se tiene que:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta}^{(m)}, \alpha, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) \left[y_{i1} - \frac{(\alpha + y_{i1})\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}} \right] \\
&= \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) \left[\frac{y_{i1}(\alpha + \lambda_{i1}^{(m)}) - (\alpha + y_{i1})\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}} \right] \\
&= \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) \left[\frac{y_{i1}\alpha + \lambda_{i1}^{(m)} y_{i1} - \alpha\lambda_{i1}^{(m)} - y_{i1}\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}} \right] \\
&= \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) \left[\frac{y_{i1}\alpha - \alpha\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}} \right] \\
&= \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) \left[\frac{\alpha}{\alpha + \lambda_{i1}^{(m)}} \right] (y_{i1} - \lambda_{i1}^{(m)})
\end{aligned}$$

De manera similar que con $\boldsymbol{\gamma}$, se reemplaza la función de estimación con una que asigna un peso a los datos faltantes, es decir:

$$\frac{\partial \ell(\boldsymbol{\beta}^{(m)}, \alpha, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{\alpha}{\alpha + \lambda_{i1}^{(m)}} \right] (y_{i1} - \lambda_{i1}^{(m)})$$

La segunda derivada parcial con respecto a $\boldsymbol{\beta}$ es:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta}^{(m)}, \alpha, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= \sum_{i=1}^{n} -\boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{(\alpha + y_{i1})\alpha \exp(\boldsymbol{x}'_{i1}\beta)}{(\alpha + \exp(\boldsymbol{x}'_{i1}\boldsymbol{\beta}))^2} \right] \boldsymbol{x}_{i1} \\
&= \sum_{i=1}^{n} -\boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{(\alpha + y_{i1})\alpha\lambda_{i1}^{(m)}}{(\alpha + \lambda_{i1}^{(m)})^2} \right] \boldsymbol{x}_{i1}
\end{aligned}$$

La matriz de información de Fisher es:

$$\begin{aligned}
\mathfrak{F}_x^{(m)} = E\left[-\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right] &= \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{(\alpha + \lambda_{i1}^{(m)})\alpha\lambda_{i1}^{(m)}}{(\alpha + \lambda_{i1}^{(m)})^2} \right] \boldsymbol{x}_{i1} \\
&= \sum_{i=1}^{n} \boldsymbol{x}'_{i1}(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{\alpha\lambda_{i1}^{(m)}}{(\alpha + \lambda_{i1}^{(m)})} \right] \boldsymbol{x}_{i1}
\end{aligned}$$

Para facilitar la notación se definen las siguientes matrices de tamaño $n \times n$.

$$\begin{aligned}
\boldsymbol{W}_1^{(m)} &= diag(w_{i1(k)}^{(m)}) \\
\boldsymbol{M}_{x1}^{(m)} &= diag\left((1 - D_{i1}^{(m)}) \frac{\alpha\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}} \right)
\end{aligned}$$

Ahora, la segunda derivada con respecto a $\boldsymbol{\beta}$ queda expresada como:

$$\frac{\partial \ell(\boldsymbol{\beta}^{(m)}, y_{i1}, D_{i1}^{(m)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \boldsymbol{x}_{i1}'(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \frac{\alpha \lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}} \boldsymbol{x}_{i1}$$

Al igual que en la estimación de γ se hace uso del método Fisher-Scoring. El score se multiplica en el numerador y el denominador por la misma expresión:

$$\begin{aligned}
\boldsymbol{U}_{x}^{(m)} &= \sum_{i=1}^{n} \boldsymbol{x}_{i1}'(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{\alpha}{\alpha + \lambda_{i1}^{(m)}} \right] (y_{i1} - \lambda_{i1}^{(m)}) \\
\boldsymbol{U}_{x}^{(m)} &= \sum_{i=1}^{n} \boldsymbol{x}_{i1}'(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{\alpha}{\alpha + \lambda_{i1}^{(m)}} \right] (y_{i1} - \lambda_{i1}^{(m)}) \frac{\lambda_{i1}^{(m)}}{\lambda_{i1}^{(m)}} \\
\boldsymbol{U}_{x}^{(m)} &= \sum_{i=1}^{n} \boldsymbol{x}_{i1}'(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{\alpha \lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}} \right] (y_{i1} - \lambda_{i1}^{(m)}) \frac{1}{\lambda_{i1}^{(m)}} \\
\boldsymbol{U}_{x}^{(m)} &= \sum_{i=1}^{n} \boldsymbol{x}_{i1}'(1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{\alpha \lambda_{i1}^{n}}{\alpha + \lambda_{i1}^{(m)}} \right] \frac{(y_{i1} - \lambda_{i1}^{(m)})}{\lambda_{i1}^{(m)}}
\end{aligned}$$

Se define el vector de tamaño $n \times 1$ como:

$$\boldsymbol{v}_{x1}^{(m)} = \left(\ldots, \frac{(y_{i1} - \lambda_{i1}^{(m)})}{\lambda_{i1}^{(m)}}, \ldots \right)'$$

Por tanto, se tiene que el score forma matricial se puede escribir como:

$$\boldsymbol{U}_{x}^{(m)} = \boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{x1}^{(m)} \boldsymbol{V}_{x1}^{(m)} \tag{3-51}$$

donde $\boldsymbol{X}_1$ es una matriz de covariables de tamaño $n \times p$. Reemplazando (3-51) en el algoritmo Scoring Fisher se tiene que:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left(\boldsymbol{X}_1' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{x1}^{(m)} \boldsymbol{W}_1^{(m)} \boldsymbol{X}_1 \right)^{-1} \boldsymbol{X}_1' \boldsymbol{W}_{i1}^{(m)} \boldsymbol{M}_{x1}^{(m)} \boldsymbol{v}_{x1}^{(m)}$$

Al realizar las operaciones se puede expresar de manera más sencilla como:

$$\boldsymbol{\beta}^{(m+1)} = \left[\boldsymbol{X}_1' \boldsymbol{W}_{i1}^{(m)} \boldsymbol{M}_{x1}^{(m)} \boldsymbol{W}_1^{(m)} \boldsymbol{X}_1 \right]^{-1} \boldsymbol{X}_{i1}' \boldsymbol{W}_1^{(m)} \boldsymbol{M}_{x1}^{(m)} \left(\boldsymbol{X}_1' \boldsymbol{\beta}^{(m)} + \boldsymbol{v}_{x1}^{(m)} \right)$$

que son las ecuaciones normales para la estimación del parámetro $\boldsymbol{\beta}$.

La estimación de α para maximizar (3-46) se realiza mediante un algoritmo Newton-Raphson. Las expresiones para el algoritmo son:

$$\frac{\partial \ln \ell}{\partial \alpha} = \sum_{i=1}^{n} (1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\frac{\Gamma'(y_{i1} + \alpha)}{\Gamma(y_{i1} + \alpha)} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \ln(\alpha) + \frac{\alpha}{\alpha} - \ln(\alpha + \lambda_{i1}^{(m)}) - \frac{(\alpha + y_{i1})}{\alpha + \lambda_{i1}^{(m)}} \right]$$

donde

$$\psi(y_{i1} + \alpha) = \frac{\Gamma'(y_{i1} + \alpha)}{\Gamma(y_{i1} + \alpha)}$$

Γ' es la función di-gamma. Por tanto, se tiene que

$$\frac{\partial \ln \ell}{\partial \alpha} = \sum_{i=1}^{n} (1 - D_{i1}^{(m)}) w_{i1(k)}^{(m)} \left[\psi(y_{i1} + \alpha) - \psi(\alpha) + \ln(\alpha) + 1 - \ln(\alpha + \lambda_{i1}^{(m)}) - \frac{(\alpha + y_{i1})}{\alpha + \lambda_{i1}^{(m)}} \right]$$

La segunda derivada es:

$$\frac{\partial^2 \ln \ell}{\partial \alpha^2} = \sum_{i=1}^{n} (1 - D_{i1}^{(m)}) w_{i1(1)}^{(m)} \left[\psi'(y_{i1} + \alpha) - \psi'(\alpha) + \frac{1}{\alpha} - \frac{2}{\alpha + \lambda_{i1}^{(m)}} + \frac{\alpha + y_{i1}}{(\alpha + \lambda_{i1}^{(m)})^2} \right]$$

Por tanto, la estimación de α es mediante el algoritmo Newton-Raphson queda expresada como:

$$\alpha^{(m+1)} = \alpha^{(m)} - \left[\frac{\dfrac{\partial \ln \ell}{\partial \alpha}}{\dfrac{\partial^2 \ln \ell}{\partial \alpha^2}} \right]$$

Ahora, dado que la convergencia del algoritmo depende de una adecuada selección del valor inicial, se propone realizar una regresión binomial negativa con los datos observados y tomar el valor α como punto de partida para el algoritmo.

Paso 2

Estimación e imputación de los datos faltantes en el tiempo 1. Se realiza el algoritmo EM para los datos en el primer tiempo, exactamente igual que el Poisson cero inflado, teniendo en cuenta como estimador de los parámetros del modelo los obtenidos en el paso 1.

Paso E. El paso E imputa los valores para los datos faltantes, utilizando el modelo propuesto por Bartlett (1937):

$$E[g(\mathbf{y}_1)|\mathbf{X}_1, \mathbf{Z}_1] = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}_1 \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{0} \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{0} \\ \mathbf{B}_1 \end{pmatrix} \boldsymbol{\tau} \tag{3-52}$$

donde $\mathbf{A}_1$ y $\mathbf{B}_1$ son matrices de tamaños $n \times (n - n_1)$ donde n es el total de datos, n_1 indica el total de datos observados y $(n - n_1)$ indican las covariables de valor faltante del primer tiempo con las matrices $\mathbf{A}_1$ y $\mathbf{B}_1$ que corresponden a las covariables del modelo Binomial Negativo y cero inflado, respectivamente, de los valores faltantes del primer tiempo. $\boldsymbol{\alpha}$ y $\boldsymbol{\tau}$ corresponden a los vectores de los $(n - n_1)$ coeficientes de regresión para las covariables de valor faltante y $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son los coeficientes estimados en el primer tiempo. Las matrices $\mathbf{X}_1$ y $\mathbf{Z}_1$ son las matrices de covariables asociadas al primer tiempo del modelo Binomial Negativo y cero inflado respectivamente. Al igual que en el paso 2 del tiempo 1 de la distribución Poisson inflada de ceros se deduce que los estimadores de mínimos cuadrados para la información faltante corresponden a los valores esperados de la variable respuesta. Es decir:

$$\widehat{\boldsymbol{\alpha}}^{(m)} = \mathbf{x}'_{i1(falt)} \widehat{\boldsymbol{\beta}}^{(m)}$$
$$\widehat{\boldsymbol{\tau}}^{(m)} = \mathbf{z}'_{i1(falt)} \widehat{\boldsymbol{\gamma}}^{(m)}$$

Teniendo en cuenta que la distribución de la variable respuesta es Binomial Negativa con exceso de ceros se tiene que los modelos a estimar son:

$$\widehat{\alpha}_i^{(m)} = \mathbf{x}'_{i1(falt)} \widehat{\boldsymbol{\beta}}^{(m)} = \ln[\widehat{\lambda}^{(m)}_{i1(falt)}] \qquad \text{si } y_{i1} > 0$$

$$\widehat{\tau}_i^{(m)} = \mathbf{z}'_{i1(falt)} \widehat{\boldsymbol{\gamma}}^{(m)} = \ln \left[\frac{\widehat{\pi}^{(m)}_{i1(falt)}}{1 - \widehat{\pi}^{(m)}_{i1(falt)}} \right] \qquad \text{si } y_{i1} = 0$$

Por lo tanto,

$$\widehat{\lambda}_{i1(falt)}^{(m)} = \exp\left(\mathbf{x}'_{i1(falt)}\boldsymbol{\beta}^{(m)}\right) \qquad \text{si } y_{i1} > 0$$
$$\widehat{\pi}_{i1(falt)}^{(m)} = \frac{\exp\left(\mathbf{z}'_{i1(falt)}\boldsymbol{\gamma}^{(m)}\right)}{1+\exp\left(\mathbf{z}'_{i1(falt)}\boldsymbol{\gamma}^{(m)}\right)} \qquad \text{si } y_{i1} = 0$$

donde $\hat{\lambda}_{i1}^{(m)}$ y $\hat{\pi}_{i1}^{(m)}$ hacen referencia al valor de la media correspondiente al individuo i en el tiempo 1, si su respuesta es diferente de cero o igual a cero respectivamente. Dada la m-ésima iteración el criterio de selección es:

$$\widehat{y}_{i1(falt)} = \begin{cases} 0 & \text{si } \widehat{\pi}_{i1(falt)}^{(m)} > p_o \\ \widehat{\lambda}_{i1(falt)}^{(m)} & \text{si } \widehat{\pi}_{i1(falt)}^{(m)} \leq p_o \end{cases}$$

Al igual que con la variable respuesta Poisson cero inflada se toma un valor p_0 de referencia para imputar el dato como cero o diferente de cero. En las variables cero infladas, generalmente según Da Costa (2003), el porcentaje de ceros es mínimo del 40 % de las observaciones. Por tanto, 0.4 puede ser un valor mínimo si no se tiene conocimiento profundo de la variable que se trabaja.

Paso M. Imputadas las observaciones se procede a la maximización partiendo del conjunto de datos completos, teniendo en cuenta el método Fisher-Scoring, con el estimador inicial dado en el paso 1. Se repite entre el paso E y M hasta lograr la convergencia. Las expresiones son:

$$E\left[D_{i1}^{(m)}|y_{i1},\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}\right] = \begin{cases} \dfrac{1}{1+\left(\dfrac{\alpha}{\alpha+\exp\left(\mathbf{x}'_{i1}\boldsymbol{\beta}^{(m)}\right)}\right)^{\alpha}\cdot\dfrac{1}{1+\exp\left(\mathbf{z}'_{i1}\boldsymbol{\gamma}^{(m)}\right)}} & \text{si } y_{i1}=0 \\ 0 & \text{si } y_{i1}>0 \end{cases}$$

Para estimar α se tiene que:

$$\alpha^{(m+1)} = \alpha^{(m)} - \left[\frac{\dfrac{\partial \ln \ell}{\partial \alpha}}{\dfrac{\partial^2 \ln \ell}{\partial \alpha^2}}\right]$$

donde:

$$\frac{\partial^2 \ln \ell}{\partial \alpha^2} = \sum_{i=1}^{n}(1-D_{i1}^{(m)})w_{i1(1)}^{(m)}\left[\psi'(y_{i1}+\alpha)-\psi'(\alpha)+\frac{1}{\alpha}-\frac{2}{\alpha+\lambda_{i1}^{(m)}}+\frac{\alpha+y_{i1}}{(\alpha+\lambda_{i1}^{(m)})^2}\right]$$

La expresión para la estimación de $\boldsymbol{\beta}$ por el algoritmo Fisher-Scoring es:

$$\boldsymbol{\beta}^{(m+1)} = \left[\boldsymbol{X}'_1\boldsymbol{W}_1^{(m)}\boldsymbol{M}_{x1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{X}_1\right]^{-1}\boldsymbol{X}'_1\boldsymbol{W}_1^{(m)}\boldsymbol{M}_{x1}^{(m)}\left(\boldsymbol{X}'_1\boldsymbol{\beta}^{(m)}+\boldsymbol{v}_{x1}^{(m)}\right)$$

donde $\boldsymbol{X}_1$ es la matriz de covariables de tamaño $n \times p$. Las matrices $\boldsymbol{W}_1^{(m)} = diag(w_{i1(k)}^{(m)})$ y $\boldsymbol{M}_{x1}^{(m)} = diag\left((1 - D_{i1}^{(m)})\dfrac{\alpha\lambda_{i1}^{(m)}}{\alpha + \lambda_{i1}^{(m)}}\right)$ de tamaños $n \times n$ y el vector:

$$\boldsymbol{v}_{x1}^{(m)} = \left(\ldots, \frac{(y_{i1} - \lambda_{i1}^{(m)})}{\lambda_{i1}^{(m)}}, \ldots\right)'$$

de tamaño $n \times 1$. Finalmente, haciendo uso del algoritmo Fisher-Scoring la expresión para $\boldsymbol{\gamma}$ es:

$$\boldsymbol{\gamma}^{(m+1)} = \left[\boldsymbol{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\boldsymbol{Z}_1'\right]^{-1}\boldsymbol{Z}_1'\boldsymbol{M}_{z1}^{(m)}\boldsymbol{W}_1^{(m)}\left(\boldsymbol{Z}_1'\boldsymbol{\gamma}^{(m)} + \boldsymbol{v}_{z1}^{(m)}\right)$$

con $W_1^{(m)} = diag\left(w_{i1(k)}^{(m)}\right)$ y $\boldsymbol{M}_{z1}^{(m)} = diag\left(\pi_{i1}^{(m)}(1 - \pi_{i1}^{(m)})\right)$ matrices de tamaño $n \times n$ y el vector

$$\boldsymbol{v}_{z1}^{(m)} = \left(\ldots, D_{i1}^{(m)} - \pi_{i1}^{(m)}, \ldots\right)'$$

de tamaño $n \times 1$.

3.2.2. Tiempo 2

Se realiza el mismo desarrollo que el llevado a cabo en el tiempo 2 del algoritmo para los datos Poisson cero inflados, se considera como covariable para el modelo las respuestas estimadas en el tiempo 1 y se definen las matrices $\mathbf{Z}_{i2}$ y $\mathbf{X}_2$ de tamaño $n \times (p+1)$ donde las primeras p columnas hacen referencias a las variables independientes y la última columna es y_{i1} las respuestas en el tiempo 1 completas. El modelo que se considera es:

$$\ln[\lambda_2] = \mathbf{X}_2'\boldsymbol{\beta} \text{ si } y_2 > 0$$
$$\ln\left[\frac{\pi_2}{1 - \pi_2}\right] = \mathbf{Z}_2'\boldsymbol{\gamma} \text{ si } y_2 = 0$$

donde $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son los parámetros a estimar en el tiempo 2.

Paso 1

El desarrollo de la log-verosimilitud y la máximización de los parámetros se realiza de igual manera que en el tiempo 2 del algoritmo para datos Poisson cero inflados.

Las expresiones de las matrices para la estimación de $\boldsymbol{\gamma}$ son: $\boldsymbol{M}_{2z}^{(m)} = diag(\pi_{i2}^{(m)}(1-\pi_{i2}^{(m)}))$ y $W_2^{(m)} = diag(w_{i2(k)}^{(m)})$ de tamaños $n \times n$, el vector correspondiente se define como:

$$\boldsymbol{v}_{z2}^{(m)} = \left(\ldots, \left(D_{i2}^{(m)} - \pi_{i2}^{(m)}\right), \ldots\right)'$$

de tamaño $n \times 1$. Las matrices anteriores se sustituyen en el score para expresar de manera matricial como:

$$\boldsymbol{U}_z^{(m)} = \mathbf{Z}_2' \boldsymbol{M}_{2z}^{(m)} \boldsymbol{W}_2^{(m)} \boldsymbol{v}_{z2}^{(m)}$$

con $\mathbf{Z}_2$ de tamaño $n \times (p+1)$ matriz de covariables. Reemplazando el score en el algoritmo Fisher-Scoring se llega a:

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + \left[\mathbf{Z}_2' \boldsymbol{M}_{2z}^{(m)} \boldsymbol{W}_2^{(m)} \mathbf{Z}_2\right]^{-1} \cdot \mathbf{Z}_2' \boldsymbol{M}_{2z}^{(m)} \boldsymbol{W}_2^{(m)} \boldsymbol{v}_{z2}^{(m)}$$

De manera similar, se deduce a partir de la log-verosimilitud para $\boldsymbol{\beta}$ el vector correspondiente es:

$$\boldsymbol{v}_{x2}^{(m)} = \left(\ldots, (y_{i2} - \lambda_{i2}^{(m)}) \frac{1}{\lambda_{i2}^{(m)}}, \ldots\right)'$$

de tamaño $n \times 1$. La matriz $\boldsymbol{M}_{x2}^{(m)} = diag\left((1 - D_{i2}^{(m)}) \dfrac{\alpha \lambda_{i2}^{(m)}}{\alpha + \lambda_{i2}^{(m)}}\right)$ de tamaño $n \times n$. Por el algoritmo Fisher-Scoring y expresando en matrices se tiene que:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}_2' \boldsymbol{M}_{x2}^{(m)} \boldsymbol{W}_2^{(m)} \mathbf{X}_2\right]^{-1} \cdot \mathbf{X}_2' \boldsymbol{M}_{x2}^{(m)} \boldsymbol{W}_2^{(m)} \boldsymbol{v}_{x2}^{(m)}$$

con $\mathbf{X}_2$ matriz de covariables de tamaño $n \times (p+1)$.

Paso 2

La estimación e imputación de datos faltantes en el tiempo 2 por medio del algoritmo EM, se realiza de la misma manera que en el paso 2 del tiempo 1.

3.2.3. Tiempo t

De manera sucesiva se tiene que para cualquier tiempo t se realiza un análisis de componentes principales con las respuestas de los tiempo $t-1$ anteriores. Se selecciona el número de componentes principales que cumplan los criterios de valor propio mayor a uno o que el porcentaje de varianza acumulado sea por lo menos del 70 %. Se toman estas componentes y se realiza el algoritmo conforme lo desarrollado en paso 1 en el tiempo t del algoritmo para los datos Poisson cero inflados y imputación se realiza teniendo en cuenta el estimador encontrado en el paso 1 anterior y se desarrolla la maximización de igual manera que el paso 2 del tiempo t del modelo Poisson cero inflado.

4. Aplicación

En esta sección se muestran los métodos para la estimación e imputación de información faltante, propuestos en la sección anterior. En la aplicación del primer método, propuesto para respuestas Poisson cero infladas, se usan los datos trabajados por Da Costa (2003), los cuales son tomados de un estudio de mejoramiento genético del maíz. En la aplicación del método para respuesta Binomial Negativa cero inflada, se consideran los datos de un estudio del forrajeo del polen presentado en Rodríguez (2014).

4.1. Poisson cero inflado: Estudio mejoramiento del maíz

El maíz desde la época prehispánica ha sido una de las mayores fuentes de alimento. El cultivo inicial fue hecho por pueblos indígenas del centro de México y fue utilizado como moneda de cambio por diversas culturas indígenas, extendiéndose su cultivo a lo largo de América. Introducido en Europa durante la colonia y extendiéndose su cultivo a lo largo del mundo en el siglo XIX.

Actualmente es el cereal con el mayor volumen de producción mundial superando el trigo y el arroz. Mueve un mercado de aproximadamente 40 billones de dólares anuales, distribuidos entre industrias de producción de alimentos para el consumo humano y materia prima para centenas de productos industrializados. EEUU es el mayor productor mundial con 386748 toneladas en el año 2016, seguido de China con 219554 toneladas y tercero Brasil con 86500 toneladas.

El maíz es posible cultivarlo desde el nivel del mar hasta los 3250 metros de altitud y se siembra principalmente sobre el Ecuador. Al ser una de las principales fuentes de alimento, el ciclo de vida de la planta es muy importante para el hombre. Entre las diferentes plagas que atacan el maíz se encuentra la Spodoptera frugiperda, conocida como oruga del cartucho del maíz, una especie que ataca decenas de culturas económicamente importantes en varios países. Hasta hoy, el control de la plaga se hace principalmente con productos químicos. En los últimos años, se ha preponderado por la disminución del uso de plaguicidas en la producción de alimentos dado los posibles efectos adversos que tienen en la salud. Es por esto que para el control de la plaga se ha planteado el mejoramiento genético del maíz.

La empresa Monsanto de Brasil es una de las líderes en la producción de soluciones agrícolas para el mejoramiento del campo. Una de las líneas de investigación es el mejoramiento genético de las plantas. Es por esto, que se diseño un experimento, a cargo del investigador Odnei Fernandes, que fue realizado en Rolandia, Estado de Paraná (Da Costa, 2003).

El experimento inició el 11 de marzo del 2001 y tuvo como objetivo evaluar la eficiencia del maíz genéticamente modificado MON810 en relación con el maíz convencional (híbrido DKB909) en el control de la Spodoptera frugiperda.

El experimento fue completamente aleatorizado, con 3 tratamientos y 8 repeticiones en parcelas de $1250m^2$, siendo evaluadas durante 9 semanas. Los tratamientos fueron:

1. Maíz genéticamente modificado MON810, Tratamiento 1.
2. Maíz convencional con aplicación de insecticidas, Tratamiento 2.
3. Maíz convencional sin aplicación de insecticidas, Tratamiento 3.

Para la aplicación de insecticidas fue establecido, que siempre que del 20 % a 30 % de las plantas tuvieran síntomas de ataques se les aplicaría el insecticida. La variable respuesta fue el número de las orugas grandes en las parcelas.

La tabla **4-1** muestra los datos del estudio del maíz:

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	1	1
5	0	0	0	0	0	0	0	0	1	1
6	0	0	0	0	0	0	0	0	1	1
7	0	0	0	0	0	1	0	1	2	1
8	0	0	0	0	0	1	0	1	3	1
9	0	0	0	0	0	0	0	0	0	2
10	0	0	0	0	0	0	0	0	0	2
11	0	0	0	0	0	0	0	0	0	2
12	0	0	0	0	0	0	0	1	0	2
13	0	0	0	1	1	1	0	1	0	2
14	0	0	0	0	1	2	1	1	0	2
15	0	0	0	0	1	2	1	2	1	2
16	0	0	0	0	2	4	1	2	3	2
17	0	0	0	0	1	4	3	2	2	3
18	0	0	0	0	1	5	4	2	3	3
19	0	0	0	0	0	5	4	2	3	3
20	0	0	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	0	0	0	0	0	8	7	4	4	3
24	0	0	0	0	0	9	7	4	4	3

Tabla 4-1.: Estudio del maíz.

4.1.1. Estimación e imputación de datos faltantes

Los datos son las mediciones durante nueve semanas del número de orugas grandes en cada una de las ocho parcelas que fueron parte del estudio. los tratamientos son: maíz geneticamente modificado T_1, maíz convencional T_2 y maíz convencional sin insecticidas T_3. Se establece que y_{it} corresponde al número de gusanos en la i-ésima parcela en la t-ésima semana, donde $i = 1, ..., 24$ y $t = 1, ..., 9$.

Haciendo uso de un algoritmo propuesto, se realiza la pérdida aleatoria de datos en porcentajes de: 20 %, 30 %, 40 % y 50 % cada uno 100 veces. Se sigue el proceso de imputación y estimación de la información faltante con respuestas Poisson cero inflada, en cada uno de los nueve tiempos. Por medio del algoritmo programado en R, se compara el número de respuestas imputadas con las respuestas originales y se hace una razón de éxito del algoritmo, se estima

un modelo longitudinal de efecto mixto con variable respuesta Poisson cero inflada y se estiman los parámetros del modelo y se comparan con el modelo con los datos completos. A modo de ilustración se muestra el desarrollo del algoritmo paso a paso en el caso que la pérdida aleatoria es del 20 %, en cada uno de los nueve tiempos que conforman las mediciones de la base de datos del maíz, recordando que el algoritmo propuesto se realiza en cada semana en dos pasos. Primer paso estimación de los parámetros del modelo y los pesos, segundo paso imputación de la información pérdida. Por tanto, se tienen 18 pasos en total para la estimación e imputación en el caso de los datos del maíz.

Paso 1. Estimación de los parámetros del modelo en el tiempo 1.

Se realizó una gran cantidad de análisis para encontrar el modelo con las variables independientes que explicaran mejor la variable respuesta. El modelo que obtuvo los mejores resultados es el que se muestra a continuación:

$$\begin{aligned} \ln\left(\frac{\pi_{i1(falt)}}{1-\pi_{i1(falt)}}\right) &= \gamma_0 + \gamma_1 T_i \\ \ln(\lambda_{i1(falt)}) &= \beta_0 + \beta_1 T_i \end{aligned}$$

donde $\pi_{i1(falt)}$ es la probabilidad de que el dato faltante en el tiempo 1 pertenezca al modelo de ceros, $\lambda_{i1(falt)}$ es la media del modelo Poisson y T_i es la covariable que hace referencia al tratamiento de la i-ésima observación. Se toma el primer tiempo como variable respuesta y el tratamiento como única covariable. Los pesos correspondientes a los individuos, se obtienen a partir del algoritmo propuesto con un estimador inicial propuesto.

La tabla **4-2** muestra el 20 % de la información perdida:

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0		0	0	0		0	0		1
2	0	0	0		0	0	0	0	0	1
3	0			0		0	0	0	0	1
4	0		0	0		0	0	0	1	1
5	0			0	0			0	1	1
6	0			0			0		1	1
7		0	0		0	1		1	2	1
8	0	0	0	0		1	0	1	3	1
9	0	0		0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11		0	0	0	0	0	0			2
12	0	0	0		0	0	0		0	2
13	0	0	0	1		1	0		0	2
14	0	0		0	1	2	1	1	0	2
15	0	0		0	1	2	1	2		2
16	0	0	0	0	2	4		2	3	2
17	0		0	0	1	4	3	2	2	3
18	0	0	0		1	5	4	2		3
19	0	0	0	0			4	2	3	3
20	0		0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23		0	0	0	0		7	4	4	3
24	0	0	0	0	0			4	4	3

Tabla 4-2.: Datos del maíz: 20 % de información perdida.

Luego de 30 iteracciones entre los pasos de Esperanza y Maximización se obtienen los estimadores para el tiempo 1, dados por:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -25.30259 \\ 0.00000 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 25.56607 \\ 0.00000 \end{pmatrix}$$

Paso 2. Imputación de los datos faltantes en el tiempo 1 y estimación de los parámetros del modelo.

En la construcción del algoritmo se utiliza la siguiente regla para la imputación de la información perdida:

$$\widehat{y}_{i1(falt)} = \begin{cases} 0 & \text{si } \widehat{\pi}_{i1(falt)} \geq 0.5, \\ [\widehat{\lambda}_{i1(falt)}] & \text{si } \widehat{\pi}_{i1(falt)} < 0.5, \end{cases}$$

La probabilidad de que un dato perdido sea del modelo cero inflado es $\widehat{\pi}_{i1(falt)}$, se propone que si el valor es mayor o igual a 0.5, se imputa como cero el valor perdido. Ahora, si $\widehat{\pi}_{i1(falt)} < 0.5$ el valor debe pertenecer al modelo Poisson y se propone que el valor imputado sea $[\widehat{\lambda}_{i1(falt)}]$ que es la función parte entera de la media del modelo Poisson.

Teniendo en cuenta los pesos estimados en el paso anterior, donde $\widehat{\pi}_{i1(falt)}$ es la probabilidad que el dato faltante sea del modelo cero inflado y $\widehat{\lambda}_{i1(falt)}$ es la media de la observación si el dato es del modelo Poisson. Los valores de $\widehat{\pi}_{1i(falt)}$ y $\widehat{\lambda}_{i1(falt)}$ se muestra en la tabla **4-3** para los datos perdidos en el tiempo 1.

ind	$\widehat{\lambda}_{i1(falt)}$	$\widehat{\pi}_{i1(falt)}$
7	0.00	1.00
11	0.00	1.00
23	0.00	1.00

Tabla 4-3.: Pesos tiempo 1.

La tabla **4-3** muestra que todas las probabilidades $\widehat{\pi}_{i1(falt)}$ de que los datos sean del modelo de cero son iguales a 1. Por tanto, los valores perdidos se imputan como 0 en el tiempo 1.

Los resultados para los estimadores de $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son respectivamente:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -25.30259 \\ 0.00000 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 25.56607 \\ 0.00000 \end{pmatrix}$$

Los estimadores anteriores son exactamente iguales que los mostrados en el paso 1. La razón de este comportamiento se debe a que la estimación de $\widehat{\beta}$ y $\widehat{\gamma}$ en el paso 1 se realiza con una regresión ponderada por unos pesos que a su vez dependen de $\widehat{\lambda}_{i1(falt)}$ y $\widehat{\pi}_{i1(falt)}$. En el paso 2 se realiza la regresión con los datos que fueron completados teniendo en cuenta $\widehat{\lambda}_{i1(falt)}$ y $\widehat{\pi}_{i1(falt)}$. Es decir, los pesos en el paso 1 garantizaron que los patrones de respuesta 'correctos' tuvieran mayor peso en la estimación de los parámetros en el primer paso, con estos pesos se estiman los valores perdidos y se realiza la estimación de los parámetros en el paso 2, esto conlleva a que los coeficientes sean muy similares entre pasos de cada tiempo.

La tabla **4-4** muestra la imputación de la información faltante para el tiempo 1.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0		0	0	0		0	0		1
2	0	0	0		0	0	0	0	0	1
3	0			0		0	0	0	0	1
4	0		0	0		0	0	0	1	1
5	0			0	0			0	1	1
6	0			0			0		1	1
7	**0**	0	0		0	1		1	2	1
8	0	0	0	0		1	0	1	3	1
9	0	0		0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0			2
12	0	0	0		0	0	0		0	2
13	0	0	0	1		1	0		0	2
14	0	0		0	1	2	1	1	0	2
15	0	0		0	1	2	1	2		2
16	0	0	0	0	2	4		2	3	2
17	0		0	0	1	4	3	2	2	3
18	0	0	0		1	5	4	2		3
19	0	0	0	0			4	2	3	3
20	0		0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0		7	4	4	3
24	0	0	0	0	0			4	4	3

Tabla 4-4.: Imputación de datos, tiempo 1.

Paso 3. Estimación de los parámetros del modelo en el tiempo 2.

Teniendo en cuenta la estimación de los valores perdidos en el tiempo 1 que fueron todos cero. Se considera el modelo de regresión

$$\begin{aligned} \ln\left(\frac{\pi_{i2(falt)}}{1-\pi_{i2(falt)}}\right) &= \gamma_0 + \gamma_1 T_i \\ \ln(\lambda_{i2(falt)}) &= \beta_0 + \beta_1 T_i \end{aligned}$$

donde $\pi_{i2(falt)}$ es la probabilidad de que el dato faltante sea del modelo de ceros, $\lambda_{i2(falt)}$ es la media del modelo Poisson y T_i es el tratamiento de la i-ésima observación.

Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -25.30259 \\ 0.00000 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 25.56607 \\ 0.00000 \end{pmatrix}$$

Paso 4. Imputación de los datos faltantes en el tiempo 2 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla utilizada en el paso anterior. Los valores de $\widehat{\pi}_{i2(falt)}$ y $\widehat{\lambda}_{i2(falt)}$ estimados en el paso anterior se muestran en la tabla **4-5** para los datos perdidos en el tiempo 2.

	$\widehat{\lambda}_{i2(falt)}$	$\widehat{\pi}_{i2(falt)}$
1	0.00	1.00
3	0.00	1.00
4	0.00	1.00
5	0.00	1.00
6	0.00	1.00
17	0.00	1.00
20	0.00	1.00

Tabla 4-5.: Pesos tiempo 2.

La tabla **4-5** muestra que todos los valores de $\widehat{\pi}_{i2(falt)}$ son iguales a uno. Por tanto, los valores perdidos en el tiempo 2 se imputan como 0.

Los valores estimados para los coeficientes del modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -25.30259 \\ 0.00000 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 25.56607 \\ 0.00000 \end{pmatrix}$$

Los estimadores anteriores coinciden con los mostrados en el paso 1 del tiempo 2. Esto sucede a la misma situación que ocurre en el tiempo 1, los parámetros $\widehat{\lambda}_{i2(falt)}$ y $\widehat{\pi}_{i2(falt)}$ que definen los valores imputados en el paso 2, son los mismos que sirven para definir los pesos en el paso 1, teniendo como consecuencia que los parámetros estimados en los dos pasos sean exactamente iguales.

Los valores estimados en el tiempo 2 se muestran en la tabla **4-6**.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0		0	0		1
2	0	0	0		0	0	0	0	0	1
3	0	**0**		0		0	0	0	0	1
4	0	**0**	0	0		0	0	0	1	1
5	0	**0**		0	0			0	1	1
6	0	**0**		0			0		1	1
7	**0**	0	0		0	1		1	2	1
8	0	0	0	0		1	0	1	3	1
9	0	0		0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0			2
12	0	0	0		0	0	0		0	2
13	0	0	0	1		1	0		0	2
14	0	0		0	1	2	1	1	0	2
15	0	0		0	1	2	1	2		2
16	0	0	0	0	2	4		2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0		1	5	4	2		3
19	0	0	0	0			4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0		7	4	4	3
24	0	0	0	0	0			4	4	3

Tabla 4-6.: Imputación de datos, tiempo 2.

Paso 5. Estimación de los parámetros del modelo en el tiempo 3.

Teniendo en cuenta que la estimación de los valores perdidos en el tiempo 1 y 2 son todos cero, se considera el modelo de regresión:

$$\begin{aligned} \ln\left(\frac{\pi_{i3(falt)}}{1-\pi_{i3(falt)}}\right) &= \gamma_0 + \gamma_1 T_i \\ \ln(\lambda_{i3(falt)}) &= \beta_0 + \beta_1 T_i \end{aligned}$$

donde $\pi_{i3(falt)}$ es la probabilidad de que el dato faltante sea del modelo de ceros, $\lambda_{i3(falt)}$ es la media del modelo Poisson y T_i es el tratamiento de la i-ésima observación.

Luego de 50 iteracciones se tiene que los coeficientes estimados del modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -25.30259 \\ 0.00000 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 25.56607 \\ 0.00000 \end{pmatrix}$$

Paso 6. Imputación de los datos faltantes en el tiempo 3 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla utilizada en el paso 2. Los valores de $\widehat{\lambda}_{i3(falt)}$ y $\widehat{\pi}_{i3(falt)}$ estimados en el paso 5 se muestran en la tabla **4-7** para los datos perdidos en el tiempo 3.

	$\widehat{\lambda}_{i3(falt)}$	$\widehat{\pi}_{i3(falt)}$
3	0.00	1.00
5	0.00	1.00
6	0.00	1.00
9	0.00	1.00
14	0.00	1.00
15	0.00	1.00

Tabla 4-7.: Pesos tiempo 3.

La tabla **4-7** muestra que todos los valores de $\widehat{\pi}_{i3(falt)}$ son iguales a 1. Por tanto, todos los valores imputados en el tiempo 3 son cero.

Imputada la información perdida en el tiempo 3 se estiman los coeficientes:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -25.30259 \\ 0.00000 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 25.56607 \\ 0.00000 \end{pmatrix}$$

Los estimadores anteriores coinciden con los mostrados en el paso 1 del tiempo 3. Esto sucede debido a que los parámetros $\widehat{\lambda}_{i2(falt)}$ y $\widehat{\pi}_{i2(falt)}$ que definen los valores imputados en el paso 2, son los mismos que sirven para definir los pesos en el paso 1, teniendo como consecuencia que los parámetros estimados en los dos pasos sean exactamente iguales.

Los valores estimados para el tiempo 3 se muestran en la tabla **4-8**.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0		0	0		1
2	0	0	0		0	0	0	0	0	1
3	0	**0**	**0**	0		0	0	0	0	1
4	0	**0**	0	0		0	0	0	1	1
5	0	**0**	**0**	0	0			0	1	1
6	0	**0**	**0**	0			0		1	1
7	**0**	0	0		0	1		1	2	1
8	0	0	0	0		1	0	1	3	1
9	0	0	**0**	0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0			2
12	0	0	0		0	0	0		0	2
13	0	0	0	1		1	0		0	2
14	0	0	**0**	0	1	2	1	1	0	2
15	0	0	**0**	0	1	2	1	2		2
16	0	0	0	0	2	4		2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0		1	5	4	2		3
19	0	0	0	0			4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0		7	4	4	3
24	0	0	0	0	0			4	4	3

Tabla 4-8.: Imputación de datos, tiempo 3.

Paso 7. Estimación de los parámetros del modelo en el tiempo 4.

Teniendo en cuenta que en los tiempos: 1, 2 y 3 los valores imputados fueron cero, se considera el modelo:

$$\begin{aligned} \ln\left(\frac{\pi_{i4(falt)}}{1-\pi_{i4(falt)}}\right) &= \gamma_0 + \gamma_1 T_i \\ \ln(\lambda_{i4(falt)}) &= \beta_0 + \beta_1 T_i \end{aligned}$$

donde $\pi_{i4(falt)}$ es la probabilidad de que el dato perdido en el tiempo 4 sea del modelo cero, $\lambda_{i4(falt)}$ es la media del modelo Poisson y T_i es el tratamiento de la i-ésima observación.

Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 18.84401 \\ -10.46154 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 31.97423 \\ -19.85805 \end{pmatrix}$$

Paso 8. Imputación de los datos faltantes en el tiempo 4 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en el paso 2. Los valores $\widehat{\pi}_{i4(falt)}$ y $\widehat{\lambda}_{i4(falt)}$ estimados en el paso 7 se muestran en la tabla **4-9** para los datos perdidos en el tiempo 4.

	$\widehat{\lambda}_{i4(falt)}$	$\widehat{\pi}_{i4(falt)}$
2	4369	0.99
7	4369	0.99
12	0.01	0.00
18	0.00	0.00

Tabla 4-9.: Pesos tiempo 4.

La tabla **4-9** muestra que la probabilidad de que los datos perdidos 2 y 7 en el tiempo 4 sea del modelo cero inflado es de 0.99. Las probabilidades de que los datos perdidos: 12 y 18 sean del modelo cero inflado es 0. Por tanto, la imputación se realiza teniendo en cuenta la parte entera de $\widehat{\lambda}_{i4(falt)}$.

Imputados los valores para los datos perdidos, los coeficientes estimados para $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ son respectivamente:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 20.08324 \\ -11.08123 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 33.84138 \\ -21.09201 \end{pmatrix}$$

Los estimadores anteriores son muy similares a los observados en el paso 1 del tiempo 4. Esto sucede debido a que los parámetros $\widehat{\lambda}_{i2(falt)}$ y $\widehat{\pi}_{i2(falt)}$ que definen los valores imputados, son los mismos que sirven para definir los pesos teniendo como consecuencia que los parámetros estimados en los dos pasos del tiempo 4 sean muy similares.

La tabla **4-10** muestra la imputación de la información faltante en el tiempo 4.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0		0	0		1
2	0	0	0	**0**	0	0	0	0	0	1
3	0	**0**	**0**	0		0	0	0	0	1
4	0	**0**	0	0		0	0	0	1	1
5	0	**0**	**0**	0	0			0	1	1
6	0	**0**	**0**	0			0		1	1
7	**0**	0	0	**0**	0	1		1	2	1
8	0	0	0	0		1	0	1	3	1
9	0	0	**0**	0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0			2
12	0	0	0	**0**	0	0	0		0	2
13	0	0	0	1		1	0		0	2
14	0	0	**0**	0	1	2	1	1	0	2
15	0	0	**0**	0	1	2	1	2		2
16	0	0	0	0	2	4		2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0	**0**	1	5	4	2		3
19	0	0	0	0			4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0		7	4	4	3
24	0	0	0	0	0			4	4	3

Tabla 4-10.: Imputación de datos, tiempo 4.

Paso 9. Estimación de los parámetros del modelo en el tiempo 5.

Teniendo en cuenta que los valores de los tiempos: 1, 2 y 3 son todos cero y que el tiempo 4 solamente tiene una observación diferente de cero. La estimación del modelo con alguno de los tiempos anteriores genera problemas de estimación del modelo, luego para garantizar la convergencia del algoritmo se considera el modelo a estimar:

$$\ln\left(\frac{\pi_{i5(falt)}}{1-\pi_{i5(falt)}}\right) = \gamma_0 + \gamma_1 T_i$$
$$\ln(\lambda_{i5(falt)}) = \beta_0 + \beta_1 T_i$$

donde $\pi_{i5(falt)}$ es la probabilidad que el dato faltante en el tiempo 5 sea del modelo cero inflado, $\lambda_{i5(falt)}$ es la media del modelo Poisson y T_i es la covariable que hace referencia al tratamiento de la i-ésima observación.
Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1.27056 \\ -0.88557 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 21.00088 \\ -11.27518 \end{pmatrix}$$

Paso 10. Imputación de los datos faltantes en el tiempo 5 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en el paso 2. Los valores $\widehat{\pi}_{i5(falt)}$ y $\widehat{\lambda}_{i5(falt)}$ estimados en el paso 9 se muestran en la tabla **4-11** para los datos en el tiempo 5:

	$\widehat{\lambda}_{i5(falt)}$	$\widehat{\pi}_{i5(falt)}$
3	1.46	0.99
4	1.46	0.99
6	1.46	0.99
8	1.46	0.99
13	0.60	0.17
19	0.25	0.00

Tabla 4-11.: Pesos tiempo 5.

La tabla **4-11** muestra que los datos: 3, 4, 6 y 8 deben pertenecen al modelo cero inflado ya que la probabilidad es del 0.99. Los datos 13 y 19 se imputan como ceros que pertenecen al modelo Poisson.

Imputados los datos en el tiempo 5, se estiman los parámetros del modelo:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1.26410 \\ -0.88333 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 20.76184 \\ -11.1596 \end{pmatrix}$$

Los estimadores anteriores son muy similares con los mostrados en el paso 1 del tiempo 5. La razón es la misma que en el tiempo 4, los parámetros $\widehat{\lambda}_{i2(falt)}$ y $\widehat{\pi}_{i2(falt)}$ que definen los valores imputados, son los mismos que sirven para definir los pesos teniendo como consecuencia que los parámetros estimados en los dos pasos del tiempo 5 sean muy similares.

Los valores estimados para el tiempo 5 se muestran en la tabla **4-12**.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0		0	0		1
2	0	0	0	**0**	0	0	0	0	0	1
3	0	**0**	**0**	0	**0**	0	0	0	0	1
4	0	**0**	0	0	**0**	0	0	0	1	1
5	0	**0**	**0**	0	0			0	1	1
6	0	**0**	**0**	0	**0**		0		1	1
7	**0**	0	0	**0**	0	1		1	2	1
8	0	0	0	0	**0**	1	0	1	3	1
9	0	0	**0**	0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0			2
12	0	0	0	**0**	0	0	0		0	2
13	0	0	0	1	**0**	1	0		0	2
14	0	0	**0**	0	1	2	1	1	0	2
15	0	0	**0**	0	1	2	1	2		2
16	0	0	0	0	2	4		2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0	**0**	1	5	4	2		3
19	0	0	0	0	**0**		4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0		7	4	4	3
24	0	0	0	0	0			4	4	3

Tabla 4-12.: Imputación de datos, tiempo 5.

Paso 11. Estimación de los parámetros del modelo en el tiempo 6.

La variable respuesta en el tiempo 5 se toma como covariable, donde de propone el modelo a estimar como:

$$\begin{aligned} \ln\left(\frac{\pi_{i6(falt)}}{1-\pi_{i6(falt)}}\right) &= \gamma_0 + \gamma_1 T_i + \gamma_2 t_5 \\ \ln(\lambda_{i6(falt)}) &= \beta_0 + \beta_1 T_i + \beta_2 t_5 \end{aligned}$$

donde $\pi_{i6(falt)}$ es la probabilidad de que el dato faltante en el tiempo 6 sea del modelo de ceros, $\lambda_{i6(falt)}$ es la media del modelo Poisson, T_i es la covariable del tratamiento asociada a

la i-ésima observación y t_5 son los valores de la variable respuesta en el tiempo 5.

Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -1.74816 \\ 1.04925 \\ 0.34064 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -0.47074 \\ 0.08908 \\ -18.09199 \end{pmatrix}$$

Paso 12. Imputación de los datos faltantes en el tiempo 6 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en el paso 2. Los valores $\widehat{\pi}_{i6(falt)}$ y $\widehat{\lambda}_{i6(falt)}$ estimados en el paso 11 se muestran en la tabla **4-13** para los datos perdidos en el tiempo 6.

	$\widehat{\lambda}_{i6(falt)}$	$\widehat{\pi}_{i6(falt)}$
1	0.49	0.40
5	0.49	0.40
6	0.49	0.40
19	4.05	0.44
23	4.05	0.44
24	4.05	0.44

Tabla 4-13.: Pesos tiempo 6.

La tabla **4-13** muestra que los datos perdidos: 1, 5 y 6 en el tiempo 6, deben ser imputados como 0. Los datos perdidos: 19, 23 y 24 se imputan como 4.

Imputados los datos se procede a la estimación de los parámetros del modelo:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -2.476223 \\ 1.302056 \\ 0.4032276 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 0.8131997 \\ -1.192431 \\ -18.27726 \end{pmatrix}$$

Los valores estimados para el tiempo 6 se muestran en la tabla **4-14**.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0	**0**	0	0		1
2	0	0	0	**0**	0	0	0	0	0	1
3	0	**0**	**0**	0	**0**	0	0	0	0	1
4	0	**0**	0	0	**0**	0	0	0	1	1
5	0	**0**	**0**	0	0	**0**		0	1	1
6	0	**0**	**0**	0	**0**	**0**	0		1	1
7	**0**	0	0	**0**	0	1		1	2	1
8	0	0	0	0	**0**	1	0	1	3	1
9	0	0	**0**	0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0			2
12	0	0	0	**0**	0	0	0		0	2
13	0	0	0	1	**0**	1	0		0	2
14	0	0	**0**	0	1	2	1	1	0	2
15	0	0	**0**	0	1	2	1	2		2
16	0	0	0	0	2	4		2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0	**0**	1	5	4	2		3
19	0	0	0	0	**0**	**4**	4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0	**4**	7	4	4	3
24	0	0	0	0	0	**4**		4	4	3

Tabla 4-14.: Imputación de datos, tiempo 6.

Paso 13. Estimación de los parámetros del modelo en el tiempo 7.

Teniendo en cuenta el algoritmo propuesto se toman las respuestas de los tiempos: 4, 5 y 6 se realiza un análisis de componentes principales para obtener una nueva covariable que retenga la máxima variabilidad contenida en los datos. El porcentaje de varianza retenida en las 100 simulaciones realizadas por el algoritmo fue en promedio del 87 % de la información, luego solo fue necesario utilizar solamente la primera componente principal para el análisis. Por lo tanto, el modelo a estimar es:

$$\ln\left(\frac{\pi_{i7(falt)}}{1-\pi_{i7(falt)}}\right) = \gamma_0 + \gamma_1 T_i + \gamma_2 C_{i1}$$
$$\ln(\lambda_{i7(falt)}) = \beta_0 + \beta_1 T_i + \beta_2 C_{i1}$$

donde $\pi_{i7(falt)}$ es la probabilidad de que el dato faltante en el tiempo 7 sea del modelo cero inflado, $\lambda_{i7(falt)}$ es la media del modelo Poisson, T_i es el tratamiento asociado a la observación i-ésima y C_{i1} es el valor del primer valor propio del análisis de componentes principales de la i-ésima parcela.

Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -4.73427 \\ 2.07799 \\ -0.03075 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -127.91443 \\ 60.65725 \\ 32.25095 \end{pmatrix}$$

Paso 14. Imputación de los datos faltantes en el tiempo 7 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en el paso 2. Los valores $\widehat{\pi}_{i7(falt)}$ y $\widehat{\lambda}_{i7(falt)}$ estimados en el paso 13 se muestran en la tabla **4-15** para los datos perdidos en el tiempo 7.

	$\widehat{\lambda}_{i7(falt)}$	$\widehat{\pi}_{i7(falt)}$
5	0.00	0.98
7	0.00	0.00
16	0.60	0.00
24	4.72	0.24

Tabla 4-15.: Pesos tiempo 7.

La tabla **4-15** muestra que los datos perdidos en el tiempo 7: 5, 7 y 16 deben ser imputados como: 0. Los dos primeros son del modelo Poisson y el tercero al modelo cero inflado. El dato 24 se imputa como 4.

Imputados los datos se procede a la estimación de los parámetros del modelo:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -2.81940 \\ 1.40817 \\ 0.06008 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 383.1398 \\ -203.5181 \\ 36.31872 \end{pmatrix}$$

Los valores estimados para el tiempo 7 se muestran en la tabla **4-16**.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0	**0**	0	0		1
2	0	0	0	**0**	0	0	0	0	0	1
3	0	**0**	**0**	0	**0**	0	0	0	0	1
4	0	**0**	0	0	**0**	0	0	0	1	1
5	0	**0**	**0**	0	0	**0**	**0**	0	1	1
6	0	**0**	**0**	0	**0**	**0**	0		1	1
7	**0**	0	0	**0**	0	1	**0**	1	2	1
8	0	0	0	0	**0**	1	0	1	3	1
9	0	0	**0**	0	0	0	0		0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0			2
12	0	0	0	**0**	0	0	0		0	2
13	0	0	0	1	**0**	1	0		0	2
14	0	0	**0**	0	1	2	1	1	0	2
15	0	0	**0**	0	1	2	1	2		2
16	0	0	0	0	2	4	**0**	2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0	**0**	1	5	4	2		3
19	0	0	0	0	**0**	**4**	4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0	**4**	7	4	4	3
24	0	0	0	0	0	**4**	**4**	4	4	3

Tabla 4-16.: Imputación de datos, tiempo 7.

Paso 15. Estimación de los parámetros del modelo en el tiempo 8.

Teniendo en cuenta el algoritmo propuesto se toman las respuestas de los tiempos: 4, 5, 6 y 7 se realiza un análisis de componentes principales para obtener una nueva covariable que retenga la máxima variabilidad contenida en los datos. Es de resaltar que el promedio de varianza retenida por la primera componente fue en promedio del 83 % en las 100 simulaciones realizadas. Luego, se utiliza solamente la primera componente para este análisis, el modelo a estimar es:

$$\ln\left(\frac{\pi_{i8(falt)}}{1-\pi_{i8(falt)}}\right) = \gamma_0 + \gamma_1 T_i + \gamma_2 C_{i2}$$
$$\ln(\lambda_{i8(falt)}) = \beta_0 + \beta_1 T_i + \beta_2 C_{i2}$$

donde $\pi_{i8(falt)}$ es la probabilidad que el dato faltante sea del modelo cero inflado en el tiempo 8, $\lambda_{i8(falt)}$ es la media del modelo Poisson, T_i es el valor del tratamiento para la i-ésima parcela y C_{i2} son los valores de la i-ésima parcela del primer valor propio del análisis de componentes principales.

Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -1.29254 \\ 0.73279 \\ -0.02134 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -17.18613 \\ 16.88631 \\ 18.84179 \end{pmatrix}$$

Paso 16. Imputación de los datos faltantes en el tiempo 8 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en el paso 2. Los valores $\widehat{\pi}_{i8(falt)}$ y $\widehat{\lambda}_{i8(falt)}$ estimados en el paso 15 se muestran en la tabla **4-17** para los datos perdidos en el tiempo 8.

	$\widehat{\lambda}_{i8(falt)}$	$\widehat{\pi}_{i8(falt)}$
6	0.55	1.00
9	1.16	1.00
11	1.16	1.00
12	1.16	1.00
13	1.18	1.00

Tabla 4-17.: Pesos tiempo 8.

La tabla **4-17** muestra que los datos perdidos en el tiempo 8: 6, 9, 11, 12 y 13 son imputados como cero.

Imputados los datos se procede a la estimación de los parámetros del modelo:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -0.24957 \\ 0.31662 \\ -0.05428 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -127.6694 \\ 90.80311 \\ 63.0691 \end{pmatrix}$$

Los valores estimados para el tiempo 8 se muestran en la tabla **4-18**.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0	**0**	0	0		1
2	0	0	0	**0**	0	0	0	0	0	1
3	0	**0**	**0**	0	**0**	0	0	0	0	1
4	0	**0**	0	0	**0**	0	0	0	1	1
5	0	**0**	**0**	0	0	**0**	**0**	0	1	1
6	0	**0**	**0**	0	**0**	**0**	0	**0**	1	1
7	**0**	0	0	**0**	0	1	**0**	1	2	1
8	0	0	0	0	**0**	1	0	1	3	1
9	0	0	**0**	0	0	0	0	**0**	0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0	**0**		2
12	0	0	0	**0**	0	0	0	**0**	0	2
13	0	0	0	1	**0**	1	0	**0**	0	2
14	0	0	**0**	0	1	2	1	1	0	2
15	0	0	**0**	0	1	2	1	2		2
16	0	0	0	0	2	4	**0**	2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0	**0**	1	5	4	2		3
19	0	0	0	0	**0**	**4**	4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0	**4**	7	4	4	3
24	0	0	0	0	0	**4**	**4**	4	4	3

Tabla 4-18.: Imputación de datos, tiempo 8.

Paso 17. Estimación de los parámetros del modelo en el tiempo 9.

Teniendo en cuenta el algoritmo propuesto se toman las respuestas de los tiempos: 4, 5, 6, 7 y 8 se realiza un análisis de componentes principales para obtener una nueva covariable que retenga la máxima variabilidad contenida en los datos. Es de resaltar que el promedio de varianza retendia por la primera componente fue del 79 % para las 100 simulaciones realizadas. Luego, solamente se utiliza la primera componente para el análisis, el modelo a estimar es:

$$\ln\left(\frac{\pi_{i9(falt)}}{1-\pi_{i9(falt)}}\right) = \gamma_0 + \gamma_1 T_i + \gamma_2 C_{i3}$$
$$\ln(\lambda_{i9(falt)}) = \beta_0 + \beta_1 T_i + \beta_2 C_{i3}$$

donde $\pi_{i9(falt)}$ es la probabilidad que el dato faltante sea del modelo cero inflado en el tiempo 9, $\lambda_{i9(falt)}$ es la media del modelo Poisson, T_i es el tratamiento de la i-ésima parcela y C_{i3} es el primer valor propio del análisis de componentes principales.
Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1.00976 \\ -0.26758 \\ -0.21679 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -8.13382 \\ 3.31004 \\ 1.26937 \end{pmatrix}$$

Paso 18. Imputación de los datos faltantes en el tiempo 9 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en paso 2. Los valores $\widehat{\pi}_{i9(falt)}$ y $\widehat{\lambda}_{i9(falt)}$ estimados en el paso 17 se muestran en la tabla **4-19** para los datos perdidos en el tiempo 9.

	$\widehat{\lambda}_{i9(falt)}$	$\widehat{\pi}_{i9(falt)}$
1	1.20	0.17
11	0.91	0.85
15	1.68	0.14
18	3.00	0.02

Tabla 4-19.: Pesos tiempo 9.

La tabla **4-19** muestra que los datos perdidos en el tiempo 9: 1, 11, 15 y 18 son imputados como: 1, 0, 1 y 3, respectivamente.

Imputados los datos se procede a la estimación de los parámetros del modelo:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.72745 \\ -0.12818 \\ -0.20142 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -576.3481 \\ 289.118 \\ 107.8854 \end{pmatrix}$$

Los valores estimados para el tiempo 9 se muestran en la tabla **4-20**.

	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	Sem9	Trat
1	0	**0**	0	0	0	**0**	0	0	**1**	1
2	0	0	0	**0**	0	0	0	0	0	1
3	0	**0**	**0**	0	**0**	0	0	0	0	1
4	0	**0**	0	0	**0**	0	0	0	1	1
5	0	**0**	**0**	0	0	**0**	**0**	0	1	1
6	0	**0**	**0**	0	**0**	**0**	0	**0**	1	1
7	**0**	0	0	**0**	0	1	**0**	1	2	1
8	0	0	0	0	**0**	1	0	1	3	1
9	0	0	**0**	0	0	0	0	**0**	0	2
10	0	0	0	0	0	0	0	0	0	2
11	**0**	0	0	0	0	0	0	**0**	**0**	2
12	0	0	0	**0**	0	0	0	**0**	0	2
13	0	0	0	1	**0**	1	0	**0**	0	2
14	0	0	**0**	0	1	2	1	1	0	2
15	0	0	**0**	0	1	2	1	2	**1**	2
16	0	0	0	0	2	4	**0**	2	3	2
17	0	**0**	0	0	1	4	3	2	2	3
18	0	0	0	**0**	1	5	4	2	**3**	3
19	0	0	0	0	**0**	**4**	4	2	3	3
20	0	**0**	0	0	0	5	5	2	4	3
21	0	0	0	0	0	4	5	3	4	3
22	0	0	0	0	0	8	6	3	6	3
23	**0**	0	0	0	0	**4**	7	4	4	3
24	0	0	0	0	0	**4**	**4**	4	4	3

Tabla 4-20.: Imputación de datos, tiempo 9.

Al comparar la tabla **4-20** con la tabla de datos originales **4-1** se observa que hay 36 observaciones que coinciden para un éxito del algoritmo en este ejemplo del 80 %.

4.1.2. Análisis de los datos: Poisson cero inflada

El estudio del mejoramiento del maíz fue realizado de manera longitudinal, por lo cual se tiene un efecto de correlación en la cantidad de larvas presentes en cada una de las parcelas medidas a lo largo del tiempo, esto conlleva a realizar un modelo lineal de efecto mixto generalizado con variable respuesta Poisson cero inflada. El objetivo de la investigación es determinar si el número de larvas presentes en cada una de las parcelas, a lo largo de nueve semanas, responde al tratamiento que recibieron cada uno de los campos de maíz. Se realizó una gran cantidad de análisis para encontrar el modelo con las variables independientes que explicaran de mejor manera la variable respuesta. El modelo que obtuvo el mejor ajuste se muestra a continuación:

$$\ln(\lambda_{it}) = \beta_0 + \beta_1 T_i + \beta_2 S_t + a_i$$
$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \gamma_0 + \gamma_1 T_i + \gamma_2 S_t$$

donde T_i indica el tratamiento que recibió cada una de las parcelas experimentales, S_t una variable que mide el efecto de las nueve semanas de medición y a_i es el efecto aleatorio que añade variación al modelo Poisson. Los coeficientes estimados, tanto para el modelo Poisson como el modelo de ceros con sus respectivos valores p se presentan en la tabla **4-21**.

Modelo Poisson				
	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercepto	-3.02861	0.73002	-4.149	0.000
Tratamiento	1.54938	0.23670	6.546	0.000
Semana	-0.04694	0.06072	-0.773	0.439
Modelo de exceso de ceros				
	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercepto	16.072	6.353	2.530	0.01142
Tratamiento	2.349	1.611	1.458	0.14487
Semana	-4.489	1.722	-2.607	0.00913

Tabla 4-21.: Estimadores de los parámetros para los datos completos: Modelo Poisson

La tabla **4-21** muestra que el efecto del tratamiento es estadísticamente significativo para contar el número de larvas en el modelo Poisson. Es decir, el tratamiento influye en el número de larvas que se encontraron en los campos de maíz. El modelo de ceros es explicado de manera significativa por el tiempo (semanas) y el tratamiento. Es decir, el número de campos con cero larvas depende del tratamiento que reciban y el tiempo en semanas que recibieron la dosis, estos resultados concuerdan con los observados en los datos.

Ahora, para medir la efectividad del modelo propuesto para completar la información faltante en las diferentes simulaciones de los escenarios del 20 %, 30 %, 40 % y 50 % de datos perdidos, se estiman los parámetros del modelo lineal de efecto mixto con los datos completos en el momento que se imputa la información faltante y se construyen intervalos de confianza al 95 % y se calcula una media recortada de los parámetros, tanto para los efectos fijos como aleatorios.

Inicialmente la varianza a_i fue de 0.40, es decir para cada parcela se tiene que el error que podemos añadir a la constante es de 0.40. La tabla **4-22** muestra los intervalos de confianza para cada uno de los escenarios simulados:

	Parte aleatoria			
Intervalos	20 %	30 %	40 %	50 %
Limite inferior	0.06	0.01	0.00	0.00
Limite superior	0.35	0.32	0.41	0.53

Tabla 4-22.: Varianza de la parte aleatoria a_i, en el modelo Poisson

En la tabla **4-22** se observa que a medida que aumenta el porcentaje de pérdida aleatoria crece el tamaño del intervalo, donde los intervalos del 40 % y 50 % contienen el valor estimado de la varianza.

La tabla **4-23** muestra los intervalos de confianza al 95 % para los parámetros de los efectos fijos del modelo y la media de las estimaciones.
En la tabla **4-23** se observa que los intervalos de confianza contienen los parámetros estimados de los datos completos que se muestran en **4-21**, es de resaltar que a medida que crece el porcentaje de pérdida aleatoria el tamaño del intervalo aumenta. Se observa que a medida que disminuye el porcentaje de información faltante el promedio de los coeficientes simulados se acerca a los valores de los parámetros observados en **4-21** y que igualmente se observan en la parte inferior de la tabla **4-23**.

Porcentaje	Coeficientes	Modelo Poisson			Modelo de exceso de ceros		
		Intercepto	Trat	Sem	Intercepto	Trat	Sem
20 %	Simulados	-2.64	1.49	-0.07	22.16	1.35	-5.81
	limite inferior	-3.44	1.32	-0.15	10.66	-0.56	-21.39
	limite superior	-1.86	1.74	0.00	106.5	20.2	-2.31
		Modelo Poisson			Modelo de exceso de ceros		
		Intercepto	Trat	Sem	Intercepto	Trat	Sem
30 %	Simulados	-2.61	1.49	-0.09	25.74	1.28	-6.93
	limite inferior	-3.57	1.23	-0.19	10.35	-0.54	-25.06
	limite superior	-1.29	1.92	-0.008	114.73	18.53	-1.62
		Modelo Poisson			Modelo de exceso de ceros		
		Intercepto	Trat	Sem	Intercepto	Trat	Sem
40 %	Simulados	-2.46	1.47	-0.10	42.52	2.97	-10.09
	limite inferior	-4.30	0.99	-0.23	9.27	-2.03	-54.49
	limite superior	-0.53	1.98	0.05	230.40	22.31	-1.14
		Modelo Poisson			Modelo de exceso de ceros		
		Intercepto	Trat	Sem	Intercepto	Trat	Sem
50 %	Simulados	-2.41	1.45	-0.13	47.32	3.27	-7.09
	limite inferior	-5.89	0.73	-0.31	8.39	-2.77	-59.86
	limite superior	-0.23	2.03	0.06	275.48	23.44	-0.57
	Reales	-3.02	1.54	-0.04	16.072	2.349	-4.489

Tabla 4-23.: Estimadores de los parámetros para los datos completos.

El promedio de razón de éxito del algoritmo, para cada uno de los porcentajes de perdida, se muestra en la tabla **4-24**:

20 %	30 %	40 %	50 %
72.21 %	72.49 %	72.96 %	72.55 %

Tabla 4-24.: Éxito del algoritmo: Respuesta Poisson cero inflada.

La tabla **4-24** muestra que el porcentaje de éxito del algoritmo se encuentra en promedio en 72 % y se enfatiza en que los intervalos de confianza para los efectos fijos que se muestran en la tabla **4-23** y que en dos de los cuatro intervalos de la varianza para a_i que se muestran en la tabla **4-22**, contienen los parámetros del modelo completo evidenciando las bondades del algoritmo propuesto para la estimación de la información faltante y que a pesar del aumento en el porcetaje de pérdida aleatoria el modelo propuesto paso a paso se mantiene la varianza de a_i en el porcentaje de éxito de las estimaciones.

4.2. Binomial Negativa Cero Inflada: estudio del forrajeo del polen

El polen de las flores sirve como fuente de alimento para las larvas de las abejas, en este proceso de recolección las hembras visitan gran cantidad de plantas y realizan de manera indirecta la polinización de las plantas, muchas de las cuales son fuente de alimento del hombre, ya que la polinización de las plantas es una parte primordial en el proceso de fertilización de las mismas. Las abejas juegan una parte fundamental en la consecución de la reproducción de las diferentes plantas que hacen parte de la cadena alimenticia del hombre. Por tanto, el éxito en el rendimiento de muchas plantas cultivadas por el hombre depende en una gran medida de las abejas.

Además, las abejas producen diferentes productos que son fuente de alimento como: el polen, la cera, los propóleos y la jalea real. Luego, el manejo y cría de abejas puede ser de gran beneficio para los agricultores ya que garantizan una fuente primaria de sustento para sus hogares dados los diferentes productos que producen las abejas y en la polinización de sus cultivos que garanticen una cosecha éxitosa.

Uno de los inconvenientes en la cría de las abejas es que los insectos pueden no tener una fuerte relación en la polinización del cultivo sembrado por el agricultor. Es decir, el agricultor realiza una labor de sembrado de una planta en particular, pero está no es muy polinizada por sus abejas. A raíz de esta problemática surge la necesidad de estudiar la relación entre las abejas y las plantas. El estudio del tipo de polen transportado por las abejas es una fuente primordial de información del tipo de planta que visita de manera asidua el insecto (Rodríguez, 2014).

Por lo anterior, es que se han realizado estudios que buscan la relación entre insectos y plantas en diferentes países. En Colombia, la información con la que se cuenta al respecto es escasa, es por esto que la investigadora Ángela Rodríguez Calderón con el apoyo de la Universidad Nacional de Colombia y el Laboratorio de Abejas de la Intitución, realizó un estudio en una zona rural del municipio de Acacías, departamento del Meta-Colombia que tenía como objetivo medir el forrajeo del polen de un tipo de abeja sin aguijón que está muy presente en esta zona del país, *Melipona fasciata* (Rodríguez, 2014).

La variable de interés fue la cantidad de abejas que regresan a la colmena con cargas de polen durante tres días, desde las 6 a.m. hasta las 5 p.m. durante cada 10 minutos. La actividad de recolección de los insectos se encuentra relacionado con la temperatura y la humedad del día, la época (lluvias o verano), la colmena (en el estudio se contaba con cinco colmenas), el estado (Estado natural en árboles y en cajas de criadero), la hora del día y el lugar (dentro de las instalaciones del laboratorio y fuera de ellas).

La tabla en el Anexo A muestra los datos completos. Dado que la variable respuesta es un conteo, la investigadora realizó una regresión con variable respuesta tipo Poisson. Es de resaltar que los datos contienen una gran cantidad de variabilidad y de ceros, por tanto se realizó una prueba de razón de verosimilitud para comprobar qué distribución es la adecuada para este tipo de datos. Para llevar a cabo está prueba se ajustan los 3 modelos: Poisson, Binomial Negativo y Binomial Negativo cero inflado con las mismas variables predictoras, para cada modelo se obtiene su respectiva log-verosimilitud, el estadístico utilizado es:

$$RV = -(2(\ell(Poisson) - \ell(Binomial)))$$
$$RV = -(2(\ell(Binomial) - \ell(BinomialCeros)))$$

donde este estadístico tiene una distribución chi-cuadrado según Alcaide (2015). La tabla **4-25** muestra la prueba para cada uno de los tres tiempos.

Regresión Tiempo 1	#Df	LogLik	Df	Chisq	Pr(>Chisq)
Respuesta Poisson	7	-360.61			
Respuesta Binomial Negativa	8	-186.17	1	348.88	0.0000
Respuesta Binomial Negativa Cero Inflada	15	-181.90	7	8.56	0.2861
Regresión Tiempo 2	#Df	LogLik	Df	Chisq	Pr(>Chisq)
Respuesta Poisson	7	-336.25			
Respuesta Binomial Negativa	8	-181.07	1	310.37	0.0000
Respuesta Binomial Negativa Cero Inflada	15	-163.47	7	35.20	0.0000
Regresión Tiempo 3	#Df	LogLik	Df	Chisq	Pr(>Chisq)
Respuesta Poisson	7	-449.94			
Respuesta Binomial Negativa	8	-178.76	1	542.36	0.0000
Respuesta Binomial Negativa Cero Inflada	13	-167.78	5	21.94	0.0005

Tabla 4-25.: Test de Log-verosimilitud.

La tabla **4-25** muestra que en el primer tiempo hay una mejora estadísticamente significativa entre el modelo con variable respuesta Poisson y Binomial Negativa (Valor p=0.0), al comparar los modelos con variable respuesta Binomial Negativa y Binomial Negativa cero inflada no hay una mejora estadístamente significativa (Valor p=0.28). En la regresión con la variable respuesta en el segundo tiempo, se tiene que el modelo Binomial Negativo cero inflado es significativamente más eficiente que el modelo Binomial negativo (Valor p=0.0). Finalmente, en la regresión con la variable respuesta en el tercer tiempo, el modelo Binomial Negativo cero inflado es más eficiente (Valor p=0.0). Luego, existe evidencia estadísticamente significativa de que los datos en dos de los tres tiempos deben ser modelados mediante una regresión Binomial Negativa cero inflada.

Estimación e Imputación de datos faltantes. La variable respuesta tiene tres mediciones en el tiempo, donde y_{it} donde $i = 1, ..., 120$ corresponde al número de insectos que regresan a la colmena con polen y $t = 1, 2, 3$ en la t-ésima ocasión. Haciendo uso del algoritmo propuesto, se realiza la pérdida aleatoria de datos en un porcentaje de: 20 %,30 %,40 % y 50 % cada uno 100 veces. Se sigue el proceso de imputación y estimación de la información faltante con respuestas Binomial Negativa Cero Inflada, en cada uno de los tres tiempos, por medio del algoritmo programado en R, se compara el número de respuestas imputadas con las respuestas originales y se hace una razón de éxito del algoritmo. Finalmente, se estima un modelo longitudinal de efecto mixto con variable respuesta Binomial Negativa cero inflada y se estiman los parámetros del modelo y se comparan con los parámetros del modelo con los datos originales.

A modo de ilustración se muestra el desarrollo del algoritmo paso a paso en que la pérdida aleatoria es del 20 %, en cada uno de los tres tiempos que conforman las mediciones de las abejas, recordando que el algoritmo propuesto se realiza en cada tiempo en dos pasos. Primer paso estimación de los parámetros del modelo y los pesos, segundo paso imputación de la información pérdida y re-estimación de los parámetros del modelo. Por tanto, se tienen 6 pasos para la estimación e imputación de los datos en el caso de los datos de las abejas.

Paso 1. Estimación de los parámetros del modelo en el tiempo 1.

Se realizó una gran cantidad de análisis para encontrar el modelo con las variables independientes que explicaran de mejor manera la variable respuesta. El modelo que obtuvo los mejores resultados se muestra a continuación:

$$\begin{aligned} \ln\left(\frac{\pi_{i(falt)1}}{1-\pi_{i(falt)1}}\right) =& \gamma_0 + \gamma_1 L_i + \gamma_2 E_i + \gamma_3 N_i + \gamma_4 H_i \\ \ln(\lambda_{i(falt)1}) =& \beta_0 + \beta_1 L_i + \beta_2 E_i + \beta_3 N_i + \beta_4 H_i \end{aligned}$$

donde $\pi_{i(falt)1}$ es la probabilidad de que el dato faltante en el tiempo 1 pertencezca al modelo de ceros, $\lambda_{i(falt)1}$ es la media del modelo Binomial Negativo, L_i es la covariable que indica el lugar en que se encuentra la colmena, E_i indica en que estado se encuentra la colmena, la variable N_i indica a a cual de las cinco colmenas pertenece la observación y H_i la hora de la recolección de los datos. Los pesos correspondientes a los individuos, se obtienen a partir del algoritmo propuesto con un estimador inicial propuesto.

Luego de 50 iteracciones entre los pasos de Esperanza y Maximización se obtienen los estimadores iniciales para el tiempo 1, que se muestran a continuación:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1.018 \\ 0.573 \\ 1.966 \\ 0.386 \\ -1.019 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -0.326 \\ -1.317 \\ -1.464 \\ -0.758 \\ 1.026 \end{pmatrix}$$

Paso 2. Imputación de los faltantes en el tiempo 1 y estimación de los parámetros del modelo.

En la construcción del algoritmo se utiliza la siguiente regla para la imputación de la información perdida:

$$\widehat{y}_{i(falt),1} = \begin{cases} 0 & \text{si } \widehat{\pi}_{i(falt)1} \geq 0.5, \\ [\widehat{\lambda}_{i(falt)1}] & \text{si } \widehat{\pi}_{i(falt)1} < 0.5, \end{cases}$$

donde se propone que la probabilidad que el dato perdido sea del modelo cero inflado $\widehat{\pi}_{i(falt)1}$ es mayor o igual a 0.5 se imputa como cero. Ahora si $\widehat{\pi}_{i(falt)1} < 0.5$ el dato debe ser del modelo Binomial Negativo. El valor perdido a imputar se propone como $[\widehat{\lambda}_{i(falt)1}]$ que es la función parte entera de la media del modelo Binomial Negativo.

Teniendo en cuenta los valores estimados en el paso anterior, donde $\widehat{\pi}_{i(falt)1}$ es la probabilidad que el dato faltante sea del modelo cero inflado y $\widehat{\lambda}_{i(falt)1}$ es la media de la observación si es del modelo Binomial Negativo. Los valores $\widehat{\pi}_{i(falt)1}$ y $\widehat{\lambda}_{i(falt)1}$ se muestran en la tabla **4-26** para los datos perdidos en el tiempo 1.

	$\widehat{y}_{i(falt)1}$	$\widehat{\lambda}_{i(falt)1}$	$\widehat{\pi}_{i(falt)1}$
1	3.00	3.61	0.31
2	0.00	0.06	0.89
3	33.00	33.19	0.07
4	12.00	12.11	0.14
5	1.00	1.61	0.40
6	0.00	0.01	0.96
7	0.00	0.00	0.98
8	7.00	7.10	0.20
9	0.00	0.05	0.90
10	0.00	0.00	0.99
11	0.00	0.00	0.99
12	0.00	0.31	0.60
13	21.00	21.20	0.06
14	0.00	0.02	0.94
15	0.00	0.00	0.99
16	12.00	12.11	0.14
17	0.00	0.03	0.92
18	2.00	2.59	0.34
19	0.00	0.00	1.00
20	2.00	2.31	0.26
21	0.00	0.14	0.69
22	0.00	0.05	0.82

Tabla 4-26.: Pesos Respuesta Binomial Negativa, tiempo 1.

La tabla **4-26** muestra los valores estimados para cada dato perdido en el tiempo 1, $\widehat{\pi}_{i(falt)1}$ la probabilidad de que el dato perdido sea del modelo de ceros y $\widehat{\lambda}_{i(falt)1}$ la media del modelo binomial negativo. Imputados los datos se procede a la estimación de los parámetros del modelo:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 2.885 \\ -0.163 \\ 1.309 \\ 0.101 \\ -0.889 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -2.853 \\ 0.183 \\ -14.008 \\ 0.255 \\ 0.0935 \end{pmatrix}$$

Paso 3. Estimación de los parámetros del modelo en el tiempo 2.

Teniendo en cuenta el algoritmo propuesto se toma como covariable los valores de la variable respuesta en el tiempo 1. Ahora, el modelo a estimar es:

$$\begin{aligned}\ln\left(\frac{\pi_{i(falt)2}}{1-\pi_{i(falt)2}}\right) =& \gamma_0 + \gamma_1 T1_i + \gamma_2 L_i + \gamma_3 E_i + \gamma_4 N_i + \gamma_5 H_i \\ \ln(\lambda_{i(falt)2}) =& \beta_0 + \beta_1 T1_i + \beta_2 L_i + \beta_3 E_i + \beta_4 N_i + \beta_5 H_i\end{aligned}$$

donde $\pi_{i(falt)2}$ es la probabilidad de que el dato faltante sea del modelo cero inflado, $\lambda_{i(falt)2}$ es la media del modelo Binomial Negativo, $T1_i$ son los valores que toma la variable respuesta en el tiempo 1, L_i es la covariable que indica en que lugar se encuentra la colmena, E_i indica en que estado se encuentra el nido, la variable N_i indica el nido en que se toma el dato y H_i muestra la hora del día en que toma la observación. Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -3.326 \\ -0.003 \\ 2.197 \\ 4.458 \\ 0.707 \\ -0.837 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 3.194 \\ -0.200 \\ -1.133 \\ -3.765 \\ -0.808 \\ 0.822 \end{pmatrix}$$

Paso 4. Imputación de los datos faltantes en el tiempo 2 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en el paso 2. Los valores $\widehat{\pi}_{i(falt)2}$ y $\widehat{\lambda}_{i(falt)2}$ se muestran en la tabla **4-27** para los datos perdidos en el tiempo 2:

	$\widehat{y}_{i(falt)2}$	$\widehat{\lambda}_{i(falt)2}$	$\widehat{\pi}_{i(falt)2}$
1	20.00	20.39	0.00
2	4.00	4.54	0.34
3	0.00	0.03	0.99
4	0.00	0.01	1.00
5	0.00	0.01	1.00
6	0.00	1.74	0.69
7	0.00	0.33	0.92
8	0.00	0.06	0.98
9	0.00	0.01	1.00
10	0.00	0.00	1.00
11	0.00	0.51	0.78
12	0.00	0.00	1.00
13	4.00	4.18	0.09
14	0.00	0.35	0.76
15	0.00	0.15	0.88
16	0.00	0.03	0.97
17	3.00	3.67	0.09
18	1.00	1.61	0.33
19	0.00	0.02	0.97
20	0.00	0.00	0.99
21	21.00	21.18	0.00
22	0.00	0.01	1.00
23	1.00	1.13	0.21
24	0.00	0.22	0.82
25	0.00	0.00	1.00
26	0.00	0.00	1.00
27	1.00	1.83	0.29
28	0.00	0.03	0.97
29	1.00	1.62	0.38
30	0.00	0.30	0.76
31	0.00	0.01	0.99

Tabla 4-27.: Pesos Respuesta Binomial Negativa, tiempo 2.

La tabla **4-27** muestra los valores imputados en el tiempo 2, la probabilidad $\widehat{\pi}_{i(falt)2}$ de que el dato perdido sea del modelo de ceros y $\widehat{\lambda}_{i(falt)2}$ la media del modelo binomial negativo. Imputados los datos se procede a la estimación de los parámetros con los datos completos:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -3.877 \\ 0.002 \\ 2.380 \\ 4.350 \\ 0.754 \\ -0.689 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -0.283 \\ -15.640 \\ 18.658 \\ -4.767 \\ 1.613 \\ -0.416 \end{pmatrix}$$

Paso 5. Estimación de los parámetros del modelo en el tiempo 3.

Teniendo en cuenta el algoritmo propuesto se toman las respuestas de los tiempos 1 y 2 y se realiza un análisis de componentes principales para obtener una nueva covariable que retenga la máxima variabilidad contenida en los datos, la primera componente retiene 76 % de la información. Es de resaltar que en las simulaciones el promedio de porcentaje retenido fue del 82 %. Luego, con la primera componente principal es suficiente para estimar el modelo propuesto:

$$\begin{aligned} \ln\left(\frac{\pi_{i(falt)3}}{1-\pi_{i(falt)3}}\right) =& \gamma_0 + \gamma_1 L_i + \gamma_2 E_i + \gamma_3 N_i + \gamma_4 H_i + \gamma_5 C1_i \\ \ln(\lambda_{i(falt)3}) =& \beta_0 + \beta_1 L_i + \beta_2 E_i + \beta_3 N_i + \beta_4 H_i + \beta_5 C1_i \end{aligned}$$

donde $\pi_{i(falt)3}$ es la probabilidad que el dato faltante sea del modelo de ceros, $\lambda_{i(falt)3}$ es la media del modelo Binomial Negativo, L_i indica en lugar en que se encuentra la colmena, E_i muestra el estado de la colmena, la variable N_i indica el nido, E_i del año, $C1_i$ es la primera componente principal y H_i es la hora del día en que se tomó la observación. Los pesos correspondientes a los individuos, se obtienen a partir del algoritmo propuesto con un estimador inicial propuesto.
Luego de 50 iteracciones se tiene que los coeficientes estimados para el modelo son:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 10.644 \\ -4.464 \\ -1.424 \\ -1.176 \\ 0.012 \\ -0.805 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -8.997 \\ 3.829 \\ 1.648 \\ 1.124 \\ -0.054 \\ 0.584 \end{pmatrix}$$

Paso 6. Imputación de los datos faltantes en el tiempo 3 y estimación de los parámetros del modelo.

La imputación de la información perdida se realiza teniendo en cuenta la regla propuesta en el paso 1. Los valores $\widehat{\pi}_{i(falt)3}$ y $\widehat{\lambda}_{i(falt)3}$ se muestran en la tabla **4-28** para los datos perdidos en el tiempo 3:

	$\widehat{y}_{i(falt)3}$	$\widehat{\lambda}_{i(falt)3}$	$\widehat{\pi}_{i(falt)3}$
1	0.00	0.00	0.99
2	1.00	1.15	0.44
3	0.00	0.55	0.66
4	0.00	0.02	0.96
5	0.00	0.01	0.97
6	0.00	0.00	0.99
7	0.00	0.00	1.00
8	8.00	8.44	0.22
9	3.00	3.71	0.35
10	0.00	0.01	0.99
11	0.00	1.15	0.62
12	0.00	0.00	1.00
13	3.00	3.21	0.34
14	0.00	0.02	0.95
15	5.00	5.76	0.20
16	2.00	2.55	0.32
17	0.00	0.99	0.62
18	0.00	0.44	0.75
19	0.00	1.24	0.51
20	0.00	0.54	0.67
21	0.00	0.00	1.00
22	0.00	0.14	0.88
23	0.00	0.01	0.99
24	0.00	1.12	0.65
25	0.00	0.04	0.96
26	0.00	0.02	0.98

Tabla 4-28.: Pesos Respuesta Binomial, tiempo 3.

La tabla **4-28** muestra los valores imputados en el tiempo 3, la probabilidad de que $\widehat{\pi}_{i(falt)3}$ el dato faltante sea del modelo de ceros y $\widehat{\lambda}_{i(falt)3}$ es la media del modelo Binomial negativo.

teniendo la información completa se procede a la estimación de los parámetros del modelo.

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 10.255 \\ -3.595 \\ -0.418 \\ -1.376 \\ 0.011 \\ -0.857 \end{pmatrix} \qquad \widehat{\boldsymbol{\gamma}} = \begin{pmatrix} -22.524 \\ 13.716 \\ 13.722 \\ 1.031 \\ -0.161 \\ -1.695 \end{pmatrix}$$

4.2.1. Análisis de los datos: Binomial Negativa cero inflada

El estudio del polen recolectado por las abejas fue realizado de manera longitudinal, por lo cual se tiene un efecto de correlación en la cantidad de polen a lo largo del día en cada una de las colmenas, esto conlleva, a realizar un modelo lineal de efecto mixto generalizado con variable respuesta Binomial Negativa cero inflada. El objetivo de la investigación es determinar si la cantidad de polen recolectado cada una de las colmenas, a lo largo del día, responde a diferentes factores.

Se realizó una gran cantidad de análisis para encontrar el modelo con las variables independientes que explicaran de mejor manera la variable respuesta. El modelo que obtuvo los mejores resultados se muestra a continuación:

$$\ln(\lambda_{it}) = \beta_0 + \beta_1 H_i + \beta_2 E_t + a_i$$
$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \gamma_0 + \gamma_1 H_i + \gamma_2 N_t$$

donde H_t indica la hora en que fue registrada la observación, E_i una variable que mide el estado de la colmena, N_i covariable que indica en cual de las cinco colmenas registradas fue tomado el dato y a_i es el efecto aleatorio que añade variación al modelo Binomial Negativo. Los coeficientes estimados, tanto para el modelo Binomial Negativo como el modelo de ceros con sus respectivos valores p se presentan en la tabla **4-29**:

Binomial Negativa				
	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercepto	2.32600	0.28799	8.077	0.00
hora	-0.75998	0.05279	-14.397	0.00
estado	1.78131	0.27115	6.569	0.00
Modelo de exceso de ceros				
	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercepto	-1.7448	1.0791	-1.617	0.10590
hora	0.6357	0.2083	3.051	0.00228
nido	-1.6115	0.7107	-2.268	0.02335

Tabla 4-29.: Parámetros estimados para los datos completos: Modelo Binomial Negativo.

La tabla **4-29** muestra que el efecto de la medición de la hora es estadísticamente significativa, de signo negativo, para explicar la cantidad de polen recolectado, este resultado concuerda con lo observado en la literatura sobre la actividad de las abejas; es mayor en las horas tempranas y disminuye conforme avanza el día. El estado de la colmena muestra que es significativa para estimar el modelo Binomial Negativo y el signo es positivo, luego un nido de abejas en buen estado influye para que la cantidad de polen recolectado por las abejas

sea mayor. El nido es estadísticamente significativo para explicar el modelo de ceros, es decir los nidos de mayor nomenclatura (4 y 5) tienen una mayor cantidad de abejas que regresan sin polen recolectado.

Ahora, para medir la efectividad del modelo propuesto para completar la información faltante en las diferentes simulaciones de los escenarios del 20 %, 30 %, 40 % y 50 % de datos perdidos, se estiman los parámetros del modelo lineal de efecto mixto con los datos completos en el momento que se imputa la información faltante y se construyen intervalos de confianza al 95 % y se calcula una media recortada de los parámetros estimados, tanto para los efectos fijos como aleatorios.

Inicialmente la varianza de a_i es de 0.14, es decir para cada colmena se tiene que el error que podemos añadir a la constante es de 0.14. La tabla **4-30** muestra los intervalos de confianza para cada uno de los escenarios simulados:

	Parte Aleatoria			
Intervalos	20 %	30 %	40 %	50 %
Limite inferior	0.09	0.13	0.21	0.31
Limite superior	0.20	0.41	0.35	0.43

Tabla 4-30.: Varianza para la parte aleatoria a_i, del modelo Binomial Negativo Cero Inflada.

En la tabla **4-30** se observa que en la simulación de los escenarios al 20 % y 30 % los intervalos contienen el parámetro estimado. Por el contrario, las simulaciones al 40 % y 50 % no contiene el coeficiente estimado. Esto se debe al parámetro de sobredispersión que caracteriza está variable aleatoria, luego se obtuvieron resultados muy diferentes en algunos datos faltantes respecto al verdadero valor del dato.

La tabla **4-31** muestra los intervalos de confianza al 95 % para los parámetros de los efectos fijos del modelo y la media de las estimaciones.
En la tabla **4-31** se observa que en las simulaciones al 20 % y 30 % los intervalos de confianza contienen los parámetros estimados de los datos completos que se muestran en **4-29**, es de resaltar que a medida que crece el porcentaje de pérdida aleatoria en algunos intervalos el tamaño del intervalo aumenta.

Porcentaje		Modelo Binomial Negativa			Modelo de exceso de ceros		
		Intercepto	Hora	Estado	Intercepto	Hora	Nido
20 %	limite inferior	2.01	-0.80	1.73	-2.23	0.43	-1.81
	limite superior	2.43	-0.23	2.01	-0.51	0.81	-1.5
		Modelo Binomial Negativa			Modelo de exceso de ceros		
		Intercepto	Hora	Estado	Intercepto	Hora	Nido
30 %	limite inferior	2.17	-0.77	1.77	-1.79	0.57	-1.65
	limite superior	2.94	-0.17	2.32	-0.45	0.89	-1.42
		Modelo Binomial Negativa			Modelo de exceso de ceros		
		Intercepto	Hora	Estado	Intercepto	Hora	Nido
40 %	limite inferior	2.51	-0.61	1.91	-1.51	0.81	-1.51
	limite superior	3.17	-0.05	2.54	-0.09	0.92	-1.37
		Modelo Binomial Negativa			Modelo de exceso de ceros		
		Intercepto	Hora	Estado	Intercepto	Hora	Nido
50 %	limite inferior	2.81	-0.41	2.37	-1.51	1.03	-1.43
	limite superior	3.43	0.01	3.12	0.10	1.09	-1.23

Tabla 4-31.: Estimadores de los parámetros para los datos completos: Binomial Negativa.

Finalmente, el promedio de razón de éxito del algoritmo, para cada uno de los porcentajes de pérdida, se muestra en la tabla **4-32**:

20 %	30 %	40 %	50 %
54.2 %	52.1	51.1	49.3

Tabla 4-32.: Porcentaje éxito: Respuesta Binomial Negativa cero inflada

La tabla **4-32** muestra que el porcentaje de éxito del algoritmo se encuentra en promedio en 51.6 %. Además, en dos de las cuatro simulaciones los intervalos de confianza para los efectos fijos que se muestran en la tabla **4-31** y en dos de los cuatro intervalos de varianza para a_i que se muestran en la tabla **4-30** contienen los parámetros del modelo completo, evidenciando que a medida que aumenta el porcentaje de datos faltantes el éxito del algoritmo disminuye, esto seguramente se debe a la sobredispersión que genera la variable aleatoria Binomial Negativa. Aunque se resalta que el porcentaje de éxito global de la propuesta es bueno en general.

5. Conclusiones

Se estimaron e imputaron datos faltantes en variables aleatorias binomial negativa cero inflada y Poisson cero infladas en cada tiempo t usando la metodología propuesta. Se propuso el uso del algoritmo EM y el método ANCOVA de Bartlett para la estimación e imputación de la información faltante. La metodología propuesta es útil en las situaciones en que las covariables son completamente observadas, permitiendo la estimación de la variable respuesta que sigue las distribuciones Poisson y Binomial Negativas cero infladas.

La metodología propuesta estima e imputa los valores faltantes haciendo uso del algoritmo EM en dos pasos para cada tiempo; en un paso inicial propone pesos específicos para los valores de la variable respuesta teniendo en cuenta si el dato es observado o faltante y realiza una regresión ponderada, en un segundo paso se imputa la información perdida y se vuelven a estimar los parámetros del modelo.

Se llevó a cabo la aplicación de la metodología propuesta, para la Poisson cero inflada, a datos reales de un estudio llevado a cabo por Da Costa (2003), se tomaron los datos completos y se realizó una pérdida aleatoria del: 20 %, 30 %, 40 % y 50 % los valores estimados tuvieron aciertos promedio del 72 % con respecto a los datos originales y los intervalos de confianza, de las simulaciones, contienen los parámetros estimados del modelo realizado con los datos del estudio del maíz.

La metodología propuesta, para la Binomial Negativa cero inflada, fue aplicada a datos reales de un estudio llevado a cabo por Rodríguez (2012), teniendo en cuenta los datos completos se realizó una pérdida aleatoria de datos del: 20 %, 30 %, 40 % y 50 % los valores estimados tuvieron aciertos promedio del 51.6 % con respecto a los datos originales y en dos de los cuatro escenarios simulados los intervalos de confianza, de las simulaciones, contienen los parámetros estimados del modelo realizado con datos del estudio del polen.

Para trabajos futuros, se recomienda extender la metodología a los casos en que la variable respuesta siga otro tipo de distribución discreta ó continua. Por ejemplo; beta inflada con ceros en el caso continuo y geométrica inflada con ceros en el caso discreto. Hacer uso de los modelos lineales de efectos mixtos generalizados o las ecuaciones de estimación generalizadas para la estimación del algoritmo paso a paso y de los pesos de los datos faltantes. Además, determinar la influencia qué se pueda presentar en la estimación del modelo, la posible

imputación de datos atípicos, así como determinar la eficacia del modelo según el tamaño de muestra.

Bibliografía

Alcaide, M. (2015). *Modelo de regresión Binomial Negativa*. [Tesis de Maestría]. Facultad de Matemáticas, Departamento de Estadística. Sevilla: Universidad de Sevilla.

Ayala, S.Y. (2006). *Estimación e imputación de datos faltantes en diseños de medidas repetidas con respuesta binaria o Poisson*. [Tesis de Maestría]. Facultad de Ciencias, Departamento de Estadística. Bogotá: Universidad Nacional de Colombia.

Ayala, S. Y. & Melo, O. O. (2007). Estimación de datos faltantes en medidas repetidas con respuesta binaria. *Revista Colombiana de Estadística*, **30**, 265-285.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied botany. *Journal of Royal Statistical*, **4**, 137-170.

Chan, J. S. & Wan, W.Y. (2011). Bayesian approach to analysing longitudinal bivariate binary data with informative dropout. *Computational Statistics*, **26**, 121-144.

Da Costa, S. (2003). *Modelos lineares generalizados mistos para dados longitudinais*. [Tesis de Doctorado]. Departamento de Estadística e Experimentación Agronómica. Sao Paulo: Universidad de Sao Paulo.

Daniels, M. & Hogan, J. (2008). *Missing Data in Longitudinal Studies: strategies for bayesian modeling and sensitivity analysis*. New York: Chapman & Hall/CRC.

Davis, C. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New-York: Springer.

Dempster, A. P., Laird, N. M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical*, **39**, 1-38.

Fang, R. (2013), *Zero-inflated negative binomial (ZINB) regression model for over-dispersed count data with excess zeros and repeated measures, an application to human microbiota sequence data*. [Master's Thesis]. Faculty of postgraduate, Biostatistical. University of Colorado.

Famoye, F. & Singh, K. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data science*, **4**, 117-130.

Fitzmaurice, G., Laird, N., & Lipsitz, S. (1994). Analysis Incomplete Longitudinal Binary Responses: A likelihood-Based Approach. *Biometrics*, **50**, 601-612.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2009). *Handbooks of Modern Statistical Methods: Longitudinal Data Analysis*. New York: Chapman & Hall/CRC.

Greene, W., (1994). *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. Working Paper # EC-94-10, School of Business, New York University.

Hall, D., (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**, 1030-1039.

Hall, D. & Shen, J. (2009). Robust Estimation for Zero-Inflated Poisson Regression. *Scandinavian Journal of Statistics*, **37**, 237-252. doi: 10.1111/j.1467-9469.2009.00657.

Ibrahim, J., (1990). Incomplete data in generalized linear models. *Journal of American Statistical Association*, **85**, 765-769.

Karazsia, B. & Van Dulmen, M. (2008). Regression Models for count data: illustrations using longitudinal predictors of childhood injury. *Journal of Pediatric Psychology*, **33**, 1076-1084.

Lambert, D. (1992). Zero-inflated Poisson regression, with and application to defects in manufacturing. *Technometrics*, **34**, 1-14.

Liang, K. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lukusa, M., Lee, S. & Li, C. (2014). A weighted approach to zero-inflated Poisson regression models with missing data in covariates. *Workshop in Symbolic Data Analysis*.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. London: Chapman Hall.

Mouatassim, Y. & Ezzahid, E. Poisson regression and Zero-inflated Poisson regression: application to private health insurance data. *European Actuarial Journal*, **2**, 187-204. doi: 10.1007/s13385-012-0056-2.

Mwalili, S., Lesaffre, E. & Declerck, D. (2007). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research*, **2**, 1-17.

Nelder, J. & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, **135**, 370-384.

Ridout, M., Demétrio, C. & Hinde, J. (1998). *Models for count data with many zeros.* [Invited paper presented at the Nineteenth International Biometric conference]. Capetown, South Africa.

Rodríguez, A.T. (2014). *Requerimientos y valor económico del servicio de polinización prestado por abejas en dos frutales promisorios colombianos, (Champa Campomanesia lineatifolia Ruiz & Pav. y Cholupa Passiflora maliformis L.).* [Tesis de Maestría]. Facultad de Ciencias, Departamento de Biología. Bogotá: Universidad Nacional de Colombia.

Rubin, D. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

Sharma, A. & Landge, V. (2013). Zero inflated negative binomial for modeling heavy vehicle crash rate on indian rural highway. *International Journal of Advances in Engineering and Technology*, **5**, 292-301.

Twisk, J. (2003). *Applied Longitudinal Data Analysis for Epidemiology.* London: Cambridge.

Yao, P. & Liu, X. (2013). Semiparametric Analysis of Longitudinal Zero-inflated count data with applications to instrumental activities of daily living. *Biometrics & Biostatistics*, **4**, doi: 10.4172/2155-6180.1000172

A. Anexo:Regresión Binomial Negativa

A.1. Datos Completos del Forrajeo del Polen

	y1	y2	y3	lugar	estado	nido	epoca
1	48	26	2	1	1	1	1
2	17	17	0	1	1	1	1
3	8	9	0	1	1	1	1
4	3	3	4	1	1	1	1
5	0	0	11	1	1	1	1
6	0	0	1	1	1	1	1
7	0	0	1	1	1	1	1
8	3	0	0	1	1	1	1
9	0	0	0	1	1	1	1
10	0	0	0	1	1	1	1
11	0	0	0	1	1	1	1
12	0	0	0	1	1	1	1
13	11	83	46	1	1	2	1
14	33	46	15	1	1	2	1
15	2	21	13	1	1	2	1
16	0	9	1	1	1	2	1
17	0	0	0	1	1	2	1
18	0	0	0	1	1	2	1
19	0	0	0	1	1	2	1
20	0	0	0	1	1	2	1
30	0	0	0	1	0	3	1
31	0	0	0	1	0	3	1
32	0	0	0	1	0	3	1
33	0	0	0	1	0	3	1
34	0	0	0	1	0	3	1
35	0	0	0	1	0	3	1
36	0	0	0	1	0	3	1
37	29	5	80	0	1	4	1

	y1	y2	y3	lugar	estado	nido	epoca
38	4	7	1	0	1	4	1
39	5	6	0	0	1	4	1
40	3	4	2	0	1	4	1
41	0	0	2	0	1	4	1
42	1	1	1	0	1	4	1
43	0	0	0	0	1	4	1
44	0	1	0	0	1	4	1
45	0	0	0	0	1	4	1
46	0	0	0	0	1	4	1
47	0	0	0	0	1	4	1
48	0	0	0	0	1	4	1
49	28	28	4	0	1	5	1
50	11	23	2	0	1	5	1
51	6	14	3	0	1	5	1
52	5	6	1	0	1	5	1
53	1	0	0	0	1	5	1
54	0	1	0	0	1	5	1
55	0	0	0	0	1	5	1
56	0	0	0	0	1	5	1
57	2	0	0	0	1	5	1
58	1	1	0	0	1	5	1
59	0	2	0	0	1	5	1
60	0	0	0	0	1	5	1
61	38	4	8	1	1	1	2
62	10	1	7	1	1	1	2
63	3	2	1	1	1	1	2
64	0	0	0	1	1	1	2
65	0	0	0	1	1	1	2
66	0	0	0	1	1	1	2
67	0	0	0	1	1	1	2
68	0	0	0	1	1	1	2
69	0	0	0	1	1	1	2
70	0	0	0	1	1	1	2
71	0	0	0	1	1	1	2
72	0	0	0	1	1	1	2
73	12	11	30	1	1	2	2
74	11	10	26	1	1	2	2
75	3	2	10	1	1	2	2
76	2	2	5	1	1	2	2
77	1	0	1	1	1	2	2
78	0	0	0	1	1	2	2

	y1	y2	y3	lugar	estado	nido	epoca
79	0	0	0	1	1	2	2
80	0	0	0	1	1	2	2
81	0	0	0	1	1	2	2
82	0	0	0	1	1	2	2
83	0	0	0	1	1	2	2
84	0	0	0	1	1	2	2
85	9	7	9	1	0	3	2
86	1	2	0	1	0	3	2
87	3	1	1	1	0	3	2
88	0	0	0	1	0	3	2
89	0	0	0	1	0	3	2
90	0	0	0	1	0	3	2
91	0	0	0	1	0	3	2
92	0	0	0	1	0	3	2
93	0	0	0	1	0	3	2
94	0	0	1	1	0	3	2
95	0	0	0	1	0	3	2
96	0	0	0	1	0	3	2
97	36	45	100	0	1	4	2
98	25	22	37	0	1	4	2
99	5	3	13	0	1	4	2
100	2	2	1	0	1	4	2
101	0	1	0	0	1	4	2
102	0	0	0	0	1	4	2
103	1	0	0	0	1	4	2
104	0	0	0	0	1	4	2
105	1	0	0	0	1	4	2
106	0	0	0	0	1	4	2
107	0	0	0	0	1	4	2
108	0	0	0	0	1	4	2
109	141	19	18	0	1	5	2
110	10	13	11	0	1	5	2
111	8	8	4	0	1	5	2
112	2	3	1	0	1	5	2
113	0	1	0	0	1	5	2
114	0	0	0	0	1	5	2
115	0	0	0	0	1	5	2
116	0	0	0	0	1	5	2
117	0	0	0	0	1	5	2
118	0	0	0	0	1	5	2
119	0	0	0	0	1	5	2
120	0	0	0	0	1	5	2

B. Código R algoritmo Poisson cero inflada: Datos del Maíz

```
library(reshape)
library(gdata)
library(splitstackshape)
library(MASS)
library(ade4)
library(pscl)
library(lmtest)
library(glmmTMB)
options(warn=-1)

porcentaje=0.2

n=nrow(datos)
missing=runif(n)<porcentaje
datos[,1][missing]=NA
missing=runif(n)<porcentaje
datos[,2][missing]=NA
missing=runif(n)<porcentaje
datos[,3][missing]=NA
missing=runif(n)<porcentaje
datos[,4][missing]=NA
missing=runif(n)<porcentaje
datos[,5][missing]=NA
missing=runif(n)<porcentaje
datos[,6][missing]=NA
missing=runif(n)<porcentaje
datos[,7][missing]=NA
missing=runif(n)<porcentaje
datos[,8][missing]=NA
missing=runif(n)<porcentaje
datos[,9][missing]=NA
```

```
perdidos=datos

ndatosperdidos=sum(is.na(perdidos))

tiempo1=cbind(datos[,1],datos[,10])
tiempo1=data.frame(tiempo1)
names(tiempo1)=c("resp","Trat")

tiempo1$indper=ifelse(!is.na(tiempo1$resp),0,1)
tiempo1$ind=1:nrow(tiempo1)
tiempo1per=tiempo1[is.na(tiempo1$resp),]

tiempo1per0=tiempo1per
tiempo1per1=tiempo1per
tiempo1per2=tiempo1per
tiempo1per3=tiempo1per
tiempo1per4=tiempo1per
tiempo1per5=tiempo1per
tiempo1per6=tiempo1per
tiempo1per7=tiempo1per
tiempo1per8=tiempo1per
tiempo1per9=tiempo1per
tiempo1per10=tiempo1per

tiempo1per0$resp=0
tiempo1per1$resp=1
tiempo1per2$resp=2
tiempo1per3$resp=3
tiempo1per4$resp=4
tiempo1per5$resp=5
tiempo1per6$resp=6
tiempo1per7$resp=7
tiempo1per8$resp=8
tiempo1per9$resp=9
tiempo1per10$resp=10

tiempo1per=rbind(tiempo1per0,tiempo1per1,tiempo1per2,
tiempo1per3,tiempo1per4,tiempo1per5,
tiempo1per6,tiempo1per7,tiempo1per8,
```

```
tiempo1per9,tiempo1per10)

tiempo1per$indper=ifelse(tiempo1per$resp==0,1,ifelse(tiempo1per$resp==1,2,
ifelse(tiempo1per$resp==2,3,
ifelse(tiempo1per$resp==3,4,ifelse(tiempo1per$resp==4,5,
ifelse(tiempo1per$resp==5,6,ifelse(tiempo1per$resp==6,7,
ifelse(tiempo1per$resp==7,8,ifelse(tiempo1per$resp==8,9,
ifelse(tiempo1per$resp==9,10,ifelse(tiempo1per$resp==10,11,"joder")))))))))))

tiempo1=rbind(tiempo1[!is.na(tiempo1$resp),],tiempo1per)
tiempo1=tiempo1[order(tiempo1$ind),]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 1
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
iterE=100
beta<- matrix(ncol=2,nrow=iterE+1)
gamma<- matrix(ncol=2,nrow=iterE+1)
beta[1,]=zeroinfl(resp~Trat,data=tiempo1)$coef$count
gamma[1,]=zeroinfl(resp~Trat,data=tiempo1)$coef$zero

k=11
N1=nrow(tiempo1)
x=cbind(1,tiempo1$Trat)
z=cbind(1,tiempo1$Trat)
q=k-1

for( i in 1:iterE){
mu=exp(x%*%beta[i,])
pi=exp(z%*%gamma[i,])/( 1+exp(z%*%gamma[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo1$indper==0,1,
ifelse(tiempo1$indper==1,pes(1),
ifelse(tiempo1$indper==2,pes(2),
ifelse(tiempo1$indper==3,pes(3),
ifelse(tiempo1$indper==4,pes(4),
ifelse(tiempo1$indper==5,pes(5),
ifelse(tiempo1$indper==6,pes(6),
```

```
ifelse(tiempo1$indper==7,pes(7),
ifelse(tiempo1$indper==8,pes(8),
ifelse(tiempo1$indper==9,pes(9),
ifelse(tiempo1$indper==10,pes(10),pes(11))))))))))
modelo=zeroinfl(resp~Trat,data=tiempo1,weights=pesos)
beta[i+1,]=modelo$coefficient$count
gamma[i+1,]=modelo$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 2
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo1=cbind(tiempo1,mu, pi)
tiempo1per=tiempo1[!tiempo1$indper==0,]
tiempo1per$resp=ifelse(tiempo1per$pi>0.5,0,as.integer(tiempo1per$mu)  )
tiempo1per=tiempo1per[tiempo1per$indper==1,]
tiempo1=rbind(tiempo1[tiempo1$indper==0,], tiempo1per)
tiempo1=tiempo1[,1:5]
tiempo1=tiempo1[order(tiempo1$ind),]

beta2=matrix(ncol=2,nrow=1)
gamma2=matrix(ncol=2, nrow=1)
modelo2=zeroinfl(resp~Trat,data=tiempo1)
beta2[1,]=modelo2$coefficient$count
gamma2[1,]=modelo2$coefficient$zero

datos=cbind(tiempo1[,1],datos[,2:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 3
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo2=cbind(datos[,1:2],datos[,10])
names(tiempo2)=c("resp1","resp2","Trat")

tiempo2$indper=ifelse(!is.na(tiempo2$resp2),0,1)
tiempo2$ind=1:nrow(tiempo2)
tiempo2per=tiempo2[is.na(tiempo2$resp2),]

tiempo2per0=tiempo2per
tiempo2per1=tiempo2per
```

```
tiempo2per2=tiempo2per
tiempo2per3=tiempo2per
tiempo2per4=tiempo2per
tiempo2per5=tiempo2per
tiempo2per6=tiempo2per
tiempo2per7=tiempo2per
tiempo2per8=tiempo2per
tiempo2per9=tiempo2per
tiempo2per10=tiempo2per

tiempo2per0$resp2=0
tiempo2per1$resp2=1
tiempo2per2$resp2=2
tiempo2per3$resp2=3
tiempo2per4$resp2=4
tiempo2per5$resp2=5
tiempo2per6$resp2=6
tiempo2per7$resp2=7
tiempo2per8$resp2=8
tiempo2per9$resp2=9
tiempo2per10$resp2=10

tiempo2per=rbind(tiempo2per0,tiempo2per1,tiempo2per2,
tiempo2per3,tiempo2per4,tiempo2per5,
tiempo2per6,tiempo2per7,tiempo2per8,
tiempo2per9,tiempo2per10)

tiempo2per$indper=ifelse(tiempo2per$resp2==0,1,ifelse(tiempo2per$resp2==1,2,
ifelse(tiempo2per$resp2==2,3,
ifelse(tiempo2per$resp2==3,4,ifelse(tiempo2per$resp2==4,5,
ifelse(tiempo2per$resp2==5,6,ifelse(tiempo2per$resp2==6,7,
ifelse(tiempo2per$resp2==7,8,ifelse(tiempo2per$resp2==8,9,
ifelse(tiempo2per$resp2==9,10,ifelse(tiempo2per$resp2==10,11,"joder")))))))))))

tiempo2=rbind(tiempo2[!is.na(tiempo2$resp2),],tiempo2per)
tiempo2=tiempo2[order(tiempo2$ind),]

iterE=100
beta3<- matrix(ncol=2,nrow=iterE+1)
gamma3<- matrix(ncol=2,nrow=iterE+1)
```

```
beta3[1,]=zeroinfl(resp2~Trat,data=tiempo2)$coefficient$count
gamma3[1,]=zeroinfl(resp2~Trat,data=tiempo2)$coefficient$zero

k=11
N1=nrow(tiempo2)
X=cbind(1,tiempo2$Trat)
z=cbind(1,tiempo2$Trat)
q=k-1

for( i in 1:iterE){
mu=exp(X%*%beta3[i,])
pi=exp(z%*%gamma3[i,])/( 1+exp(z%*%gamma3[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo2$indper==0,1,
ifelse(tiempo2$indper==1,pes(1),
ifelse(tiempo2$indper==2,pes(2),
ifelse(tiempo2$indper==3,pes(3),
ifelse(tiempo2$indper==4,pes(4),
ifelse(tiempo2$indper==5,pes(5),
ifelse(tiempo2$indper==6,pes(6),
ifelse(tiempo2$indper==7,pes(7),
ifelse(tiempo2$indper==8,pes(8),
ifelse(tiempo2$indper==9,pes(9),
ifelse(tiempo2$indper==10,pes(10),pes(11)))))))))))
modelo3=zeroinfl(resp2~Trat,data=tiempo2,weights=pesos)
beta3[i+1,]=modelo3$coefficient$count
gamma3[i+1,]=modelo3$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 4
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo2=cbind(tiempo2,mu,pi)
tiempo2per=tiempo2[!tiempo2$indper==0,]
tiempo2per$resp2=ifelse(tiempo2per$pi>0.10,0,as.integer(tiempo2per$mu)  )
tiempo2per=tiempo2per[tiempo2per$indper==1,]
tiempo2=rbind(tiempo2[tiempo2$indper==0,], tiempo2per)
tiempo2=tiempo2[,1:5]
tiempo2=tiempo2[order(tiempo2$ind),]
```

```
beta4=matrix(ncol=2,nrow=1)
gamma4=matrix(ncol=2,nrow=1)
modelo4=zeroinfl(resp2~Trat,data=tiempo2)
beta4[1,]=modelo4$coefficient$count
gamma4[1,]=modelo4$coefficient$zero

datos=cbind(tiempo2[,1:2],datos[,3:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 5
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo3=cbind(datos[,1:3],datos[,10])
names(tiempo3)=c("resp1", "resp2", "resp3", "Trat")

tiempo3$indper=ifelse(!is.na(tiempo3$resp3),0,1)
tiempo3$ind=1:nrow(tiempo3)
tiempo3per=tiempo3[is.na(tiempo3$resp3),]

tiempo3per0=tiempo3per
tiempo3per1=tiempo3per
tiempo3per2=tiempo3per
tiempo3per3=tiempo3per
tiempo3per4=tiempo3per
tiempo3per5=tiempo3per
tiempo3per6=tiempo3per
tiempo3per7=tiempo3per
tiempo3per8=tiempo3per
tiempo3per9=tiempo3per
tiempo3per10=tiempo3per

tiempo3per0$resp3=0
tiempo3per1$resp3=1
tiempo3per2$resp3=2
tiempo3per3$resp3=3
tiempo3per4$resp3=4
tiempo3per5$resp3=5
tiempo3per6$resp3=6
tiempo3per7$resp3=7
```

```
tiempo3per8$resp3=8
tiempo3per9$resp3=9
tiempo3per10$resp3=10

tiempo3per=rbind(tiempo3per0,tiempo3per1,tiempo3per2,
tiempo3per3,tiempo3per4,tiempo3per5,
tiempo3per6,tiempo3per7,tiempo3per8,
tiempo3per9,tiempo3per10)

tiempo3per$indper=ifelse(tiempo3per$resp3==0,1,ifelse(tiempo3per$resp3==1,2,
ifelse(tiempo3per$resp3==2,3,
ifelse(tiempo3per$resp3==3,4,ifelse(tiempo3per$resp3==4,5,
ifelse(tiempo3per$resp3==5,6,ifelse(tiempo3per$resp3==6,7,
ifelse(tiempo3per$resp3==7,8,ifelse(tiempo3per$resp3==8,9,
ifelse(tiempo3per$resp3==9,10,ifelse(tiempo3per$resp3==10,11,"joder")))))))))))

tiempo3=rbind(tiempo3[!is.na(tiempo3$resp3),],tiempo3per)
tiempo3=tiempo3[order(tiempo3$ind),]

N=nrow(tiempo3)
x=cbind(1,tiempo3$Trat)
z=cbind(1,tiempo3$Trat)

beta5=matrix(ncol=2,nrow=iterE+1)
beta5[1,]=beta4[1,]
gamma5=matrix(ncol=2,nrow=iterE+1)
gamma5[1,]=gamma4[1,]

for( i in 1:iterE){
mu=exp(x%*%beta5[i,])
pi=exp(z%*%gamma5[i,])/( 1+exp(z%*%gamma5[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo3$indper==0,1,
ifelse(tiempo3$indper==1,pes(1),
ifelse(tiempo3$indper==2,pes(2),
ifelse(tiempo3$indper==3,pes(3),
ifelse(tiempo3$indper==4,pes(4),
ifelse(tiempo3$indper==5,pes(5),
```

```
ifelse(tiempo3$indper==6,pes(6),
ifelse(tiempo3$indper==7,pes(7),
ifelse(tiempo3$indper==8,pes(8),
ifelse(tiempo3$indper==9,pes(9),
ifelse(tiempo3$indper==10,pes(10),pes(11)))))))))))
modelo5=zeroinfl(resp3~Trat,data=tiempo3,weights=pesos)
beta5[i+1,]=modelo5$coefficient$count
gamma5[i+1,]=modelo5$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 6
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo3=cbind(tiempo3,mu,pi)
tiempo3per=tiempo3[!tiempo3$indper==0,]
tiempo3per$resp3=ifelse(tiempo3per$pi>0.10,0,as.integer(tiempo3per$mu)  )
tiempo3per=tiempo3per[tiempo3per$indper==1,]
tiempo3=rbind(tiempo3[tiempo3$indper==0,], tiempo3per)

tiempo3=tiempo3[,1:6]
tiempo3=tiempo3[order(tiempo3$ind),]

beta6=matrix(ncol=2,nrow=1)
gamma6=matrix(ncol=2,nrow=1)
modelo6=zeroinfl(resp3~Trat,data=tiempo3)
beta6[1,]=modelo6$coefficient$count
gamma6[1,]=modelo6$coefficient$zero

datos=cbind(tiempo3[,1:3],datos[,4:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 7
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo4=cbind(datos[,1:4],datos[,10])
names(tiempo4)=c("resp1", "resp2", "resp3","resp4","Trat")

tiempo4$indper=ifelse(!is.na(tiempo4$resp4),0,1)
tiempo4$ind=1:nrow(tiempo4)
tiempo4per=tiempo4[is.na(tiempo4$resp4),]
```

```
tiempo4per0=tiempo4per
tiempo4per1=tiempo4per
tiempo4per2=tiempo4per
tiempo4per3=tiempo4per
tiempo4per4=tiempo4per
tiempo4per5=tiempo4per
tiempo4per6=tiempo4per
tiempo4per7=tiempo4per
tiempo4per8=tiempo4per
tiempo4per9=tiempo4per
tiempo4per10=tiempo4per

tiempo4per0$resp4=0
tiempo4per1$resp4=1
tiempo4per2$resp4=2
tiempo4per3$resp4=3
tiempo4per4$resp4=4
tiempo4per5$resp4=5
tiempo4per6$resp4=6
tiempo4per7$resp4=7
tiempo4per8$resp4=8
tiempo4per9$resp4=9
tiempo4per10$resp4=10

tiempo4per=rbind(tiempo4per0,tiempo4per1,tiempo4per2,
tiempo4per3,tiempo4per4,tiempo4per5,
tiempo4per6,tiempo4per7,tiempo4per8,
tiempo4per9,tiempo4per10)

tiempo4per$indper=ifelse(tiempo4per$resp4==0,1,ifelse(tiempo4per$resp4==1,2,
ifelse(tiempo4per$resp4==2,3,
ifelse(tiempo4per$resp4==3,4,ifelse(tiempo4per$resp4==4,5,
ifelse(tiempo4per$resp4==5,6,ifelse(tiempo4per$resp4==6,7,
ifelse(tiempo4per$resp4==7,8,ifelse(tiempo4per$resp4==8,9,
ifelse(tiempo4per$resp4==9,10,ifelse(tiempo4per$resp4==10,11,"joder")))))))))))

tiempo4=rbind(tiempo4[!is.na(tiempo4$resp4),],tiempo4per)
tiempo4=tiempo4[order(tiempo4$ind),]

N=nrow(tiempo4)
```

```
x=cbind(1,tiempo4$Trat)
z=cbind(1,tiempo4$Trat)

beta7=matrix(ncol=2,nrow=iterE+1)
beta7[1,]=beta6[1,]
gamma7=matrix(ncol=2,nrow=iterE+1)
gamma7[1,]=gamma6[1,]

for( i in 1:iterE){
mu=exp(x%*%beta7[i,])
pi=exp(z%*%gamma7[i,])/( 1+exp(z%*%gamma7[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo4$indper==0,1,
ifelse(tiempo4$indper==1,pes(1),
ifelse(tiempo4$indper==2,pes(2),
ifelse(tiempo4$indper==3,pes(3),
ifelse(tiempo4$indper==4,pes(4),
ifelse(tiempo4$indper==5,pes(5),
ifelse(tiempo4$indper==6,pes(6),
ifelse(tiempo4$indper==7,pes(7),
ifelse(tiempo4$indper==8,pes(8),
ifelse(tiempo4$indper==9,pes(9),
ifelse(tiempo4$indper==10,pes(10),pes(11)))))))))))
modelo7=zeroinfl(resp4~Trat,data=tiempo4,weights=pesos)
beta7[i+1,]=modelo7$coefficient$count
gamma7[i+1,]=modelo7$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 8
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo4=cbind(tiempo4,mu,pi)
tiempo4per=tiempo4[!tiempo4$indper==0,]
tiempo4per$resp4=ifelse(tiempo4per$pi>0.90,0,as.integer(tiempo4per$mu)  )
tiempo4per=tiempo4per[tiempo4per$indper==1,]
tiempo4=rbind(tiempo4[tiempo4$indper==0,], tiempo4per)

tiempo4=tiempo4[,1:7]
tiempo4=tiempo4[order(tiempo4$ind),]
```

```
beta8=matrix(ncol=2,nrow=1)
gamma8=matrix(ncol=2,nrow=1)
modelo8=zeroinfl(resp4~Trat,data=tiempo4)
beta8[1,]=modelo8$coefficient$count
gamma8[1,]=modelo8$coefficient$zero

datos=cbind(tiempo4[,1:4], datos[,5:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 9
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo5=cbind(datos[,1:5],datos[,10])
names(tiempo5)=c("resp1", "resp2", "resp3","resp4", "resp5","Trat")

tiempo5$indper=ifelse(!is.na(tiempo5$resp5),0,1)
tiempo5$ind=1:nrow(tiempo5)
tiempo5per=tiempo5[is.na(tiempo5$resp5),]

tiempo5per0=tiempo5per
tiempo5per1=tiempo5per
tiempo5per2=tiempo5per
tiempo5per3=tiempo5per
tiempo5per4=tiempo5per
tiempo5per5=tiempo5per
tiempo5per6=tiempo5per
tiempo5per7=tiempo5per
tiempo5per8=tiempo5per
tiempo5per9=tiempo5per
tiempo5per10=tiempo5per

tiempo5per0$resp5=0
tiempo5per1$resp5=1
tiempo5per2$resp5=2
tiempo5per3$resp5=3
tiempo5per4$resp5=4
tiempo5per5$resp5=5
tiempo5per6$resp5=6
tiempo5per7$resp5=7
```

```
tiempo5per8$resp5=8
tiempo5per9$resp5=9
tiempo5per10$resp5=10

tiempo5per=rbind(tiempo5per0,tiempo5per1,tiempo5per2,
tiempo5per3,tiempo5per4,tiempo5per5,
tiempo5per6,tiempo5per7,tiempo5per8,
tiempo5per9,tiempo5per10)

tiempo5per$indper=ifelse(tiempo5per$resp5==0,1,ifelse(tiempo5per$resp5==1,2,
ifelse(tiempo5per$resp5==2,3,
ifelse(tiempo5per$resp5==3,4,ifelse(tiempo5per$resp5==4,5,
ifelse(tiempo5per$resp5==5,6,ifelse(tiempo5per$resp5==6,7,
ifelse(tiempo5per$resp5==7,8,ifelse(tiempo5per$resp5==8,9,
ifelse(tiempo5per$resp5==9,10,ifelse(tiempo5per$resp5==10,11,"joder")))))))))))

tiempo5=rbind(tiempo5[!is.na(tiempo5$resp5),],tiempo5per)
tiempo5=tiempo5[order(tiempo5$ind),]

N=nrow(tiempo5)
x=cbind(1,tiempo5$Trat)
z=cbind(1,tiempo5$Trat)

iterE=100
beta9=matrix(ncol=2,nrow=iterE+1)
beta9[1,]=zeroinfl(resp5~Trat,data=tiempo5)$coefficient$count
gamma9=matrix(ncol=2,nrow=iterE+1)
gamma9[1,]=zeroinfl(resp5~Trat,data=tiempo5)$coefficient$zero

for( i in 1:iterE){
mu=exp(x%*%beta9[i,])
pi=exp(z%*%gamma9[i,])/( 1+exp(z%*%gamma9[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo5$indper==0,1,
ifelse(tiempo5$indper==1,pes(1),
ifelse(tiempo5$indper==2,pes(2),
ifelse(tiempo5$indper==3,pes(3),
ifelse(tiempo5$indper==4,pes(4),
```

```
ifelse(tiempo5$indper==5,pes(5),
ifelse(tiempo5$indper==6,pes(6),
ifelse(tiempo5$indper==7,pes(7),
ifelse(tiempo5$indper==8,pes(8),
ifelse(tiempo5$indper==9,pes(9),
ifelse(tiempo5$indper==10,pes(10),pes(11))))))))))
modelo9=zeroinfl(resp5~Trat,data=tiempo5,weights=pesos)
beta9[i+1,]=modelo9$coefficient$count
gamma9[i+1,]=modelo9$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 10
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo5=cbind(tiempo5,mu,pi)
tiempo5per=tiempo5[!tiempo5$indper==0,]
tiempo5per$resp5=ifelse(tiempo5per$pi>0.90,0,as.integer(tiempo5per$mu)  )
tiempo5per=tiempo5per[tiempo5per$indper==1,]
tiempo5=rbind(tiempo5[tiempo5$indper==0,], tiempo5per)

tiempo5=tiempo5[,1:8]
tiempo5=tiempo5[order(tiempo5$ind),]

beta10=matrix(ncol=2,nrow=1)
gamma10=matrix(ncol=2,nrow=1)
modelo10=zeroinfl(resp5~Trat,data=tiempo5)
beta10[1,]=modelo10$coefficient$count
gamma10[1,]=modelo10$coefficient$zero

datos=cbind(tiempo5[,1:5],datos[,6:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 11
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo6=cbind(datos[,1:6],datos[,10])
names(tiempo6)=c("resp1", "resp2", "resp3","resp4", "resp5","resp6","Trat")

tiempo6$indper=ifelse(!is.na(tiempo6$resp6),0,1)
tiempo6$ind=1:nrow(tiempo6)
tiempo6per=tiempo6[is.na(tiempo6$resp6),]
```

```
tiempo6per0=tiempo6per
tiempo6per1=tiempo6per
tiempo6per2=tiempo6per
tiempo6per3=tiempo6per
tiempo6per4=tiempo6per
tiempo6per5=tiempo6per
tiempo6per6=tiempo6per
tiempo6per7=tiempo6per
tiempo6per8=tiempo6per
tiempo6per9=tiempo6per
tiempo6per10=tiempo6per

tiempo6per0$resp6=0
tiempo6per1$resp6=1
tiempo6per2$resp6=2
tiempo6per3$resp6=3
tiempo6per4$resp6=4
tiempo6per5$resp6=5
tiempo6per6$resp6=6
tiempo6per7$resp6=7
tiempo6per8$resp6=8
tiempo6per9$resp6=9
tiempo6per10$resp6=10

tiempo6per=rbind(tiempo6per0,tiempo6per1,tiempo6per2,
tiempo6per3,tiempo6per4,tiempo6per5,
tiempo6per6,tiempo6per7,tiempo6per8,
tiempo6per9,tiempo6per10)

tiempo6per$indper=ifelse(tiempo6per$resp6==0,1,ifelse(tiempo6per$resp6==1,2,
ifelse(tiempo6per$resp6==2,3,
ifelse(tiempo6per$resp6==3,4,ifelse(tiempo6per$resp6==4,5,
ifelse(tiempo6per$resp6==5,6,ifelse(tiempo6per$resp6==6,7,
ifelse(tiempo6per$resp6==7,8,ifelse(tiempo6per$resp6==8,9,
ifelse(tiempo6per$resp6==9,10,ifelse(tiempo6per$resp6==10,11,"joder")))))))))))

tiempo6=rbind(tiempo6[!is.na(tiempo6$resp6),],tiempo6per)
tiempo6=tiempo6[order(tiempo6$ind),]
```

```
N=nrow(tiempo6)
x=cbind(1,tiempo6$Trat, tiempo6$resp5)
z=cbind(1,tiempo6$Trat, tiempo6$resp5)

iterE=100
beta11=matrix(ncol=3,nrow=iterE+1)
beta11[1,]=zeroinfl(resp6~Trat+resp5,data=tiempo6)$coefficient$count
gamma11=matrix(ncol=3,nrow=iterE+1)
gamma11[1,]=zeroinfl(resp6~Trat+resp5,data=tiempo6)$coefficient$zero

for( i in 1:iterE){
mu=exp(x%*%beta11[i,])
pi=exp(z%*%gamma11[i,])/( 1+exp(z%*%gamma11[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo6$indper==0,1,
ifelse(tiempo6$indper==1,pes(1),
ifelse(tiempo6$indper==2,pes(2),
ifelse(tiempo6$indper==3,pes(3),
ifelse(tiempo6$indper==4,pes(4),
ifelse(tiempo6$indper==5,pes(5),
ifelse(tiempo6$indper==6,pes(6),
ifelse(tiempo6$indper==7,pes(7),
ifelse(tiempo6$indper==8,pes(8),
ifelse(tiempo6$indper==9,pes(9),
ifelse(tiempo6$indper==10,pes(10),pes(11)))))))))))
modelo11=zeroinfl(resp6~Trat+resp5,data=tiempo6,weights=pesos)
beta11[i+1,]=modelo11$coefficient$count
gamma11[i+1,]=modelo11$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 12
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo6=cbind(tiempo6,mu,pi)
tiempo6per=tiempo6[!tiempo6$indper==0,]
tiempo6per$resp6=ifelse(tiempo6per$pi>0.80,0,as.integer(tiempo6per$mu)  )
tiempo6per=tiempo6per[tiempo6per$indper==1,]
tiempo6=rbind(tiempo6[tiempo6$indper==0,], tiempo6per)
```

```
tiempo6=tiempo6[,1:10]
tiempo6=tiempo6[order(tiempo6$ind),]

beta12=matrix(ncol=3,nrow=1)
gamma12=matrix(ncol=3,nrow=1)
modelo12=zeroinfl(resp6~Trat+resp5,data=tiempo6)
beta12[1,]=modelo12$coefficient$count
gamma12[1,]=modelo12$coefficient$zero

datos=cbind(tiempo6[,1:6],datos[,7:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 13
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo7=cbind(datos[,1:7],datos[,10])
names(tiempo7)=c("resp1", "resp2", "resp3","resp4", "resp5","resp6","resp7","Trat")

tiempo7$indper=ifelse(!is.na(tiempo7$resp7),0,1)
tiempo7$ind=1:nrow(tiempo7)
tiempo7per=tiempo7[is.na(tiempo7$resp7),]

tiempo7per0=tiempo7per
tiempo7per1=tiempo7per
tiempo7per2=tiempo7per
tiempo7per3=tiempo7per
tiempo7per4=tiempo7per
tiempo7per5=tiempo7per
tiempo7per6=tiempo7per
tiempo7per7=tiempo7per
tiempo7per8=tiempo7per
tiempo7per9=tiempo7per
tiempo7per10=tiempo7per

tiempo7per0$resp7=0
tiempo7per1$resp7=1
tiempo7per2$resp7=2
tiempo7per3$resp7=3
tiempo7per4$resp7=4
tiempo7per5$resp7=5
```

```
tiempo7per6$resp7=6
tiempo7per7$resp7=7
tiempo7per8$resp7=8
tiempo7per9$resp7=9
tiempo7per10$resp7=10

tiempo7per=rbind(tiempo7per0,tiempo7per1,tiempo7per2,
tiempo7per3,tiempo7per4,tiempo7per5,
tiempo7per6,tiempo7per7,tiempo7per8,
tiempo7per9,tiempo7per10)

tiempo7per$indper=ifelse(tiempo7per$resp7==0,1,ifelse(tiempo7per$resp7==1,2,
ifelse(tiempo7per$resp7==2,3,
ifelse(tiempo7per$resp7==3,4,ifelse(tiempo7per$resp7==4,5,
ifelse(tiempo7per$resp7==5,6,ifelse(tiempo7per$resp7==6,7,
ifelse(tiempo7per$resp7==7,8,ifelse(tiempo7per$resp7==8,9,
ifelse(tiempo7per$resp7==9,10,ifelse(tiempo7per$resp7==10,11,"joder")))))))))))

tiempo7=rbind(tiempo7[!is.na(tiempo7$resp7),],tiempo7per)
tiempo7=tiempo7[order(tiempo7$ind),]

uno=data.frame(tiempo7$resp4,tiempo7$resp5, tiempo7$resp6)
c4=princomp(uno,scannf=F)$scores[,1]

N=nrow(tiempo7)
x=cbind(1,tiempo7$Trat, c4)
z=cbind(1,tiempo7$Trat, c4)

iterE=100
beta13=matrix(ncol=3,nrow=iterE+1)
beta13[1,]=beta12[1,]
gamma13=matrix(ncol=3,nrow=iterE+1)
gamma13[1,]=gamma12[1,]

for( i in 1:iterE){
mu=exp(x%*%beta13[i,])
pi=exp(z%*%gamma13[i,])/( 1+exp(z%*%gamma13[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
```

```
pesos=ifelse(tiempo7$indper==0,1,
ifelse(tiempo7$indper==1,pes(1),
ifelse(tiempo7$indper==2,pes(2),
ifelse(tiempo7$indper==3,pes(3),
ifelse(tiempo7$indper==4,pes(4),
ifelse(tiempo7$indper==5,pes(5),
ifelse(tiempo7$indper==6,pes(6),
ifelse(tiempo7$indper==7,pes(7),
ifelse(tiempo7$indper==8,pes(8),
ifelse(tiempo7$indper==9,pes(9),
ifelse(tiempo7$indper==10,pes(10),pes(11))))))))))))
modelo13=zeroinfl(resp7~Trat+c4,data=tiempo7,weights=pesos)
beta13[i+1,]=modelo13$coefficient$count
gamma13[i+1,]=modelo13$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 14
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo7=cbind(tiempo7,c4,mu,pi)
tiempo7per=tiempo7[!tiempo7$indper==0,]
tiempo7per$resp7=ifelse(tiempo7per$pi>0.80,0,as.integer(tiempo7per$mu)  )
tiempo7per=tiempo7per[tiempo7per$indper==1,]
tiempo7=rbind(tiempo7[tiempo7$indper==0,], tiempo7per)

tiempo7=tiempo7[,1:11]
tiempo7=tiempo7[order(tiempo7$ind),]

beta14=matrix(ncol=3,nrow=1)
gamma14=matrix(ncol=3,nrow=1)
modelo14=zeroinfl(resp7~Trat+abs(c4),data=tiempo7)
beta14[1,]=modelo14$coefficient$count
gamma14[1,]=modelo14$coefficient$zero

datos=cbind(tiempo7[,1:7],datos[,8:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 15
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo8=cbind(datos[,1:8],datos[,10])
```

```
names(tiempo8)=c("resp1", "resp2", "resp3","resp4", "resp5","resp6","resp7",
"resp8","Trat")

tiempo8$indper=ifelse(!is.na(tiempo8$resp8),0,1)
tiempo8$ind=1:nrow(tiempo8)
tiempo8per=tiempo8[is.na(tiempo8$resp8),]

tiempo8per0=tiempo8per
tiempo8per1=tiempo8per
tiempo8per2=tiempo8per
tiempo8per3=tiempo8per
tiempo8per4=tiempo8per
tiempo8per5=tiempo8per
tiempo8per6=tiempo8per
tiempo8per7=tiempo8per
tiempo8per8=tiempo8per
tiempo8per9=tiempo8per
tiempo8per10=tiempo8per

tiempo8per0$resp8=0
tiempo8per1$resp8=1
tiempo8per2$resp8=2
tiempo8per3$resp8=3
tiempo8per4$resp8=4
tiempo8per5$resp8=5
tiempo8per6$resp8=6
tiempo8per7$resp8=7
tiempo8per8$resp8=8
tiempo8per9$resp8=9
tiempo8per10$resp8=10

tiempo8per=rbind(tiempo8per0,tiempo8per1,tiempo8per2,
tiempo8per3,tiempo8per4,tiempo8per5,
tiempo8per6,tiempo8per7,tiempo8per8,
tiempo8per9,tiempo8per10)

tiempo8per$indper=ifelse(tiempo8per$resp8==0,1,ifelse(tiempo8per$resp8==1,2,
ifelse(tiempo8per$resp8==2,3,
ifelse(tiempo8per$resp8==3,4,ifelse(tiempo8per$resp8==4,5,
ifelse(tiempo8per$resp8==5,6,ifelse(tiempo8per$resp8==6,7,
```

```
ifelse(tiempo8per$resp8==7,8,ifelse(tiempo8per$resp8==8,9,
ifelse(tiempo8per$resp8==9,10,ifelse(tiempo8per$resp8==10,11,"joder"))))))))))

tiempo8=rbind(tiempo8[!is.na(tiempo8$resp8),],tiempo8per)
tiempo8=tiempo8[order(tiempo8$ind),]

dos=data.frame(tiempo8$resp4,tiempo8$resp5, tiempo8$resp6,tiempo8$resp7)
c5=princomp(dos,scannf=F)$scores[,1]

N=nrow(tiempo8)
x=cbind(1,tiempo8$Trat, c5)
z=cbind(1,tiempo8$Trat, c5)

iterE=100
beta15=matrix(ncol=3,nrow=iterE+1)
beta15[1,]=beta14[1,]
gamma15=matrix(ncol=3,nrow=iterE+1)
gamma15[1,]=gamma14[1,]

for( i in 1:iterE){
mu=exp(x%*%beta15[i,])
pi=exp(z%*%gamma15[i,])/( 1+exp(z%*%gamma15[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo8$indper==0,1,
ifelse(tiempo8$indper==1,pes(1),
ifelse(tiempo8$indper==2,pes(2),
ifelse(tiempo8$indper==3,pes(3),
ifelse(tiempo8$indper==4,pes(4),
ifelse(tiempo8$indper==5,pes(5),
ifelse(tiempo8$indper==6,pes(6),
ifelse(tiempo8$indper==7,pes(7),
ifelse(tiempo8$indper==8,pes(8),
ifelse(tiempo8$indper==9,pes(9),
ifelse(tiempo8$indper==10,pes(10),pes(11)))))))))))
modelo15=zeroinfl(resp8~Trat+floor(c5),data=tiempo8,weights=pesos)
beta15[i+1,]=modelo15$coefficient$count
gamma15[i+1,]=modelo15$coefficient$zero
}
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 16
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo8=cbind(tiempo8,c5,mu,pi)
tiempo8per=tiempo8[!tiempo8$indper==0,]
tiempo8per$resp8=ifelse(tiempo8per$pi>0.80,0,as.integer(tiempo8per$mu)  )
tiempo8per=tiempo8per[tiempo8per$indper==1,]
tiempo8=rbind(tiempo8[tiempo8$indper==0,], tiempo8per)

tiempo8=tiempo8[,1:12]
tiempo8=tiempo8[order(tiempo8$ind),]

beta16=matrix(ncol=3,nrow=1)
gamma16=matrix(ncol=3,nrow=1)
modelo16=zeroinfl(resp8~Trat+c5,data=tiempo8)
beta16[1,]=modelo16$coefficient$count
gamma16[1,]=modelo16$coefficient$zero

datos=cbind(tiempo8[,1:8],datos[,9:10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 17
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo9=cbind(datos[,1:9],datos[,10])
names(tiempo9)=c("resp1", "resp2", "resp3","resp4", "resp5","resp6","resp7",
"resp8","resp9","Trat")

tiempo9$indper=ifelse(!is.na(tiempo9$resp9),0,1)
tiempo9$ind=1:nrow(tiempo9)
tiempo9per=tiempo9[is.na(tiempo9$resp9),]

tiempo9per0=tiempo9per
tiempo9per1=tiempo9per
tiempo9per2=tiempo9per
tiempo9per3=tiempo9per
tiempo9per4=tiempo9per
tiempo9per5=tiempo9per
tiempo9per6=tiempo9per
tiempo9per7=tiempo9per
```

```
tiempo9per8=tiempo9per
tiempo9per9=tiempo9per
tiempo9per10=tiempo9per

tiempo9per0$resp9=0
tiempo9per1$resp9=1
tiempo9per2$resp9=2
tiempo9per3$resp9=3
tiempo9per4$resp9=4
tiempo9per5$resp9=5
tiempo9per6$resp9=6
tiempo9per7$resp9=7
tiempo9per8$resp9=8
tiempo9per9$resp9=9
tiempo9per10$resp9=10

tiempo9per=rbind(tiempo9per0,tiempo9per1,tiempo9per2,
tiempo9per3,tiempo9per4,tiempo9per5,
tiempo9per6,tiempo9per7,tiempo9per8,
tiempo9per9,tiempo9per10)

tiempo9per$indper=ifelse(tiempo9per$resp9==0,1,ifelse(tiempo9per$resp9==1,2,
ifelse(tiempo9per$resp9==2,3,
ifelse(tiempo9per$resp9==3,4,ifelse(tiempo9per$resp9==4,5,
ifelse(tiempo9per$resp9==5,6,ifelse(tiempo9per$resp9==6,7,
ifelse(tiempo9per$resp9==7,8,ifelse(tiempo9per$resp9==8,9,
ifelse(tiempo9per$resp9==9,10,ifelse(tiempo9per$resp9==10,11,"joder")))))))))))

tiempo9=rbind(tiempo9[!is.na(tiempo9$resp9),],tiempo9per)
tiempo9=tiempo9[order(tiempo9$ind),]

tres=data.frame(tiempo9$resp4,tiempo9$resp5, tiempo9$resp6,tiempo9$resp7,tiempo9$resp8)
c6=princomp(tres,scannf=F)$scores[,1]

N=nrow(tiempo9)
x=cbind(1,tiempo9$Trat, c6)
z=cbind(1,tiempo9$Trat, c6)

iterE=100
beta17=matrix(ncol=3,nrow=iterE+1)
```

```
beta17[1,]=beta16[1,]
gamma17=matrix(ncol=3,nrow=iterE+1)
gamma17[1,]=gamma16[1,]

for( i in 1:iterE){
mu=exp(x%*%beta17[i,])
pi=exp(z%*%gamma17[i,])/( 1+exp(z%*%gamma17[i,]))
pes=function(q){
ifelse(q==1, pi+(1-pi)*exp(-mu),(1-pi)*exp(-mu)*mu^{q-1}/factorial(q-1))
}
pesos=ifelse(tiempo9$indper==0,1,
ifelse(tiempo9$indper==1,pes(1),
ifelse(tiempo9$indper==2,pes(2),
ifelse(tiempo9$indper==3,pes(3),
ifelse(tiempo9$indper==4,pes(4),
ifelse(tiempo9$indper==5,pes(5),
ifelse(tiempo9$indper==6,pes(6),
ifelse(tiempo9$indper==7,pes(7),
ifelse(tiempo9$indper==8,pes(8),
ifelse(tiempo9$indper==9,pes(9),
ifelse(tiempo9$indper==10,pes(10),pes(11)))))))))))
modelo17=zeroinfl(resp9~Trat+c6,data=tiempo9,weights=pesos)
beta17[i+1,]=modelo17$coefficient$count
gamma17[i+1,]=modelo17$coefficient$zero
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 18
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo9=cbind(tiempo9,c6,mu,pi)
tiempo9per=tiempo9[!tiempo9$indper==0,]
tiempo9per$resp9=ifelse(tiempo9per$pi>0.80,0,as.integer(tiempo9per$mu)  )
tiempo9per=tiempo9per[tiempo9per$indper==1,]
tiempo9=rbind(tiempo9[tiempo9$indper==0,], tiempo9per)

tiempo9=tiempo9[,1:13]
tiempo9=tiempo9[order(tiempo9$ind),]

beta18=matrix(ncol=3,nrow=1)
gamma18=matrix(ncol=3,nrow=1)
modelo18=zeroinfl(resp9~Trat+c6,data=tiempo9)
```

```
beta18[1,]=modelo18$coefficient$count
gamma18[1,]=modelo18$coefficient$zero

datos=cbind(tiempo9[,1:9],datos[,10])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "y4","y5","y6","y7","y8","y9", "Trat")

ndatosnoexitosos=length(which((originales-datos)!=0))

datos2=data.frame(melt(datos,id=c("Trat")),rep(seq(1:24),9),c(rep(1,24),rep(2,24),rep(3,24
datos2=datos2[,-2]
names(datos2)=c("trat","resp","ind","sem")
final=glmmTMB(resp~trat+sem+(1|ind),zi=~trat+sem,data=datos2,family=poisson)
summary(final)

ndatosperdidos
ndatosnoexitosos
```

C. Código R algoritmo Binomial Negativa cero inflada: Datos del Polen

```
library(gdata)
library(splitstackshape)
library(MASS)
library(ade4)
library(pscl)
library(lmtest)
library(reshape)
library(glmmTMB)

options(warn=-1)

porcentaje=0.2

n=nrow(datos)
missing=runif(n)<porcentaje
datos[,1][missing]=NA
missing=runif(n)<porcentaje
datos[,2][missing]=NA
missing=runif(n)<porcentaje
datos[,3][missing]=NA
missing=runif(n)<porcentaje

perdidos=datos

ndatosperdidos=sum(is.na(perdidos))

tiempo1=cbind(datos[,1],datos[,4:7])
tiempo1=data.frame(tiempo1)
names(tiempo1)=c("resp","lugar", "estado", "nido", "hora")
```

```
tiempo1$indper=ifelse(!is.na(tiempo1$resp),0,1)
tiempo1$ind=1:nrow(tiempo1)
tiempo1per=tiempo1[is.na(tiempo1$resp),]

tiempo1per0=tiempo1per
tiempo1per1=tiempo1per
tiempo1per2=tiempo1per
tiempo1per3=tiempo1per
tiempo1per4=tiempo1per
tiempo1per5=tiempo1per
tiempo1per6=tiempo1per
tiempo1per7=tiempo1per
tiempo1per8=tiempo1per
tiempo1per9=tiempo1per
tiempo1per10=tiempo1per
tiempo1per11=tiempo1per
tiempo1per12=tiempo1per
tiempo1per13=tiempo1per
tiempo1per14=tiempo1per
tiempo1per15=tiempo1per
tiempo1per16=tiempo1per
tiempo1per17=tiempo1per
tiempo1per18=tiempo1per
tiempo1per19=tiempo1per
tiempo1per20=tiempo1per
tiempo1per21=tiempo1per
tiempo1per22=tiempo1per
tiempo1per23=tiempo1per
tiempo1per24=tiempo1per
tiempo1per25=tiempo1per
tiempo1per26=tiempo1per
tiempo1per27=tiempo1per
tiempo1per28=tiempo1per
tiempo1per29=tiempo1per
tiempo1per30=tiempo1per
tiempo1per31=tiempo1per
tiempo1per32=tiempo1per
tiempo1per33=tiempo1per
tiempo1per34=tiempo1per
```

```
tiempo1per35=tiempo1per
tiempo1per36=tiempo1per
tiempo1per37=tiempo1per
tiempo1per38=tiempo1per
tiempo1per39=tiempo1per
tiempo1per40=tiempo1per
tiempo1per41=tiempo1per
tiempo1per42=tiempo1per
tiempo1per43=tiempo1per
tiempo1per44=tiempo1per
tiempo1per45=tiempo1per
tiempo1per46=tiempo1per
tiempo1per47=tiempo1per
tiempo1per48=tiempo1per
tiempo1per49=tiempo1per
tiempo1per50=tiempo1per

tiempo1per0$resp=0
tiempo1per1$resp=1
tiempo1per2$resp=2
tiempo1per3$resp=3
tiempo1per4$resp=4
tiempo1per5$resp=5
tiempo1per6$resp=6
tiempo1per7$resp=7
tiempo1per8$resp=8
tiempo1per9$resp=9
tiempo1per10$resp=10
tiempo1per11$resp=11
tiempo1per12$resp=12
tiempo1per13$resp=13
tiempo1per14$resp=14
tiempo1per15$resp=15
tiempo1per16$resp=16
tiempo1per17$resp=17
tiempo1per18$resp=18
tiempo1per19$resp=19
tiempo1per20$resp=20
tiempo1per21$resp=21
tiempo1per22$resp=22
```

```
tiempo1per23$resp=23
tiempo1per24$resp=24
tiempo1per25$resp=25
tiempo1per26$resp=26
tiempo1per27$resp=27
tiempo1per28$resp=28
tiempo1per29$resp=29
tiempo1per30$resp=30
tiempo1per31$resp=31
tiempo1per32$resp=32
tiempo1per33$resp=33
tiempo1per34$resp=34
tiempo1per35$resp=35
tiempo1per36$resp=36
tiempo1per37$resp=37
tiempo1per38$resp=38
tiempo1per39$resp=39
tiempo1per40$resp=40
tiempo1per41$resp=41
tiempo1per42$resp=42
tiempo1per43$resp=43
tiempo1per44$resp=44
tiempo1per45$resp=45
tiempo1per46$resp=46
tiempo1per47$resp=47
tiempo1per48$resp=48
tiempo1per49$resp=49
tiempo1per50$resp=50

tiempo1per=rbind(tiempo1per0,tiempo1per1,tiempo1per2,
tiempo1per3,tiempo1per4,tiempo1per5,
tiempo1per6,tiempo1per7,tiempo1per8,
tiempo1per9,tiempo1per10,
tiempo1per11,tiempo1per12,tiempo1per13,tiempo1per14,
tiempo1per15,tiempo1per16,tiempo1per17,tiempo1per18,
tiempo1per19,tiempo1per21,tiempo1per22,tiempo1per23,
tiempo1per24,tiempo1per25,tiempo1per26,tiempo1per27,
tiempo1per28,tiempo1per29,tiempo1per30,tiempo1per31,
tiempo1per32,tiempo1per33,tiempo1per34,tiempo1per35,
tiempo1per36,tiempo1per37,tiempo1per38,tiempo1per39,
```

```
tiempo1per40,tiempo1per41,tiempo1per42,tiempo1per43,
tiempo1per44,tiempo1per45,tiempo1per46,tiempo1per47,
tiempo1per48,tiempo1per49,tiempo1per50)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

tiempo1per$indper=ifelse(tiempo1per$resp==0,1,ifelse(tiempo1per$resp==1,2,
ifelse(tiempo1per$resp==2,3,
ifelse(tiempo1per$resp==3,4,  ifelse(tiempo1per$resp==4,5,
ifelse(tiempo1per$resp==5,6,  ifelse(tiempo1per$resp==6,7,
ifelse(tiempo1per$resp==7,8,  ifelse(tiempo1per$resp==8,9,
ifelse(tiempo1per$resp==9,10, ifelse(tiempo1per$resp==10,11,
ifelse(tiempo1per$resp==11,12,ifelse(tiempo1per$resp==12,13,
ifelse(tiempo1per$resp==13,14,ifelse(tiempo1per$resp==14,15,
ifelse(tiempo1per$resp==15,16,ifelse(tiempo1per$resp==16,17,
ifelse(tiempo1per$resp==17,18,ifelse(tiempo1per$resp==18,19,
ifelse(tiempo1per$resp==19,20,ifelse(tiempo1per$resp==20,21,
ifelse(tiempo1per$resp==21,22,ifelse(tiempo1per$resp==22,23,
ifelse(tiempo1per$resp==23,24,ifelse(tiempo1per$resp==24,25,
ifelse(tiempo1per$resp==25,26,ifelse(tiempo1per$resp==26,27,
ifelse(tiempo1per$resp==27,28,ifelse(tiempo1per$resp==28,29,
ifelse(tiempo1per$resp==29,30,ifelse(tiempo1per$resp==30,31,
ifelse(tiempo1per$resp==31,32,ifelse(tiempo1per$resp==32,33,
ifelse(tiempo1per$resp==33,34,ifelse(tiempo1per$resp==34,35,
ifelse(tiempo1per$resp==35,36,ifelse(tiempo1per$resp==36,37,
ifelse(tiempo1per$resp==37,38,ifelse(tiempo1per$resp==38,39,
ifelse(tiempo1per$resp==39,40,ifelse(tiempo1per$resp==40,41,
ifelse(tiempo1per$resp==41,42,ifelse(tiempo1per$resp==42,43,
ifelse(tiempo1per$resp==43,44,ifelse(tiempo1per$resp==44,45,
ifelse(tiempo1per$resp==45,46,ifelse(tiempo1per$resp==46,47,
ifelse(tiempo1per$resp==47,48,ifelse(tiempo1per$resp==48,49,
ifelse(tiempo1per$resp==49,50,51))))))))))))))))))))))))))))))))))))))))))))))))))

tiempo1=rbind(tiempo1[!is.na(tiempo1$resp),],tiempo1per)
tiempo1=tiempo1[order(tiempo1$ind),]
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 1
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
iterE=30
beta<- matrix(ncol=5,nrow=iterE+1)
```

```
gamma<- matrix(ncol=5,nrow=iterE+1)
theta=matrix(ncol=1,nrow=iterE+1)
modelo=zeroinfl(resp~lugar+estado+nido+hora,dist="negbin",data=tiempo1)
beta[1,]=modelo$coef$count
gamma[1,]=modelo$coef$zero
theta[1,]=modelo$theta

k=49
N1=nrow(tiempo1)
x=cbind(1,tiempo1$lugar, tiempo1$estado, tiempo1$nido, tiempo1$hora)
z=cbind(1,tiempo1$lugar, tiempo1$estado, tiempo1$nido, tiempo1$hora)
q=k-1

for( i in 1:iterE){
lambda=exp(x%*%beta[i,])
pi=exp(z%*%gamma[i,])/( 1+exp(z%*%gamma[i,]))
pes=function(q){
ifelse(q==0,pi+(1-pi)*(theta[i,]/(theta[i,]+lambda))^{theta[i,]},
(1-pi)*(gamma(q+theta[i,])/(gamma(q+1)*gamma(theta[i,])))*
((theta[i,]/(theta[i,]+lambda))^{theta[i,]})*
(lambda/(theta[i,]+lambda))^{q}
)
}
pesos=ifelse(tiempo1$indper==0,1,pes(as.numeric(tiempo1$resp)))
modelo1=zeroinfl(resp~lugar+estado+nido+hora,dist="negbin",data=tiempo1,weights=pesos)
beta[i+1,]=modelo1$coefficient$count
gamma[i+1,]=modelo1$coefficient$zero
theta[i+1,]=modelo$theta
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 2
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo1=cbind(tiempo1,lambda, pi)
tiempo1per=tiempo1[!tiempo1$indper==0,]
tiempo1per$resp=ifelse(tiempo1per$pi>0.5,0,as.integer(tiempo1per$lambda)  )
tiempo1per=tiempo1per[tiempo1per$indper==1,]
tiempo1=rbind(tiempo1[tiempo1$indper==0,], tiempo1per)
tiempo1=tiempo1[,1:7]
tiempo1=tiempo1[order(tiempo1$ind),]
```

```
beta2=matrix(ncol=5,nrow=1)
gamma2=matrix(ncol=5, nrow=1)
theta2=matrix(ncol=1,nrow=1)
modelo2=zeroinfl(resp~lugar+estado+nido+hora,dist="negbin",data=tiempo1)
beta2[1,]=modelo2$coefficient$count
gamma2[1,]=modelo2$coefficient$zero
theta2[1,]=modelo2$theta

datos=cbind(tiempo1[,1],datos[,2:7])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3","lugar", "estado", "nido","hora")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 3
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo2=cbind(datos[,1:2],datos[,4:7])
names(tiempo2)=c("resp1","resp2","lugar", "estado", "nido", "hora")

tiempo2$indper=ifelse(!is.na(tiempo2$resp2),0,1)
tiempo2$ind=1:nrow(tiempo2)
tiempo2per=tiempo2[is.na(tiempo2$resp2),]

tiempo2per0=tiempo2per
tiempo2per1=tiempo2per
tiempo2per2=tiempo2per
tiempo2per3=tiempo2per
tiempo2per4=tiempo2per
tiempo2per5=tiempo2per
tiempo2per6=tiempo2per
tiempo2per7=tiempo2per
tiempo2per8=tiempo2per
tiempo2per9=tiempo2per
tiempo2per10=tiempo2per
tiempo2per11=tiempo2per
tiempo2per12=tiempo2per
tiempo2per13=tiempo2per
tiempo2per14=tiempo2per
tiempo2per15=tiempo2per
tiempo2per16=tiempo2per
tiempo2per17=tiempo2per
tiempo2per18=tiempo2per
```

```
tiempo2per19=tiempo2per
tiempo2per20=tiempo2per
tiempo2per21=tiempo2per
tiempo2per22=tiempo2per
tiempo2per23=tiempo2per
tiempo2per24=tiempo2per
tiempo2per25=tiempo2per
tiempo2per26=tiempo2per
tiempo2per27=tiempo2per
tiempo2per28=tiempo2per
tiempo2per29=tiempo2per
tiempo2per30=tiempo2per
tiempo2per31=tiempo2per
tiempo2per32=tiempo2per
tiempo2per33=tiempo2per
tiempo2per34=tiempo2per
tiempo2per35=tiempo2per
tiempo2per36=tiempo2per
tiempo2per37=tiempo2per
tiempo2per38=tiempo2per
tiempo2per39=tiempo2per
tiempo2per40=tiempo2per
tiempo2per41=tiempo2per
tiempo2per42=tiempo2per
tiempo2per43=tiempo2per
tiempo2per44=tiempo2per
tiempo2per45=tiempo2per
tiempo2per46=tiempo2per
tiempo2per47=tiempo2per
tiempo2per48=tiempo2per
tiempo2per49=tiempo2per
tiempo2per50=tiempo2per

tiempo2per0$resp2=0
tiempo2per1$resp2=1
tiempo2per2$resp2=2
tiempo2per3$resp2=3
tiempo2per4$resp2=4
tiempo2per5$resp2=5
tiempo2per6$resp2=6
```

```
tiempo2per7$resp2=7
tiempo2per8$resp2=8
tiempo2per9$resp2=9
tiempo2per10$resp2=10
tiempo2per11$resp2=11
tiempo2per12$resp2=12
tiempo2per13$resp2=13
tiempo2per14$resp2=14
tiempo2per15$resp2=15
tiempo2per16$resp2=16
tiempo2per17$resp2=17
tiempo2per18$resp2=18
tiempo2per19$resp2=19
tiempo2per20$resp2=20
tiempo2per21$resp2=21
tiempo2per22$resp2=22
tiempo2per23$resp2=23
tiempo2per24$resp2=24
tiempo2per25$resp2=25
tiempo2per26$resp2=26
tiempo2per27$resp2=27
tiempo2per28$resp2=28
tiempo2per29$resp2=29
tiempo2per30$resp2=30
tiempo2per31$resp2=31
tiempo2per32$resp2=32
tiempo2per33$resp2=33
tiempo2per34$resp2=34
tiempo2per35$resp2=35
tiempo2per36$resp2=36
tiempo2per37$resp2=37
tiempo2per38$resp2=38
tiempo2per39$resp2=39
tiempo2per40$resp2=40
tiempo2per41$resp2=41
tiempo2per42$resp2=42
tiempo2per43$resp2=43
tiempo2per44$resp2=44
tiempo2per45$resp2=45
tiempo2per46$resp2=46
```

```
tiempo2per47$resp2=47
tiempo2per48$resp2=48
tiempo2per49$resp2=49
tiempo2per50$resp2=50

tiempo2per=rbind(tiempo2per0,tiempo2per1,tiempo2per2,
tiempo2per3,tiempo2per4,tiempo2per5,
tiempo2per6,tiempo2per7,tiempo2per8,
tiempo2per9,tiempo2per10,
tiempo2per11,tiempo2per12,tiempo2per13,tiempo2per14,
tiempo2per15,tiempo2per16,tiempo2per17,tiempo2per18,
tiempo2per19,tiempo2per21,tiempo2per22,tiempo2per23,
tiempo2per24,tiempo2per25,tiempo2per26,tiempo2per27,
tiempo2per28,tiempo2per29,tiempo2per30,tiempo2per31,
tiempo2per32,tiempo2per33,tiempo2per34,tiempo2per35,
tiempo2per36,tiempo2per37,tiempo2per38,tiempo2per39,
tiempo2per40,tiempo2per41,tiempo2per42,tiempo2per43,
tiempo2per44,tiempo2per45,tiempo2per46,tiempo2per47,
tiempo2per48,tiempo2per49,tiempo2per50)

tiempo2per$indper=ifelse(tiempo2per$resp2==0,1,ifelse(tiempo2per$resp2==1,2,
ifelse(tiempo2per$resp2==2,3,
ifelse(tiempo2per$resp2==3,4,ifelse(tiempo2per$resp2==4,5,
ifelse(tiempo2per$resp2==5,6,ifelse(tiempo2per$resp2==6,7,
ifelse(tiempo2per$resp2==7,8,ifelse(tiempo2per$resp2==8,9,
ifelse(tiempo2per$resp2==9,10,ifelse(tiempo2per$resp2==10,11,
ifelse(tiempo2per$resp2==11,12,ifelse(tiempo2per$resp2==12,13,
ifelse(tiempo2per$resp2==13,14,ifelse(tiempo2per$resp2==14,15,
ifelse(tiempo2per$resp2==15,16,ifelse(tiempo2per$resp2==16,17,
ifelse(tiempo2per$resp2==17,18,ifelse(tiempo2per$resp2==18,19,
ifelse(tiempo2per$resp2==19,20,ifelse(tiempo2per$resp2==20,21,
ifelse(tiempo2per$resp2==21,22,ifelse(tiempo2per$resp2==22,23,
ifelse(tiempo2per$resp2==23,24,ifelse(tiempo2per$resp2==24,25,
ifelse(tiempo2per$resp2==25,26,ifelse(tiempo2per$resp2==26,27,
ifelse(tiempo2per$resp2==27,28,ifelse(tiempo2per$resp2==28,29,
ifelse(tiempo2per$resp2==29,30,ifelse(tiempo2per$resp2==30,31,
ifelse(tiempo2per$resp2==31,32,ifelse(tiempo2per$resp2==32,33,
ifelse(tiempo2per$resp2==33,34,ifelse(tiempo2per$resp2==34,35,
ifelse(tiempo2per$resp2==35,36,ifelse(tiempo2per$resp2==36,37,
ifelse(tiempo2per$resp2==37,38,ifelse(tiempo2per$resp2==38,39,
```

```
ifelse(tiempo2per$resp2==39,40,ifelse(tiempo2per$resp2==40,41,
ifelse(tiempo2per$resp2==41,42,ifelse(tiempo2per$resp2==42,43,
ifelse(tiempo2per$resp2==43,44,ifelse(tiempo2per$resp2==44,45,
ifelse(tiempo2per$resp2==45,46,ifelse(tiempo2per$resp2==46,47,
ifelse(tiempo2per$resp2==47,48,ifelse(tiempo2per$resp2==48,49,
ifelse(tiempo2per$resp2==49,50,51)))))))))))))))))))))))))))))))))))))))))))))))))))))

tiempo2=rbind(tiempo2[!is.na(tiempo2$resp2),],tiempo2per)
tiempo2=tiempo2[order(tiempo2$ind),]

iterE=30
beta3<- matrix(ncol=6,nrow=iterE+1)
gamma3<- matrix(ncol=6,nrow=iterE+1)
theta3=matrix(ncol=1,nrow=iterE+1)
modelo3=zeroinfl(resp2~resp1+lugar+estado+nido+hora,dist="negbin",data=tiempo2)
beta3[1,]=modelo3$coefficient$count
gamma3[1,]=modelo3$coefficient$zero
theta3[1,]=modelo3$theta

k=50
N1=nrow(tiempo2)
X=cbind(1,tiempo2$resp1, tiempo2$lugar, tiempo2$estado,tiempo2$nido, tiempo2$hora)
z=cbind(1,tiempo2$resp1, tiempo2$lugar, tiempo2$estado,tiempo2$nido, tiempo2$hora)
q=k-1

for( i in 1:iterE){
lambda=exp(X%*%beta3[i,])
pi=exp(z%*%gamma3[i,])/( 1+exp(z%*%gamma3[i,]))
pes=function(q){
ifelse(q==0,pi+(1-pi)*(theta3[i,]/(theta3[i,]+lambda))^{theta3[i,]},
(1-pi)*(gamma(q+theta3[i,])/(gamma(q+1)*gamma(theta3[i,])))*
((theta3[i,]/(theta3[i,]+lambda))^{theta3[i,]})*
(lambda/(theta3[i,]+lambda))^{q}
)
}
pesos=ifelse(tiempo2$indper==0,1,pes(as.numeric(tiempo2$resp2)))
modelo4=zeroinfl(resp2~resp1+lugar+estado+nido+hora,dist="negbin",data=tiempo2,weights
beta3[i+1,]=modelo4$coefficient$count
gamma3[i+1,]=modelo4$coefficient$zero
theta3[i+1,]=modelo4$theta
```

```
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 4
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo2=cbind(tiempo2,lambda,pi)
tiempo2per=tiempo2[!tiempo2$indper==0,]
tiempo2per$resp2=ifelse(tiempo2per$pi>0.10,0,as.integer(tiempo2per$lambda)  )
tiempo2per=tiempo2per[tiempo2per$indper==1,]
tiempo2=rbind(tiempo2[tiempo2$indper==0,], tiempo2per)
tiempo2=tiempo2[,1:8]
tiempo2=tiempo2[order(tiempo2$ind),]

beta4=matrix(ncol=6,nrow=1)
gamma4=matrix(ncol=6,nrow=1)
theta4=matrix(ncol=1,nrow=1)
modelo5=zeroinfl(resp2~resp1+lugar+estado+nido+hora,dist="negbin",data=tiempo2)
beta4[1,]=modelo5$coefficient$count
gamma4[1,]=modelo5$coefficient$zero
theta4[1,]=modelo5$theta

datos=cbind(tiempo2[,1:2],datos[,3:7])
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3","lugar", "estado", "nido", "hora")
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 5
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo3=datos
names(tiempo3)=c("resp1", "resp2", "resp3","lugar", "estado", "nido", "hora")

tiempo3$indper=ifelse(!is.na(tiempo3$resp3),0,1)
tiempo3$ind=1:nrow(tiempo3)
tiempo3per=tiempo3[is.na(tiempo3$resp3),]

tiempo3per0=tiempo3per
tiempo3per1=tiempo3per
tiempo3per2=tiempo3per
tiempo3per3=tiempo3per
tiempo3per4=tiempo3per
tiempo3per5=tiempo3per
tiempo3per6=tiempo3per
```

```
tiempo3per7=tiempo3per
tiempo3per8=tiempo3per
tiempo3per9=tiempo3per
tiempo3per10=tiempo3per
tiempo3per11=tiempo3per
tiempo3per12=tiempo3per
tiempo3per13=tiempo3per
tiempo3per14=tiempo3per
tiempo3per15=tiempo3per
tiempo3per16=tiempo3per
tiempo3per17=tiempo3per
tiempo3per18=tiempo3per
tiempo3per19=tiempo3per
tiempo3per20=tiempo3per
tiempo3per21=tiempo3per
tiempo3per22=tiempo3per
tiempo3per23=tiempo3per
tiempo3per24=tiempo3per
tiempo3per25=tiempo3per
tiempo3per26=tiempo3per
tiempo3per27=tiempo3per
tiempo3per28=tiempo3per
tiempo3per29=tiempo3per
tiempo3per30=tiempo3per
tiempo3per31=tiempo3per
tiempo3per32=tiempo3per
tiempo3per33=tiempo3per
tiempo3per34=tiempo3per
tiempo3per35=tiempo3per
tiempo3per36=tiempo3per
tiempo3per37=tiempo3per
tiempo3per38=tiempo3per
tiempo3per39=tiempo3per
tiempo3per40=tiempo3per
tiempo3per41=tiempo3per
tiempo3per42=tiempo3per
tiempo3per43=tiempo3per
tiempo3per44=tiempo3per
tiempo3per45=tiempo3per
tiempo3per46=tiempo3per
```

```
tiempo3per47=tiempo3per
tiempo3per48=tiempo3per
tiempo3per49=tiempo3per
tiempo3per50=tiempo3per

tiempo3per0$resp3=0
tiempo3per1$resp3=1
tiempo3per2$resp3=2
tiempo3per3$resp3=3
tiempo3per4$resp3=4
tiempo3per5$resp3=5
tiempo3per6$resp3=6
tiempo3per7$resp3=7
tiempo3per8$resp3=8
tiempo3per9$resp3=9
tiempo3per10$resp3=10
tiempo3per11$resp3=11
tiempo3per12$resp3=12
tiempo3per13$resp3=13
tiempo3per14$resp3=14
tiempo3per15$resp3=15
tiempo3per16$resp3=16
tiempo3per17$resp3=17
tiempo3per18$resp3=18
tiempo3per19$resp3=19
tiempo3per20$resp3=20
tiempo3per21$resp3=21
tiempo3per22$resp3=22
tiempo3per23$resp3=23
tiempo3per24$resp3=24
tiempo3per25$resp3=25
tiempo3per26$resp3=26
tiempo3per27$resp3=27
tiempo3per28$resp3=28
tiempo3per29$resp3=29
tiempo3per30$resp3=30
tiempo3per31$resp3=31
tiempo3per32$resp3=32
tiempo3per33$resp3=33
tiempo3per34$resp3=34
```

```
tiempo3per35$resp3=35
tiempo3per36$resp3=36
tiempo3per37$resp3=37
tiempo3per38$resp3=38
tiempo3per39$resp3=39
tiempo3per40$resp3=40
tiempo3per41$resp3=41
tiempo3per42$resp3=42
tiempo3per43$resp3=43
tiempo3per44$resp3=44
tiempo3per45$resp3=45
tiempo3per46$resp3=46
tiempo3per47$resp3=47
tiempo3per48$resp3=48
tiempo3per49$resp3=49
tiempo3per50$resp3=50

tiempo3per=rbind(tiempo3per0,tiempo3per1,tiempo3per2,
tiempo3per3,tiempo3per4,tiempo3per5,
tiempo3per6,tiempo3per7,tiempo3per8,
tiempo3per9,tiempo3per10,
tiempo3per11,tiempo3per12,tiempo3per13,tiempo3per14,
tiempo3per15,tiempo3per16,tiempo3per17,tiempo3per18,
tiempo3per19,tiempo3per21,tiempo3per22,tiempo3per23,
tiempo3per24,tiempo3per25,tiempo3per26,tiempo3per27,
tiempo3per28,tiempo3per29,tiempo3per30,tiempo3per31,
tiempo3per32,tiempo3per33,tiempo3per34,tiempo3per35,
tiempo3per36,tiempo3per37,tiempo3per38,tiempo3per39,
tiempo3per40,tiempo3per41,tiempo3per42,tiempo3per43,
tiempo3per44,tiempo3per45,tiempo3per46,tiempo3per47,
tiempo3per48,tiempo3per49,tiempo3per50)

tiempo3per$indper=ifelse(tiempo3per$resp3==0,1,ifelse(tiempo3per$resp3==1,2,
ifelse(tiempo3per$resp3==2,3,
ifelse(tiempo3per$resp3==3,4,ifelse(tiempo3per$resp3==4,5,
ifelse(tiempo3per$resp3==5,6,ifelse(tiempo3per$resp3==6,7,
ifelse(tiempo3per$resp3==7,8,ifelse(tiempo3per$resp3==8,9,
ifelse(tiempo3per$resp3==9,10,ifelse(tiempo3per$resp3==10,11,
ifelse(tiempo3per$resp3==11,12,ifelse(tiempo3per$resp3==12,13,
ifelse(tiempo3per$resp3==13,14,ifelse(tiempo3per$resp3==14,15,
```

```
ifelse(tiempo3per$resp3==15,16,ifelse(tiempo3per$resp3==16,17,
ifelse(tiempo3per$resp3==17,18,ifelse(tiempo3per$resp3==18,19,
ifelse(tiempo3per$resp3==19,20,ifelse(tiempo3per$resp3==20,21,
ifelse(tiempo3per$resp3==21,22,ifelse(tiempo3per$resp3==22,23,
ifelse(tiempo3per$resp3==23,24,ifelse(tiempo3per$resp3==24,25,
ifelse(tiempo3per$resp3==25,26,ifelse(tiempo3per$resp3==26,27,
ifelse(tiempo3per$resp3==27,28,ifelse(tiempo3per$resp3==28,29,
ifelse(tiempo3per$resp3==29,30,ifelse(tiempo3per$resp3==30,31,
ifelse(tiempo3per$resp3==31,32,ifelse(tiempo3per$resp3==32,33,
ifelse(tiempo3per$resp3==33,34,ifelse(tiempo3per$resp3==34,35,
ifelse(tiempo3per$resp3==35,36,ifelse(tiempo3per$resp3==36,37,
ifelse(tiempo3per$resp3==37,38,ifelse(tiempo3per$resp3==38,39,
ifelse(tiempo3per$resp3==39,40,ifelse(tiempo3per$resp3==40,41,
ifelse(tiempo3per$resp3==41,42,ifelse(tiempo3per$resp3==42,43,
ifelse(tiempo3per$resp3==43,44,ifelse(tiempo3per$resp3==44,45,
ifelse(tiempo3per$resp3==45,46,ifelse(tiempo3per$resp3==46,47,
ifelse(tiempo3per$resp3==47,48,ifelse(tiempo3per$resp3==48,49,
ifelse(tiempo3per$resp3==49,50,51))))))))))))))))))))))))))))))))))))))))))))))))))

tiempo3=rbind(tiempo3[!is.na(tiempo3$resp3),],tiempo3per)
tiempo3=tiempo3[order(tiempo3$ind),]

uno=data.frame(tiempo3$resp1,tiempo3$resp2)
c1=princomp(uno,scannf=F)$scores[,1]

N=nrow(tiempo3)
x=cbind(1,tiempo3$lugar, tiempo3$estado,tiempo3$nido,c1, tiempo3$hora)
z=cbind(1,tiempo3$lugar, tiempo3$estado,tiempo3$nido,c1, tiempo3$hora)

iterE=100
beta5=matrix(ncol=6,nrow=iterE+1)
gamma5=matrix(ncol=6,nrow=iterE+1)
theta5=matrix(ncol=1,nrow=iterE+1)
modelo6=zeroinfl(resp3~lugar+estado+nido+c1+hora,dist="negbin",data=tiempo3)
beta5[1,]=modelo6$coefficient$count
gamma5[1,]=modelo6$coefficient$zero
theta5[1,]=modelo6$theta

for( i in 1:iterE){
lambda=exp(x%*%beta5[i,])
```

```
pi=exp(z%*%gamma5[i,])/( 1+exp(z%*%gamma5[i,]))
pes=function(q){
ifelse(q==0,pi+(1-pi)*(theta5[i,]/(theta5[i,]+lambda))^{theta5[i,]},
(1-pi)*(gamma(q+theta5[i,])/(gamma(q+1)*gamma(theta5[i,])))*
((theta5[i,]/(theta5[i,]+lambda))^{theta5[i,]})*
(lambda/(theta5[i,]+lambda))^{q}
)
}
pesos=ifelse(tiempo3$indper==0,1,pes(as.numeric(tiempo3$resp3)))
modelo7=zeroinfl(resp3~lugar+estado+nido+c1+hora,dist="negbin",data=tiempo3,weights=pe
beta5[i+1,]=modelo7$coefficient$count
gamma5[i+1,]=modelo7$coefficient$zero
theta5[i+1,]=modelo7$theta
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Paso 6
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tiempo3=cbind(tiempo3,c1,lambda,pi)
tiempo3per=tiempo3[!tiempo3$indper==0,]
tiempo3per$resp3=ifelse(tiempo3per$pi>0.10,0,as.integer(tiempo3per$lambda)  )
tiempo3per=tiempo3per[tiempo3per$indper==1,]
tiempo3=rbind(tiempo3[tiempo3$indper==0,], tiempo3per)

tiempo3=tiempo3[,1:10]
tiempo3=tiempo3[order(tiempo3$ind),]

beta6=matrix(ncol=6,nrow=1)
gamma6=matrix(ncol=6,nrow=1)
theta6=matrix(ncol=1,nrow=1)
modelo8=zeroinfl(resp3~lugar+estado+nido+floor(c1)+hora,dist="negbin",data=tiempo3)
beta6[1,]=modelo8$coefficient$count
gamma6[1,]=modelo8$coefficient$zero
theta6[1,]=modelo8$theta

datos=tiempo3[,1:7]
datos=data.frame(datos)
names(datos)=c("y1", "y2", "y3", "lugar", "estado", "nido", "hora")

ndatosnoexitosos=length(which((originales-datos)!=0))
```

```
arandano=melt(datos[,1:4],id=c("lugar"))
arandano=arandano[,-2]
names(arandano)=c("lugar","resp")
ind=c(rep(seq(1:120),3))
estado=rep(datos$estado,3)
nido=rep(datos$nido,3)
hora=c(rep(rep(seq(1:12),10),3))
datos2=data.frame(arandano,ind,estado,nido,hora)

final=glmmTMB(resp~hora+estado+(1|ind),zi=~hora+nido,data=datos2,family=nbinom2)
summary(final)

ndatosperdidos
ndatosnoexitosos
```