UNIVERSIDAD NACIONAL DEL COMAHUE

-FACULTAD DE ECONOMÍA-

MAESTRÍA EN ESTADÍSTICA APLICADA

Tesis Final

Estrategias de Modelación Para Datos de Conteo con Exceso de Ceros. Una Ilustración asociada a Ecología de Poblaciones

Alumna: Lic. María Fernanda Menni

Director: Dra. María del Pilar Díaz

Co-Director: MSc. Ing. Darío Fernández

AÑO 2011

Dedico este trabajo a mis hijos, sobre todo a Juan. I., que vivió el proceso desde el más temprano inicio. Y a mis padres que lo hicieron posible siempre, gracias.

AGRADECIMIENTOS

En este proceso de aprendizaje, quiero agradecer a las personas que me acompañaron e hicieron posible la realización de esta tarea, que por momentos fue ardua pero que logre disfrutarla de principio a fin gracias a todos ellos.

A mi directora, María del Pilar Díaz por sus correcciones y su acompañamiento pese a la distancia.

A Darío Fernández, por toda la atención y contención que me brindó en este proceso, por su calidez y su paciencia infinita.

A la Experimental del INTA Alto Valle, que posibilitó mediante la beca de formación profesional el acceso a esta instancia.

Gracias Cecilia Gittins, mi socia, por tu generosidad, tus aportes, tus correcciones, y tu puesta a punto cuando fue necesario, y por tanto más C.

A Alejandro Giayetto, por su buena voluntad, por sus contribuciones, lecturas y relecturas.

A Silivina Garrido, por el material, por su ayuda y buena disposición para todo.

A mi tía Ana María, por su ánimo y su permanente fe en mí.

A todos los que no nombré explícitamente pero están siempre presente dentro y fuera de la institución. Gracias totales.

RESUMEN

A partir de las distintas estrategias propuestas entre los modelos lineales generalizados para variables aleatorias discretas, como es la captura de *Cydia pomonella* (L.) en trampas con distintos tipos de cebos, habiéndose tenido en cuenta la especie donde está ubicada la trampa, así como la generación a la que pertenece la captura, presentados en este trabajo, se observo que el modelo que mejor explica el fenómeno es el Binomial negativo con exceso de ceros (ZINB). La manifestación de superdispersión, pudo ser capturada a través del modelo propuesto. Dando la posibilidad de estimar de manera precisa los parámetros del modelo. Pudiendo así representar el comportamiento del fenómeno para decidir acciones de control y de esta manera evitar las pérdidas económicas que genera la plaga en la región del Alto Valle del río Negro y del Neuquén.

Palabras claves: ***modelos lineales generalizados; superdispersión; Cydia pomonella (L.); ZINB***

ABSTRACT

Since the proposed strategies among the generalized linear models for the discrete random variables, as in the case of captures of by the use of traps with different types of baits, taking into account the place where the traps were located, and also the generation to which the trap belongs to, we observed that the model that better explained the phenomenon was the zero inflated negative binomial model (ZINB). The display of overdispersion, could be captured through the proposed model. Giving the possibility to estimate in a precise way the parameters of the model. This allowed the representation of the phenomenon behavior which allowed to decide control actions and in this way to avoid the economic losses that this plague generates in the region of Alto Valle of Río Negro and Neuquén.

key words: ***generalized linear models; overdispersion; Cydia pomonella (L.); ZINB***

INDICE

INDICE DE TABLAS Y FIGURAS

Tablas

Figuras

INTRODUCCIÓN

Según Linsdey (1995), un recuento o conteo es el número de eventos de una misma variable ocurridos sobre la misma unidad observacional en un intervalo temporal o espacial definido. Los datos de conteo particularmente han sido manejados como generados por familias normales y cuando los supuestos requeridos por la teoría normal no eran satisfechos, las transformaciones de variables constituían una opción valiosa para poder aplicar la metodología clásica (Sturman, 1999). El uso de transformaciones se adecua solamente cuando la escala elegida permite que los supuestos se cumplan y tiene la dificultad que las conclusiones a las que se llega se aplican solamente para las poblaciones transformadas (Mead,Curnow y Hasted, 1993; Díaz, 2008).

Los datos de recuento o conteo, muchas veces generan exceso de ceros asociado a un doble proceso relacionado con la distribución específica de los mismos, produciendo una discrepancia entre la varianza nominal y la real, generalmente la variación de la varianza excede la media, esto es, $Var(y) \rangle \ \mu \ \Rightarrow Var(y) = \phi \cdot \mu$, siendo ϕ un factor multiplicativo de dispersión. Este fenómeno es llamado superdispersión (McCullagh y Nelder, 1989)y permite ampliar la mirada de la modelación respetando la naturaleza *per se* de la distribución de la variable aleatoria en estudio. Estos autores sostienen que este fenómeno es lo usual y la varianza nominal la excepción, la incidencia y el grado de superdispersión que se puede encontrar, depende del campo de aplicación.

El problema que se presentará es modelar el comportamiento promedio de variables aleatorias que representan conteos de individuos, cuyas observaciones contienen gran cantidad de ceros (exceso de observaciones nulas), muchas veces, no debidas necesariamente a la inexistencia del fenómeno que se está estudiando.

Como objetivo general se propone encontrar el modelo de Poisson con superdispersión que explique de manera adecuada la relación media-varianza de

la variable aleatoria que representa el conteo de individuos pertenecientes a poblaciones ecológicas, con patrones de distribución agregada.

Como objetivo específico, la propuesta es modelar la respuesta del conteo de *C. pomonella (L.)*por unidad de trampa, evaluada en la región del Alto Valle del río Negro y del Neuquén y estimar los cambios producidos, a través de la distribución de Poisson y sus generalizaciones en presencia de extra-variación.

> *“Ninguna parte substancial del Universo es lo suficientemente simple como para ser asida y controlada sin una adecuada dosis de abstracción. La abstracción consiste en reemplazar esa parte del Universo que nos interesa, por una representación similar pero de estructura más simple. Por ello, las representaciones abstractas o modelos, son una necesidad para el conocimiento científico”.* (Naylor et al. 1966)

LOS PRINCIPIOS DE LA MODELACIÓN

Un modelo es un esquema analítico deliberadamente simplificado, que representa de manera esquemática y aproximada la realidad observada de un fenómeno complejo como para representarlo idénticamente. (Chiang, 1987).

Un modelo estadístico es aquel que describe un experimento cuyo resultado no queda determinado inequívocamente, a partir de las condiciones en que se lo lleve a cabo. Se conoce como experimento aleatorio, al fenómeno empírico que admite dos o más resultados posibles y no se tienen elementos de juicio suficientes como para poder predecir con exactitud cuál o cuáles de ellos ocurrirán, aunque se los repita bajo las mismas condiciones (Capriglioni, 2005).

En el ámbito científico, la complejidad de los problemas ecológicos y ambientales es ampliamente reconocida. El uso de métodos de análisis y modelado acordes con esa complejidad, son muy útiles para servir de guía en la comprensión del problema y en estudiar posibles soluciones. Los modelos en este entorno coadyuvan a comprender mejor la repartición poblacional en el espacio así como la distribución de las mismas. (Raventós, Segarra y Acevedo, 2005).

Generalmente la simplificación de la realidad se debe a diversas razones, frecuentemente son de tipo económica –cuestiones de costos a la hora de llevar a cabo los diseños experimentales propuestos –; imposibilidad práctica de muestrear poblaciones -de insectos en nuestro caso-; o bien la temporalidad, como estudios con la necesidad de evaluar patrones a largo plazo cuyos resultados no coinciden con los tiempos del investigador; por estos motivos, entre

otros la modelación como herramienta teórica-práctica se vuelve fundamental en algunas áreas de interés y una forma de comprender la realidad con mayor profundidad. Elaborar un modelo, no es un fin en si mismo, sino un medio para aprehender mejor el fenómeno objeto de estudio (Losilla Vidal, 1995; Vives Brosa, 2002). Siguiendo a Jorgensen (1989), la modelación es un arte donde deben combinar el "background" teórico, la experiencia y el error propio del tipo de ensayo que se analiza.

El modelado estadístico es un proceso polietápico intrínsecamente iterativo, en el que se procede a reducir el error en forma progresiva, mediante la comparación de múltiples pares de modelos que pretenden dar cuenta de la misma realidad empírica. El objetivo básico del modelado estadístico es derivar, a partir de la variabilidad observada, un modelo donde la proporción de variabilidad sistemática sea relativamente grande respecto a la variabilidad aleatoria, esto es, un modelo que represente óptimamente la relación entre una variable de respuesta Y, y un número de variables explicativas, minimizando el componente aleatorio (Losilla Vidal, 2002).

Estadísticamente, un modelo se lo puede expresar como una relación funcional que permita explicar una variable cuantitativa a partir de una o más variables que puedan estar relacionadas (Gujarati, 2004). El modelo en su forma más simple es un modelo lineal y contiene un componente determinístico y otro estocástico $Y = \beta X + \varepsilon$, el componente determinístico -βX- hace referencia a la parte sistemática del modelo y contempla las variables explicativas del mismo. El componente aleatorio o estocástico -ε- es llamado error aleatorio o residuo y contempla todos los factores que no se han tenido en cuenta en la parte sistemática, ya sea por omisión o por imposibilidad de observación. La teoría clásica ha basado todos los supuestos imponiendo a la parte estocástica del sistema, un comportamiento normal, varianza homogénea, así como independencia entre las observaciones del mismo; expresado esto de manera

matemática: $Y \approx iid\mathrm{N}\left(E(Y)=\mu_Y;\sigma^2_{Y_i}=\sigma^2_Y\right)$[1]. La técnica analítica asociada es la teoría de mínimos cuadrados clásicos u ordinarios (MCC) la cual asume un único componente de error (Nelder y Wedderburn, 1972). Según estos autores, en presencia de múltiples errores se desarrollaron nuevas metodologías para los diseños experimentales y los modelos de sobrevida; así como el avance de técnicas para modelos no normales incluyó el análisis probit –siendo la binomial, la distribución asumida al componente aleatorio- también como para las tablas de contingencia, la distribución admitida es de tipo multinomial y la parte sistemática del modelo tiene efectos de interés usualmente multiplicativos. Estos modelos fueron agrupados, por los mismos autores, bajo la distribución de Poisson con restricciones, con la característica que el componente sistemático continúa teniendo una estructura lineal. A estos se le sumaron posteriormente modelos basados en la distribución Ji cuadrado y gamma, con la misma característica en su parte sistemática, sumando la condición que el componente aleatorio pertenezca a la familia exponencial, los autores designaron a este grupo de modelos como *Generalized Linear Models*, en nuestro lenguaje Modelos Lineales Generalizados (MLG), cuya estimación se realiza a través de la metodología de Máxima Verosimilitud (MV) que puede ser obtenida a partir de la técnica iterativa de mínimos cuadrados ponderados.

Evolución Histórica

En 1937, Bartlett utiliza la transformación log [y/(1-y)] para analizar proporciones. También Fisher y Yates sugieren en 1938 el uso de esta transformación para analizar datos binarios. El término logit fue introducido por Joseph Berkson en 1944 para designar esta transformación y sus trabajos popularizaron la utilización de la regresión logística. Jerome Cornfield utilizó la

[1] En un modelo lineal la parte estocástica del sistema es el término del error aleatorio; la variable respuesta es aleatoria por ser combinación lineal del mismo en este entorno.

regresión logística para el cálculo de *odds ratio* como valores aproximados del riesgo relativo en estudios de casos y controles.

En 1972, Nelder y Wedderburn unificaron los avances logrados en el campo de los modelos no normales de Dyke y Patterson (1952) y Bliss (1935) y propusieron una alternativa al enfoque clásico, permitiendo otras posibilidades de distribución de la variable respuesta. El proceso de estimación de los parámetros de interés se basa en el método de Máxima Verosimilitud (McCullagh y Nelder, 1989; Dobson, 1990; Ato, 2000). A su vez, como medida de la bondad de ajuste se deja de lado la suma de cuadrados residual y se emplea la *Deviance* utilizada en el ajuste del modelo y en las etapas de diagnóstico.

En 1974, Wedderburn proporcionó la base teórica para los modelos de cuasi-verosimilitud, generalizando lo anterior para datos correlacionados. Jorgensen (1983) introduce la construcción de los modelos de dispersión dando un mayor alcance a la distribución de la variable respuesta. Liang y Zeger (1986) extienden esta metodología a través de las ecuaciones generalizadas de estimación, popularizadas como GEE (*Generalized Estimating Equations*), permitiendo el planteo de estudios longitudinales, temporales o espaciales, asociados a variables aleatorias no normales correlacionadas.

Hastie y Tibshirani (1990) avanzan sobre los Modelos Aditivos Generalizados (GAM) suponiendo que el predictor lineal puede ser formado por funciones semiparamétricas. En 1993, Breslow y Clayton construyen el marco teórico para los Modelos Lineales Generalizados Mixtos (GLMM) pudiendo incluir los efectos aleatorios de tipo normal en el predictor lineal.

A partir del 2004 Skrondal y Rabe-Hesketh continúan los avances teóricos de modelos para este tipo de variables, entre otros, los modelos GLLAMM (*Generalized Linear Latent and Mixed Models*).

"En lugar de transformar los datos para que cumplan las asunciones del modelo de regresión lineal, deberían considerarse modelos de regresión que capturen las características naturales de los datos". (Gardner et al. 1995)

MODELOS LINEALES GENERALIZADOS (MLG)

La generalización de los modelos lineales ha permitido la integración del modelado de datos categóricos y cuantitativos en un mismo entorno (Losilla Vidal, 2002). A su vez, integra mecanismos para abordar la presencia tanto de relaciones lineales como no lineales, entre variables de respuesta y variables explicativas. Los MLG se basan en la familia exponencial (McCullagh y Nelder, 1989).

Las estimaciones y predicciones se enmarcan en los métodos de Máxima Verosimilitud (MV), diferenciándose del método Mínimos Cuadrados Clásicos u ordinarios (MCC) de la teoría clásica.

Suponiendo que las observaciones y -como componente aleatorio *per sé*- provienen de una distribución de probabilidad perteneciente a la familia exponencial

$$f_y(y;\theta,\phi) = \exp\left[\frac{1}{a(\phi)}\{y\theta - b(\theta)\} + c(\phi, y)\right],$$

siendo $a(\cdot), b(\cdot), c(\cdot)$ funciones conocidas, θ un parámetro de posición o canónico y ϕ un parámetro o factor de escala, con $a(\phi) > 0$ generalmente positivo (McCullagh y Nelder, 1989; Demétrio y Hinde, 1998). Quedando incluidas también las familias normales, para σ conocido, bajo esta expresión.

La utilidad de esta parametrización reside en poder expresar la función generadora de momentos como:

$$M_y(t;\theta,\phi)=\exp\left[\frac{1}{a(\phi)}\{b(a(\phi)t+\theta)-b(\theta)\}\right],$$

para posibilitar la obtención a través de esta, del primer momento absoluto –la esperanza matemática- y del momento de segundo orden centrado –la varianza-, al menos de manera asintótica. Estos momentos son los que caracterizan, de manera reducida, toda distribución de probabilidad.

Los cumulantes corresponden a otro tipo de estadísticos de alto orden derivados a partir del $\ln$ de la función generadora de momentos. Su forma matemática es:

$$\ln M_x(t;\theta,\phi)=\varphi(t;\theta,\phi)=\frac{1}{a(\phi)}[b(a(\phi)t+\theta)-b(\theta)]$$

,

Las expresiones de la primera y segunda derivada del log-MV nos permitirá hallar la media y la varianza de $\mathbf{y}$ teniendo en cuenta el factor de escala $a(\phi)$; usando los resultados de Nelder y Wedderburn, (1972, p. 371) para dichas expresiones,

$$\mathrm{E}\left(\partial L/\partial\theta\right)=0, \tag{1}$$

$$\mathrm{E}\left(\partial^2 L/\partial\theta^2\right)=-\mathrm{E}\left(\partial L/\partial\theta\right)^2, \tag{2}$$

entonces se obtiene,

$$\partial L/\partial\theta=\frac{1}{a(\phi)}\{y-b'(\theta)\},$$

que a partir de (1) implica

$$\mu = \mathrm{E}(Y) = b'(\theta);$$

y por lo tanto

$$\partial L \big/ \partial \theta = \frac{1}{a(\phi)}\{y - \mu\}$$

.

De la forma (2), referida a la derivada segunda, se obtiene:

$$\frac{1}{a(\phi)} b''(\theta) = [a(\phi)]^2 V(y) \Rightarrow V(Y) = a(\phi) b''(\theta)$$

,

$$\partial^2 L \big/ \partial \theta^2 = -a(\phi) V(\mu) \quad \text{tal que} \quad V(\mu) = d\mu \big/ d\theta$$

.

Se determinó la media y la varianza de la variable $\mathbf{Y}$ a partir de los resultados conocidos de la función de log-verosimilitud. Habitualmente a la derivada de la función de log-verosimilitud respecto al parámetro θ, se la denomina *score* y se lo denota como $\mathbf{U}$, se presentan a continuación las propiedades de los *scores.* (Nelder y Wedderburn, 1972; Montero-Mercadé, 2004)

Propiedad de primer orden:

$$\mathrm{E}\left[\frac{\partial L}{\partial \theta}\right] = \mathrm{E}(\mathbf{U}) = 0 \qquad (1)$$

,

Propiedad de segundo orden:

$$\begin{aligned} &\mathrm{E}\left[\frac{\partial^2 L}{\partial \theta^2}\right] + \mathrm{E}\left[\left(\frac{\partial L}{\partial \theta}\right)^2\right] = \mathrm{E}(\mathbf{U}') + \mathrm{E}\left[\mathbf{U}^2\right] = 0 \\ &\Rightarrow V[\mathbf{U}] = \mathrm{E}\left[\mathbf{U}^2\right] = -\mathrm{E}[\mathbf{U}'] \end{aligned} \qquad (2)$$

.

Al aplicar los resultados previos, al logaritmo de la función de la familia exponencial:

$$f_y(y;\theta,\phi)=\exp\left[\frac{1}{a(\phi)}\{y\theta-b(\theta)\}+c(\phi,y)\right]$$

,

$$L(y;\theta,\phi)=\log f_y(y;\theta,\phi)=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)$$

.

se obtiene de (1):

$$\mathrm{E}\left[\frac{\partial L}{\partial\theta}\right]=\mathrm{E}(\mathbf{U})=\mathrm{E}\left[\frac{y-b'(\theta)}{a(\phi)}\right]=\frac{\mu-b'(\theta)}{a(\phi)}=0$$
$$b'(\theta)=\mu=\mathrm{E}(Y)$$

,

la esperanza de la variable aleatoria, siendo esta la media, denominada *mu* - μ -.

De la propiedad (2):

$$\mathrm{E}\left[\frac{\partial^2 L}{\partial\theta^2}\right]+\mathrm{E}\left[\left(\frac{\partial L}{\partial\theta}\right)^2\right]=\mathrm{E}(\mathbf{U}')+\mathrm{E}\left[\mathbf{U}^2\right]=$$
$$=\mathrm{E}\left[\frac{-b''(\theta)}{a(\phi)}\right]+\mathrm{E}\left[\left(\frac{y-b'(\theta)}{a(\phi)}\right)^2\right]=$$
$$=\mathrm{E}\left[\frac{-b''(\theta)}{a(\phi)}\right]+\frac{1}{a(\phi)^2}\mathrm{E}\left[\left(Y-\mathrm{E}[Y]^2\right)\right]=$$
$$=\frac{-b''(\theta)}{a(\phi)}+\frac{1}{a(\phi)^2}V[Y]=0$$
$$\Rightarrow V[Y]=a(\phi)b''(\theta)$$

,

$$b''(\theta)=\frac{\partial^2 b(\theta)}{\partial\theta^2}=\frac{\partial b'(\theta)}{\partial\theta}=\frac{\partial\mu}{\partial\theta}$$

,

se deduce la función de varianza de la variable aleatoria, depende del parámetro canónico y por lo tanto de la media, así podemos denotarla como $V(\mu)$.

Aplicando lo anteriormente demostrado a la distribución particular de Poisson, queda expresado como:

$$b(\theta)=\mu=e^{\theta}\begin{cases} b'(\theta)=e^{\theta}=\mu=\mathrm{E}(Y) \\ a(\phi)\cdot b''(\theta)=e^{\theta}=\mu=Var(Y)\Leftrightarrow\phi=1 \end{cases},$$

cuando el efecto de superdispersión está presente, obsérvese que el parámetro de escala es $\phi\neq 1$, modificando la varianza a través del mismo, pero no así el promedio:

$$b(\theta)=\mu=e^{\theta}\begin{cases} b'(\theta)=e^{\theta}=\mu=\mathrm{E}(Y) \\ \phi\cdot b''(\theta)=\phi\cdot e^{\theta}=\phi\cdot Var(\mu)=Var(Y) \end{cases}.$$

Los MLG pueden ser resumidos en un trinomio. Los tres elementos del mismo son la aleatoriedad de la respuesta observada en unidades independientes y_i, siendo el componente aleatorio; el componente sistemático, lineal en parámetros, conocido como predictor lineal $\eta=\mathrm{X}\beta$ y la función de enlace, $g(.)\diagup\ \eta_i=g(\mu_i)$, la cual es una función monótona y diferenciable, que articula ambas componentes (McCullagh y Nelder, 1989).

El Modelo Lineal Generalizado Poisson

Según el trinomio presentado de manera previa:

I. $\mathbf{Y} \approx P(\boldsymbol{\mu};\varphi)$

II. $\boldsymbol{\eta} = \mathbf{X}'\boldsymbol{\beta}$

III. $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \ln(\boldsymbol{\mu})$

siendo I. la especificación de la distribución de la variable aleatoria; Poisson, en esta presentación, con parámetro μ y ϕ como parámetro de escala. En la expresión II. se observa el comportamiento del predictor como combinación lineal entre las variables explicativas $\mathbf{X}$ y los parámetros $\boldsymbol{\beta}$ a ser estimados, las variables independientes pueden ser cuantitativas y producir variaciones simples en el modelo, o de tipo cualitativas y producir un set de variables *dummys* cuyos valores serán 1 para un nivel particular de estudio y 0 para los niveles restantes. Y en el enunciado III, la función monótona y diferenciable, siendo la descriptora de la relación entre la distribución propuesta –componente aleatoria- y el predictor lineal –componente sistemático-.

La distribución de Poisson $f(y,\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \mathrm{I}_{\{0,1,2..n\}}(y)$, como distribución natural para variables de tipo positivas, como los conteos, puede escribirse en términos de la familia exponencial de la siguiente manera:

$$f_{y}(y;\theta,\phi) =$$

$$f(y,\mu) = \exp\left[y\ln(\mu) - \mu - \ln(y!)\right],$$

luego,

$$c(y,\phi) = -\ln(y!) \Leftrightarrow \phi = 1,$$

$$\theta = \ln(\mu) \Leftrightarrow \mu = e^{\theta}.$$

La función de enlace: $\theta = \ln(\mu) = \eta = g(\mu)$,

$\Rightarrow$ inversa de la función de enlace $g^{-1}(\mu) = \eta^{-1} = \exp(\mu)$.

Derivando la función de enlace:

$$\theta = \ln(\mu) = \eta = g(\mu),$$

$$\frac{\partial \theta}{\partial \mu} = \frac{\partial[\ln(\mu)]}{\partial \mu} = \frac{1}{\mu},$$

$$\frac{1}{\mu} = \frac{1}{\exp(x'\beta)} = \mu^{-1} = \exp(x'\beta)^{-1},$$

parametrización que se utilizará en el algoritmo de estimación de MV; así como en la generación de la *Deviance* como medida de la diferencia entre el modelo saturado y el modelo máximo verosímil, un indicativo de la "bondad del ajuste" del modelo que se propone y un criterio alternativo de selección de modelos, estos temas se profundizarán posteriormente.

$$b(\theta) = \mu = e^{\theta} \begin{cases} b'(\theta) = e^{\theta} = \mu = \mathrm{E}(Y) \\ a(\phi) \cdot b''(\theta) = e^{\theta} = \mu = Var(Y) \Leftrightarrow \phi = 1 \end{cases},$$

este fenómeno de igualdad entre media y varianza es conocido como equidispersión, por contrario, cuando el valor de la varianza excede el valor de la media surge el término de superdispersión. Obsérvese que al estar presente el fenómeno de superdispersión, el parámetro de escala es $\phi \neq 1$, modifica la varianza a través del mismo:

$$b(\theta) = \mu = e^{\theta} \begin{cases} b'(\theta) = e^{\theta} = \mu = \mathrm{E}(Y) \\ \phi \cdot b''(\theta) = \phi \cdot e^{\theta} = \phi \cdot Var(\mu) = Var(Y) \end{cases},$$

$a(\phi)$ tomará distinta forma según se piense la función de varianza esté comportándose en el modelo; según McCullagh y Nelder (1989) $a(\phi) = \phi / w$, siendo w la ponderación conocida impuesta a priori, que varía con cada observación.

La media condicional depende de los predictores de acuerdo con el siguiente modelo:

$$\mu = \mathrm{E}(Y|X) = \exp(x_i' \beta_i),$$

al aplicar la función exponencial, se consigue que $\mu_i \geq 0$, siendo esta una de las propiedades de la escala de medida de las variables positivas de conteo (Long, 1997). La función $\mu_i = \exp(x_i' \beta_i)$ es la inversa de $\log(\mu_i) = x_i' \beta_i$ y se la considera una expresión multiplicativa dado las propiedades de los exponentes, a diferencia de los modelos lineales clásicos, que manifiestan los efectos aditivos:

$$\mu_i = \exp(x_i' \beta_i) = \exp(\beta_0 + \beta_j x_j) = \exp(\beta_0) \cdot \exp(\beta_j x_j),$$

esta propiedad es importante, dado que las relaciones que pretende modelar el modelo de regresión Poisson son de tipo no lineal, de forma que posibilita la interpretación de los parámetros directamente sobre la escala de la variable respuesta (Long, 1997; Díaz, 1999).

Si bien la principal característica del modelo de Poisson es la capacidad de capturar la naturaleza discreta y no negativa de la variable de conteo, la limitación proviene de la rigidez impuesta por sus supuestos (Winkelmann, 2000). Una particularidad de este modelo específico es que la relación media-varianza del conjunto de datos observados no se ajusta a la relación media-varianza que caracteriza la distribución subyacente teórica. Por este motivo, la varianza observada es mayor que la varianza nominal, es decir la varianza definida por la distribución de referencia, generando el problema de superdispersión (Losilla, 2002).

Surgen así ampliaciones como el modelo de Poisson generalizado, popularizado binomial negativo; el de Poisson truncado, el modelo de Poisson con ceros aumentados (*Zero Inflated Poisson –ZIP-*) entre otros. Las diferencias entre estos modelos dependen de cómo se plantea la función explícita de probabilidad para la variable cuando esta toma el valor cero (Ridout, *et al* 1991), ya que en consecuencia, queda modificada la varianza asociada al mismo y sus estimaciones. Esto será ampliado en el Cap. V.

Como el objetivo general propuesto fue encontrar el modelo apropiado que explique de manera adecuada la relación media-varianza de la variable aleatoria que representa el conteo de individuos pertenecientes a poblaciones ecológicas, con patrones de distribución agregada. En este trabajo se focalizará en las poblaciones animales –invertebrados, específicamente artrópodos-, en la que los individuos viven en el espacio, en hábitats discontinuos, que pueden ser divididos en unidades discretas como hojas, flores, suelo, frutos o plantas enteras tomadas como unidades de hábitat, siendo habitual que se analicen mediante el uso de las frecuencias obtenidas por los conteos del número de individuos por unidad de hábitat (Cadahia, 1977).

Las poblaciones de invertebrados, específicamente, los lepidópteros tortricidos (como *Cydia pomonella*), presentan generalmente un patrón de distribución agregada (Capuccino y Price, 1995). Este patrón genera en el muestreo estacional una gran cantidad de ceros, ya que dependen directamente tanto de las condiciones climáticas como de la disponibilidad del recurso; así también como comportamientos no lineales. Este tipo de poblaciones permitirán ajustar los modelos propuestos en este trabajo.

"…pese a tan ilustres modos de errar, no [se] ha descifrado el laberinto singular y plural, arduo y distinto…" (Borges, 1969)

GÉNESIS DE LA SUPERDISPERSIÓN

La superdispersión es un fenómeno que se da frecuentemente, por distintos motivos de mala especificación en el trinomio del modelo. Cuando la varianza real supera a la varianza nominal, en otras palabras, la varianza condicional sobrepasa la media condicional y consecuentemente el parámetro de escala $\phi \neq 1$ estamos ante el fenómeno denominado superdispersión. (McCullagh y Nelder, 1989; Hinde y Demetrio, 2005) Es muy común que se dé en la práctica, la incidencia y el grado encontrado dependen mucho del campo de aplicación.

Las razones de la superdispersión suelen ser muy diversas, entre las que se citan:

- Función de media incorrecta

  La función de media del MRP es:

$$\mathrm{E}(Y|X) = \mu_i = \exp(x_i\beta),$$

si denotamos la función media verdadera como μ_0 y el valor esperado respecto a la densidad verdadera como E_0,

$$\mathrm{E}_0(Y|X) = \mu_0 = \mathbf{f}(x_0,\beta_0),$$

la función de media está especificada erróneamente si no existe ninguna β que cumpla $\mu_i = \mathbf{f}(x_i,\beta_0)$ de forma que $\mu_i \neq \mu_0$.

Los errores de especificación pueden ser debidos a la correlación entre las observaciones, implicando redundancia en la información aportada; variables omitidas en el modelo; predictor no lineal en parámetros; error de especificación en la función de enlace, entre otros.

- Heterogeneidad no observada

Cuando las variables regresoras no explican la heterogeneidad individual, estamos frente a esta situación. Las observaciones difieren aleatoriamente de una forma que no es recogida exhaustivamente por las variables explicativas del modelo.

Si especificamos el modelo verdadero como:

$$\tilde{\mu}_i = \exp\left(\mathbf{x_i}\boldsymbol{\beta} + \mathbf{z_i}\boldsymbol{\gamma}\right),$$

siendo $\mathbf{z_i}$ un vector de variables que no han sido incluidas en el modelo usado, podemos derivar la ecuación, siguiendo a Winkelmann (2000) como

$$\tilde{\mu}_i = \mu_i u_i \quad , u_i > 0,$$

donde $\mu_i = \exp\left(x_i\beta\right)$ y $u_i = \exp\left(z_i\gamma\right)$, asumiendo que el término estocástico $u_i > 0$ es independiente de las regresoras. El parámetro $\tilde{\mu}_i$ se vuelve una variable aleatoria. Las dos fuentes de variación en el parámetro de Poisson $\tilde{\mu}_i$ interactúan de forma multiplicativa; la primera fuente de variación es sistemática y depende de las variables explicativas x_i, mientras que la segunda fuente está causada por un efecto aleatorio individual u_i independiente de x_i. Sea $\ln \varepsilon_i = \ln u_i$, entonces $\tilde{\mu}_i = \exp\left(x_i\beta + \varepsilon_i\right)$ de forma que el error es aditivo en la escala logarítmica (Linsday, 1998; Díaz, 1999).

Si se asume que $\mathrm{E}(u\ )=1$ y $Var(u\ )=\sigma_u^2$. Los momentos que caracterizan a la variable aleatoria y_i son:

$$\mathrm{E}(Y|X)=\mu$$

,

$$Var(Y|X)=\mu\ +\mu^2\sigma^2$$

,

de esta forma la relación media varianza no se mantiene, tal que $Var(Y|X)=\mu\ +\mu^2\sigma^2 > \mathrm{E}(Y|X)=\mu$ y así se manifiesta el fenómeno de superdispersión.

- Exceso de ceros

El exceso de ceros es una de las fuentes de especificación errónea más frecuente, consiste en la presencia de un exceso de valores nulos respecto a la probabilidad predicha por la distribución de Poisson. No necesariamente implica ausencia del fenómeno bajo estudio.

Lambert (1992) introdujo un modelo de regresión Poisson de ceros modificados, en el cual:

$$P[Y=y_i]\begin{cases}\pi_i+(1-\pi_i)\exp(-\mu_i) & si \quad y_i=0\\ (1-\pi_i)\dfrac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} & si \quad y_i>0\end{cases}$$

,

se propone capturar la influencia de las variables explicativas en la probabilidad de ceros adicionales, los momentos característicos de este modelo son

$$\mathrm{E}(Y)=\mu\,, Var(Y)=\mu+\left(\frac{\pi}{1-\pi}\right)\cdot\mu^2$$

.

Este es el caso en el cual nos vamos a detener, dado el interés práctico de los objetivos propuesto. Comenzamos definiendo la función de probabilidad Poisson para casos en los que existe una gran cantidad de ceros que sesgan el promedio; modelo ZIP (*Zero Inflated Poisson*).

$$P[Y=y]\begin{cases} p+(1-p)\exp(-\mu) & si \quad y=0 \\ (1-p)\dfrac{\exp(-\mu)\mu^{y}}{y!} & si \quad y>0 \end{cases} \quad siendo \quad \mu=\lambda .$$

Esperanza matemática para el modelo ZIP

$$\begin{aligned} \mathrm{E}(Y) &= \sum_{y=0}^{\infty} yP[Y=y] = \sum_{y=1}^{\infty} yP[Y=y] \\ &= \sum_{y=1}^{\infty} y(1-p)e^{-\lambda}\frac{\lambda^{y}}{y!} \\ &= (1-p)e^{-\lambda}\lambda \sum_{(y-1)=0}^{\infty} \frac{\lambda^{(y-1)}}{(y-1)!} \\ &= (1-p)e^{-\lambda}\lambda e^{\lambda} \\ &= (1-p)\lambda = \mu \end{aligned} .$$

Hallando la varianza

$$\begin{aligned} \mathrm{E}(Y^{2}) &= \sum_{y=0}^{\infty} y^{2}P[Y=y] = \sum_{y=1}^{\infty} y^{2}(1-p)e^{-\lambda}\frac{\lambda^{y}}{y!} \\ &= (1-p)\sum_{(y-1)=0}^{\infty} ye^{-\lambda}\frac{\lambda^{y}}{(y-1)!} \\ si(y-1) &= m \Rightarrow (1-p)\sum_{m=0}^{\infty}(m+1)e^{-\lambda}\frac{\lambda^{m+1}}{m!} \\ &= (1-p)\left(\sum_{m=0}^{\infty} me^{-\lambda}\frac{\lambda^{m}}{m!} + \sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^{m}}{m!}\right) \\ &= (1-p)\left(\lambda\sum_{m=0}^{\infty} me^{-\lambda}\frac{\lambda^{m}}{m!} + \lambda\sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^{m}}{m!}\right) \\ &= (1-p)(\lambda^{2}+\lambda) \end{aligned} ,$$

de tal forma que

$$
\begin{aligned}
Var(Y) &= \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \\
&= (1-p)(\lambda^2 + \lambda) - ((1-p)\lambda)^2 = \\
&= (1-p)\lambda + \frac{p}{1-p}((1-p)\lambda)^2 = \\
&= \mu + \frac{p}{1-p}\mu^2
\end{aligned}
,
$$

nótese que la varianza de este modelo es mayor que la media de la distribución de Poisson. Cuanto mayor es la probabilidad de exceso de ceros, mayor es la varianza de la variable. A medida que *p* se aproxima a cero, la varianza se aproxima a μ y retornamos al modelo Poisson de referencia.

- Función de Varianza incorrecta

Como se ha mencionado previamente, la ausencia nominal de Poisson implica una violación del supuesto distribucional. La función de varianza del modelo de referencia es la que coincide con la esperanza matemática del mismo y la que denota equidispersión, siendo la identidad $Var(Y|X) = \mathrm{E}(Y|X) = \exp(x_i\beta)$, cuando no se da esta relación pueden ocurrir dos situaciones, superdispersión $Var(Y|X) > \mathrm{E}(Y|X)$ que es generalmente lo más habitual, o infradispersión $Var(Y|X) < \mathrm{E}(Y|X)$ que raramente se da. Ambas situaciones aparecen cuando en el mecanismo generador de datos subyacente, la función que relaciona la media condicional con la varianza condicional no es la función identidad. Puede ser cualquier función arbitraria que contempla variables explicativas adicionales u_i quedando expresado como:

$$
Var(Y|X,u) = f[\exp(x_i\beta), u_i] = \exp(x_i\beta)u_i
,
$$

esta modificación en la función de la varianza ha sido incorporada en diversos modelos de los cuales el más frecuentemente aplicado es el de Binomial Negativa,

que constituye un caso particular del hipermodelo denominado Negbin *k* (Cameron y Trivedi, 1998; Ridout, Hinde y Demetrio, 1998) este modelo puede ser expresado como $Var(Y|X) = \mu_i + \alpha\mu^{2-k}, \alpha \in \Re^+, k \in \Re$, adviértase que en el caso que k=0 se vuelve al modelo Binomial Negativo estándar $Var(Y|X) = \mu_i + \alpha\mu_i^2$, o bien al modelo propuesto por Lambert (1992) si $\alpha = \left(\frac{\pi}{1-\pi}\right)$ y en el caso que k=1 $Var(Y \mid X) = \mu_i + \alpha\mu_i$.

El reconocimiento del fenómeno de superdispersión en un modelo, se realiza a través del valor de Pearson (χ^2) y la *Deviance*[2] generalizada, en relación con los grados de libertad (Nelder y Wedderburn, 1972; Hilbe, 2008), siempre que este cociente exceda la unidad: $\frac{\chi^2}{gl} > 1$ o bien $\frac{D}{gl} > 1$, ya que en el caso de independencia y homogeneidad este valor es igual a la unidad (Wedderburn, 1974; Cox, 1989). El valor de Pearson (χ^2) y la *Deviance* son medidas de dispersión, utilizadas tanto para reconocer superdispersión de manera generalizada en el modelo, como para observar la bondad de ajuste del mismo, formalmente:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu})^2}{V(\mu_i)} = (Z - \hat{\eta})' \hat{W} \hat{\phi} (Z - \hat{\eta}),$$

estadístico de Pearson generalizado, (Cordeiro, 1986) y la *Deviance:*

$$\begin{aligned} D(y;\hat{\mu}) &= \sum 2w_i \left\{ y_i \left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right) \right\} \\ &= 2\left(\tilde{L}_{(n)} - \hat{L}_{(p)}\right), \end{aligned}$$

$\tilde{\theta} =$ estimación bajo el modelo saturado

$\hat{\theta} =$ estimación bajo el modelo corriente

[2]Conocida también como Discrepancia, tiene una distribución asintóticamente χ^2.

Deviance Escalada o Estandarizada:

$$D(y;\hat{\mu})/\phi = \sum 2w_i\left\{y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)/\phi\right\}$$
$$= 2\left(\tilde{L}_{(n)} - \hat{L}_{(p)}\right)$$

.

Se generan estadísticos de bondad de ajuste asociados a la distribución que se esté utilizando para el modelo propuesto. Particularmente para el modelo de Poisson, queda:

$$D(y;\hat{\mu}) = \sum 2\left\{y\log\left(y/\hat{\mu}\right) - \left(y - \hat{\mu}\right)\right\}$$

.

Según Hinde y Demétrio (2005) la omisión de superdispersión trae serias consecuencias, entre las que se pueden destacar los errores estándar obtenidos, estos podrían ser incorrectos y seriamente subestimados, con lo cual habría malos resultados sobre la significatividad individual de los parámetros de interés. A su vez, la interpretación del modelo sería desatinada y las predicciones poco precisas.

"...la modelación es un arte donde deben combinar el "background" teórico, la experiencia y el error propio del tipo de ensayo que se analiza". (Jorgensen, 1989)

MODELOS PARA DATOS DE CONTEO CON SUPERDISPERSIÓN

El supuesto más relevante en el modelo de Poisson es la equidispersión o en otros términos la igualdad entre la media y la varianza. Hemos introducido en capítulos previos, que cuando la varianza excede el promedio estamos en presencia de superdispersión.

Si bien la principal característica del modelo de Poisson es la capacidad de capturar la naturaleza discreta y no negativa de la variable de conteo, la limitación proviene de la rigidez impuesta por sus supuestos (Winkelmann, 2000). Surgen así distintas ampliaciones y modificaciones del modelo que permiten relajar estos supuestos, en presencia de superdispersión.

1. Modelo de Regresión Binomial Negativa (MRBN)

El modelo de RBN es conocido como el modelo paramétrico estándar para datos de conteo con presencia de superdispersión (Cameron y Trivedi 1998; Navarro y Utzet 2001). Este modelo se puede derivar de distintas maneras, la más habitual es la caracterizada por un conjunto de datos distribuidos a partir de la distribución de Poisson, en el cual la media está especificada de manera incompleta a causa de la existencia de heterogeneidad no observada. Se la considera así como una variable aleatoria con una distribución gamma a nivel poblacional (Mc.Cullagh y Nelder, 1989; Cameron y Trivedi, 1998; Veronesi, 2001).

En el modelo de Poisson, la media condicional de y es,

$$\mathrm{E}(Y|X) = \mu = \exp(x\ \beta) \qquad (1)$$

la variación en μ es introducida a través de la heterogeneidad observada, de forma que diferentes valores de x resultan en diferentes valores de μ; así diferentes observaciones x_i tienen la misma media μ_i. En contraste, en el MRBN la media μ es reemplazada por la variable aleatoria $\tilde{\mu}$ (Long, 1997) obteniéndose la siguiente función estocástica:

$$\tilde{\mu}_i = \exp(x_i\beta + \varepsilon_i), \qquad (2)$$

asumiendo que $Cov(\varepsilon, x) = 0$, implicando que el error aleatorio ε y la variable explicativa x no están correlacionados. La variación en $\tilde{\mu}$ se debe tanto a la variación en x_i entre los individuos, como a la heterogeneidad no observada introducida a través de ε. Así para una combinación de valores en las variables independientes, existe una distribución de diversas $\tilde{\mu}$ en lugar de una única μ.

De (1) y (2) se observa la relación entre $\tilde{\mu}$ y la μ de origen (McCullagh y Nelder, 1989):

$$\tilde{\mu}_i = \exp(x_i\beta + \varepsilon_i) = \exp(x_i\beta)\exp(\varepsilon_i),$$

$$= \mu_i \exp(\varepsilon_i) = \mu_i\delta_i,$$

al asumir la media del término de error igual a uno (Long, 1997) $E(\delta_i) = 1$, se obtiene la caracterización del modelo de regresión Poisson

$$\mathrm{E}(\tilde{\mu}_i) = \mathrm{E}(\mu_i\delta_i) = \mu_i\mathrm{E}(\delta_i) = \mu_i.$$

Por otro lado, la distribución de las observaciones dada x y δ es también Poisson:

$$\Pr(Y|X,\delta)=\frac{\exp(-\tilde{\mu}_i)\tilde{\mu}_i^{y_i}}{y_i!}=\frac{\exp(-\mu_i\delta_i)(\mu_i\delta_i)^{y_i}}{y_i!}\,.$$

Sin embargo, ya que δ es desconocido no se puede calcular $\Pr(y|x,\delta)$. Para calcular $\Pr(Y|X)$ sin tener en cuenta δ, se promedia $\Pr(y|x\delta)$ por la probabilidad de cada valor de δ. Si g es la función de densidad de probabilidad de δ, entonces la densidad marginal de y_i se puede obtener integrando con respecto a δ_i (Long, 1997; Cameron y Trivedi, 1986):

$$\Pr(Y|X)=\int_0^\infty[\Pr(y_i|x_i,\delta_i)]g(\delta_i)d\delta_i=\int_0^\infty\frac{e^{-\exp(x_i\beta+\delta_i)}\exp(x_i\beta+\delta_i)^{y_i}}{y_i!}g(\delta_i)d\delta_i\,,$$

esta expresión es la función de distribución de Poisson compuesta según Cameron y Trivedi (1986), su utilidad reside en la generalización natural de los modelos de Poisson básicos y su aplicación se debe generalmente a una necesidad de mayor flexibilidad, especialmente en situaciones de superdispersión.

La ecuación de la distribución de Poisson compuesta, calcula la probabilidad de y_i como una combinación de dos distribuciones de probabilidad (Long, 1997). A su vez, la forma de la ecuación depende de la selección de g, es decir de la función de densidad de probabilidad que se asuma para δ_i. Según Long (1997) lo más usual es asumir que δ_i sigue una distribución gamma con parámetro v_i de la siguiente forma:

$$g(\delta_i)=\frac{v_i^{v_i}}{\Gamma(v_i)}\delta_i^{v_i-1}\exp(-\delta_i v_i)\qquad \forall v_i>0\,,$$

definiendo a la función gamma como $\Gamma(v)=\int_0^\infty t^{v-1}e^{-1}dt$. A partir de estos supuestos, la integración de la ecuación de la regresión de Poisson compuesta conduce a una

distribución binomial negativa. Podemos así definir la distribución de probabilidad binomial negativa, según Long (1997) como:

$$\Pr(Y|X) = \frac{\Gamma(y_i + v_i)}{\Gamma(y+1)\Gamma(v_i)}\left(\frac{v_i}{v_i + \mu_i}\right)^{v_i}\left(\frac{\mu_i}{v_i + \mu_i}\right)^{y_i} \qquad \forall y_i \geq 0 \quad .$$

El valor esperado de Y para la distribución binomial negativa es el mismo que para la distribución de Poisson:

$$\mathrm{E}(Y|X) = \exp(x\beta) = \mu \quad ,$$

pero la varianza condicional sí difiere en relación a la de la distribución de Poisson:

$$Var(Y|X) = \mu\left(1 + \frac{\mu}{v}\right) = \exp(x\ \beta)\left(1 + \frac{\exp(x\ \beta)}{v}\right) \quad ,$$

dado que $\mu > 0$ y $v > 0$, la varianza condicional de Y en el MRBN será mayor que la media condicional $\mathrm{E}(Y|X) = \exp(x\ \beta) = \mu$. Pero al aumentar v la distribución tiende a la equidispersión ya que $Var(Y|X) \longrightarrow \mu$. Una varianza condicional elevada en Y incrementa la frecuencia de valores altos y bajos. De esta forma, en situaciones de superdispersión, la distribución binomial negativa corrige la probabilidad asociada a valores bajos de conteos que, habitualmente presentan un ajuste deficiente a través del modelo de regresión Poisson (Long, 1997; Veronesi, 2001).

Si v varía entre los individuos se genera un problema de indeterminación, dado que existen más parámetros que observaciones. Como solución más apropiada se asume que v es común para todos los individuos (Long, 1997) implicando varianza constante,

$$v\ = \alpha^{-1} \qquad \forall \alpha > 0 \quad ,$$

siendo α el parámetro de superdispersión, obsérvese que al incrementarse se incrementa la varianza condicional de Y .

Al reexpresar la función de densidad de la siguiente manera:

$$\Pr(Y|X) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i},$$

se reexpresa consecuentemente la varianza

$$Var(Y|X) = \mu\left(1+\frac{\mu}{\alpha^{-1}}\right) = \mu\left(1+\alpha\mu\right) = \mu + \alpha\mu^2,$$

quedando explícito que la varianza varía conjuntamente varía la media (Pinto y Ponce de Leon, 2006).

2. Modelos con Varianza Generalizada

Los modelos generalizados de datos de conteo tienen una característica en común, la falta de determinación con respecto al origen del error de especificación que produce la superdispersión. Una estrategia para evitar la restricción impuesta por el modelo de regresión Poisson es el uso de una función de varianza generalizada como otra forma para adquirir la varianza de la variable aleatoria que se está modelando, que sin perder precisión, sigue caracterizando la distribución (Hinde y Demetrio, 2005)

2.1. Regresión de Poisson generalizada

A partir de la distribución de Poisson generalizada (Lambert, 1992; Singh y Famoye, 1993) se constituye una alternativa al modelo de conteos de eventos.

Este admite tanto superdispersión como infradispersión y tiene la característica de anidar al modelo de regresión Poisson como un caso especial.

Se determina la distribución de Poisson generalizada, a partir de su función de densidad, y los momentos matemáticos de primer y segundo orden condicionados.

Función de densidad:

$$f(Y)=\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i}\frac{(1+\alpha\mu_i)^{y_i-1}}{y_i!}\exp\left(-\frac{\mu_i(1+\alpha\mu_i)}{1+\alpha\mu_i}\right).$$

Media condicional:

$$\mathrm{E}(Y|X)=\mu .$$

Varianza condicional:

$$Var(Y|X)=\mu\ (1+\alpha\mu\ )^2,$$

siendo α el parámetro de dispersión, tal que:

$\alpha<0$ señalainfradispersión.

$\alpha>0$ señalasuperdispersión.

$\alpha=0$ señalaequidispersión, en cuyo caso la función de densidad queda reducida al modelo de regresión de Poisson. (Singh y Famoye, 1993).

2.2. Regresión de Poisson Robusta

Cuando el mecanismo generador de los datos no es del tipo Poisson, una de las soluciones más usadas se basa en introducir un parámetro de dispersión que sea capaz de modelar la naturaleza de dicha variación, dado que la estructura

de media está relacionada con la estructura de la varianza (Winkelmann y Zimermmann, 1995).

Asumiendo la forma funcional de la varianza como $Var(Y|X) = \phi\mu$, se asegura una estimación semi-paramétrica o robusta, sin la necesidad de un conocimiento exhaustivo de la distribución generadora de las observaciones. Y la distribución será caracterizada a partir de los momentos de primer y segundo orden (Cordeiro*et. al*, 1993; Orbe, 2001).

Los estimadores semi-paramétricos, según Winkelmann (2000) utilizados de manera frecuente como el quasi máximo verosimil (*quasi-maximum likelihood* -QML-) tienden a ser en general inconsistentes e ineficientes. Sin embargo Gouriroux*et. al* (1984) indican que si la media está especificada correctamente y el modelo forma parte de la familia exponencial lineal, el estimador QML es consistente y lo denominan pseudo-máximo verosimil (PML).

Dada la pertenencia de la distribución de Poisson a la familia exponencial de distribuciones, las desviaciones de la función de varianza estándar no afectan a la consistencia de los parámetros estimados, siempre que la media esté especificada de manera correcta, de este modo se asume una media de Poisson mientras que se relaja la restricción de equidispersión. Los errores estándar de los parámetros del modelo de regresión Poisson deben ser adaptados en presencia de superdispersión, así la matriz de varianzas estimada bajo máxima verosimilitud resulta adecuada (Winkelmann y Zimmermann, 1995; Cameron y Trivedi, 1998).

Como estrategia, Winkelmann y Zimmermann (1995) proponen calcular de manera asintótica los errores estándar, partiendo del supuesto de consistencia de las estimaciones de los parámetros. Esto equivale a la estimación PML y lo denominan regresión de Poisson robusta.

Si la media está bien especificada, se cumple que $\mathrm{E}(Y|X) = \exp(x_i\beta) = \mu_i$ y entonces los estimadores pseudo-máximo-verosímiles (PML) adoptan la siguiente forma:

$$\hat{\beta} \overset{a}{\sim} \mathrm{N}\left[\beta, Var_{PML}\left(\hat{\beta}\right)\right],$$

siendo

$$Var_{PML}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{n} \mu_i x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} \omega_i x_i x_i'\right) \left(\sum_{i=1}^{n} \mu_i x_i x_i'\right)^{-1} \Big/ \; \omega = V\hat{a}r\left(Y \middle| X\right).$$

Respecto al término ω se pueden distinguir tres supuestos sobre la función de varianza (Winkelmann, 2000):

. Función de varianza en caso de haber ausencia de supuesto (Breslow, 1990)

$$\omega = Var\left(Y \middle| X\right) = \left(y_i - \hat{\mu}_i\right)^2.$$

. Función de varianza en caso que el supuesto sea la linealidad (Mc.Cullagh y Nelder, 1989)

$$\omega = Var\left(Y \middle| X\right) = \hat{\delta}^2 \hat{\mu}_i \Big/ \; \hat{\delta}^2 = \left[\frac{1}{n-k} \sum_{i=1}^{n} \frac{\left(y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i}\right].$$

. Función de varianza en caso que el supuesto sea un polinomio de grado dos (Gourieroux*et. al*, 1984)

$$\omega = Var\left(Y \middle| X\right) = \hat{\mu}_i + \hat{\delta}^2 \hat{\mu}_i^{\,2}.$$

3. Modelos de ceros Modificados

Los modelos de ceros modificados, son aquellos que presentan un desequilibrio respecto a la cantidad de ceros ya sea por exceso o por defecto (Winkelmann y Zimmermann, 1995). Se pondrá el énfasis en el exceso, ya que se presentará el problema de modelar el comportamiento promedio de variables

aleatorias que representan conteos de individuos, cuyas observaciones contienen gran cantidad de ceros (exceso de observaciones nulas).

Los modelos de ceros modificados, varían la estructura de media para modelar de manera explícita el exceso de ceros, asumiendo que los valores nulos pueden ser generados a partir de un proceso diferente al de los conteos estrictamente positivos. A diferencia del MRBN y del MRPG que solo incrementan la varianza sin modificar la media (Long, 1997; Cameron y Trivedi, 1998).

El sustento de los modelos con ceros aumentados[3] es que la probabilidad binomial rige el resultado binario de la presencia del valor cero, o bien de un valor positivo (Cameron y Trivedi, 1998).

Si c_i es una variable de selección binaria tal que permite la separación de valores de conteos cero y de valores de conteos estrictamente positivos de la siguiente manera:

$$y_i = \begin{cases} 0 & \text{si} \quad c_i = 1 \\ y* & \text{si} \quad c_i = 0 \end{cases}$$

entonces, si la probabilidad de ocurrencia de $c_i = 1$ está representada por π_i, la función de probabilidad de y_i es, según Winkelmann (2000):

$$f(Y) = \pi_i(1-d_i) + (1-\pi_i)g(y_i) \Big/ y_i = 1,2,...,$$

siendo $d_i = (1-c_i) = \min\{y_i,1\}$ y $g(y_i)$ un modelo de conteo habitual como el MRP o el MRBN.

En estos modelos de ceros aumentados se obtienen dos tipos de ceros, asociados a dos procesos generadores: por un lado -la mayoría- provienen de $c_i = 1$ y por otro lado, cuando $c_i = 0$. Esto es, el valor cero realizado por una razón

[3] El significado de la palabra “aumentado” proviene de la traducción de la palabra *inflated* que hace referencia a la inflación en el sentido de aumento desmedido.

estructural, o bien, el otro tipo de cero, porque verdaderamente ocurrió la ausencia. Cómo usualmente no es posible conocer con exactitud qué tipo de proceso generó el valor cero, la distinción entre los dos grupos es una forma de heterogeneidad inobservable discreta (Long, 1997; Winkelmann, 2000; Pron, 2007).

De este modo, la probabilidad total de ceros es una combinación de probabilidades de cada grupo, ponderada por la probabilidad de un individuo de estar en el grupo (Long, 1997; Cameron y Trivedi, 1998;Winkelmann, 2000; Pron, 2007).

3.1. MRP de ceros Aumentados (ZIP, *Zero Inflated Poisson*)

En 1992, Lambert introdujo formalmente el modelo de Poisson de ceros aumentados con la intención de capturar la influencia de las variables explicativas en la probabilidad de ceros adicionales, entonces retomando lo anteriormente expuesto sobre modelos de ceros adicionales:

$$\pi_i = F(z_i, \gamma) = \frac{\exp(z_i\gamma)}{1+\exp(z_i\gamma)},$$

la probabilidad de los valores generados a partir del segundo proceso, estará dado por la distribución de referencia, en este caso la distribución Poisson:

$$f(y_i, \mu_i) = \Pr(Y = y_i|\mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}.$$

Combinando ambos procesos, las probabilidades de los valores de conteo vienen dadas por: (Lambert, 1992; Long, 1997)

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

siendo la esperanza:

$$\begin{aligned}\mathrm{E}(Y) &= 0\cdot\left[\pi+(1-\pi)e^{-\mu}\right]+\sum_{y=1}^{n} y(1-\pi)\frac{e^{-\mu}\mu^{y}}{y!}\\ &= (1-\pi)\sum_{y=1}^{n} y\frac{e^{-\mu}\mu^{y}}{y!} = (1-\pi)\sum_{y=0}^{n} y\frac{e^{-\mu}\mu^{y}}{y!}\end{aligned},$$

si se considera que $Y \approx \mathrm{Poi}(\mu)$

$$\mathrm{E}(Y)=\sum_{y=0}^{n} y\frac{e^{-\mu}\mu^{y}}{y!}=\mu,$$

y se obtiene así la esperanza del modelo ZIP

$$\mathrm{E}(Y)=(1-\pi)\mu.$$

Operando algebraicamente se obtiene la varianza:

$$\begin{aligned}Var(Y) &= \mathrm{E}(Y^2)-[\mathrm{E}(Y)]^2\\ &= (1-\pi)(\mu+\mu^2)-(1-\pi)^2\mu^2\\ &= \mu+\mu^2-\pi\mu-\pi\mu^2-(1-2\pi+\pi^2)\mu^2\\ &= \mu+\mu^2-\pi\mu-\pi\mu^2-\mu^2+2\pi\mu^2-\pi^2\mu^2\\ &= \mu-\pi\mu+\pi\mu^2-\pi^2\mu^2\\ &= \mu(1-\pi+\pi\mu-\pi^2\mu)\\ &= \mu[1(1-\pi)+\pi\mu(1-\pi)]\\ &= \mu(1-\pi)\cdot(1+\pi\mu)\end{aligned},$$

$$Var(Y)=\mu(1-\pi)\cdot(1+\pi\mu).$$

Si $\pi=0$ se tiene el modelo de regresión estándar, de otro modo, la varianza excede el promedio. Esto es debido a que en esta distribución, tanto el valor

esperado como la varianza, contemplan no solamente el promedio, sino también la proporción de ceros.

Para el ajuste del modelo ZIP se debe utilizar dos funciones de enlace (Ridout*et al.*, 1998), una para la parte con exceso de ceros (a) y la otra para la parte que no contempla el exceso de ceros (b).

a) Función de enlace entre la media y el predictor lineal para la parte con exceso de ceros, dada por la función logit:

$$\text{logit}(\boldsymbol{\pi}) = \ln\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \mathbf{G}\gamma \quad ,$$

siendo $\mathbf{G}$ una matriz de variables explicativas y γ el vector de los parámetros a ser estimados de variables asociadas a la parte con exceso de ceros del modelo.

b) Función de enlace para la parte sin exceso de ceros, es la misma que la utilizada por el modelo de Poisson habitual:

$$\ln(\boldsymbol{\mu}) = \mathbf{B}\beta \quad ,$$

siendo $\mathbf{B}$ la matriz de variables asociadas y β el vector de parámetros a ser estimado correspondiente a la parte sin exceso de ceros del modelo.

3.2. MRBN de ceros Aumentados (ZINB)

Una variable aleatoria Y tiene distribución Binomial Negativa con exceso de ceros (ZINB), con probabilidades π_i para los ceros estructurales y $(1-\pi_i)$ para los ceros muestrales (Long, 1997) si:

$$P[Y = y_i]\begin{cases} \pi_i + (1-\pi_i)\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} & si \quad y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\dfrac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} & si \quad y_i > 0 \end{cases} .$$

Análogamente al modelo ZIP, el valor esperado y la varianza toman la siguiente forma, considerando la estructura de ceros incluida en el modelo:

$$\mathrm{E}(Y|X) = \mu(1-\pi) = \mu - \mu\pi .$$

La media condicional del modelo es modificada mediante la reducción del número esperado en $\mu\pi$ y la varianza a su vez, se modifica de la siguiente forma:

$$Var(Y|X) = \mu(1-\pi) + (1+\mu(\pi+\alpha)) .$$

Si $\pi = 0$, se vuelve al modelo de RBN, pero si $\pi > 0$ la dispersión es mayor que el modelo estándar (Long, 1997; Cameron y Trivedi, 1998;Winkelmann, 2000; Pron, 2007).

"Una teoría es una buena teoría siempre que satisfaga dos requisitos: debe describir con precisión un amplio conjunto de observaciones, sobre la base de un modelo que contenga sólo unos pocos parámetros arbitrarios y debe ser capaz de predecir positivamente los resultados de observaciones futuras". (Hawking,1988)

MÉTODOS Y ALGORITMOS DE ESTIMACIÓN

En este capítulo, se consideran varias posibilidades para estimar modelos con superdispersión. Entre ellos, se distingue el método de Máxima Verosimilitud (ML)[4] (Fisher, 1922), el cual se utiliza si hay una especificación completa sobre la distribución probabilística del mismo. El método de máxima verosimilitud es un método clásico de estimación de parámetros, asociado a funciones de densidad o probabilidad de variables aleatorias. Requiere como característica, que los parámetros se encuentren en el espacio paramétrico o en el rango de variación natural de los mismos, y contiene así propiedades óptimas de los buenos estimadores, tales como consistencia y eficiencia asintótica (Nelder y Wedderburn, 1972; Montero Mercadé, 2004; Hinde y Demetrio, 2005; Cordeiro y Demetrio, 2007).

Cuando se tiene una especificación no completa sobre la forma de la función de varianza, entonces se necesita un método de estimación que quede representado a partir de los dos primeros momentos, como el de cuasi verosimilitud (MQL) entre otros (Hinde y Demetrio, 2005), los cuales se profundizarán a medida que se avance en la exposición del capítulo.

[4] ML por el vocabulario inglés *Maximun Llikelihood.*

1. Máxima Verosimilitud (ML)

Suponiendo un conjunto de observaciones $y_1,...y_n$, siendo estas una realización particular de un vector aleatorio $\mathbf{Y^T}=(\mathbf{y_1,...,y_n})$, todas ellas con la misma distribución de probabilidad, mutuamente independientes y pertenecientes a la familia exponencial y cada una dependiente de un parámetro único $\boldsymbol{\theta}^{\mathbf{T}}=(\boldsymbol{\theta}_{\mathbf{1}},...,\boldsymbol{\theta}_{\mathbf{n}})$ y parámetro de escala común y conocido ϕ. Se utiliza el método ML para la estimación de los parámetros lineales $\beta_i \diagup i=1,2,..,p$ del modelo (Montero Mercadé, 2004).

La función de log-verosimilitud del conjunto de observaciones tiene por expresión:

$$L(y;\theta,\phi)=\mathbf{log}\prod_{i=1,..,n} f_y(y_i;\theta_i,\phi)=\sum_{i=1,..,n}\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+\sum_{i=1,..,n}c(y_i,\phi)$$

,

$$l(\beta)=\frac{1}{\phi}\sum_{i=1}^{n}y_i\theta_i-b(\theta_i)+\sum_{i=1}^{n}c(y_i,\phi)$$

.

Siendo $\theta_i=q(\mu_i)$, $\mu_i=g^{-1}(\eta_i)$ y $\eta_i=\sum_{r=1}^{p}x_{ir}\beta_r$.

De la función expuesta, se puede calcular por la regla de la cadena el denominado vector *score total,* que se denota como $\mathbf{U}_r$, a la derivada parcial de la función de log-verosimilitud respecto al parámetro θ_i, de dimensión p (Nelder y Wedderburn, 1972; Montero-Mercadé, 2004, Hinde y Demetrio, 2005).

$$\mathbf{U}(\beta)=\frac{\partial l(\beta)}{\partial\beta}$$

,

con elemento típico $U_r(\beta)=\frac{\partial l(\beta)}{\partial \beta_r}=\sum_{i=1}^{n}\frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r}$, ya que $l(\beta)=f(\theta_1,\theta_2,...,\theta_i,...,\theta_n)$,

$$\longrightarrow \theta_i=\int V_i^{-1}d\mu_i=q(\mu_i),$$

$$\longrightarrow \mu_i=g^{-1}(\eta_i)=h(\eta_i),$$

$$\longrightarrow \eta_i=\sum_{r=1}^{p}x_{ir}\beta_r,$$

y recordando que ser parte de la familia exponencial tiene la siguiente implicancia $\mu_i=b'(\theta_i)$ y $\frac{\partial \mu_i}{\partial \theta_i}=V_i$ y por lo tanto

$$U_r=\frac{1}{\phi}\sum_{i=1}^{n}(y_i-\mu_i)\frac{1}{V_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ir} \;/\; r=1,2,...,p.$$

La estimación $\hat{\beta}_i$, de los parámetros β_i por medio del método de máxima verosimilitud se obtiene igualando $\mathbf{U}_r$ a cero. Generalmente $\mathbf{U}_r=0 \;/\; r=1,2,...,p$ son funciones no lineales y tienen que ser resueltas numéricamente por procesos iterativos del tipo Newton-Raphson (Hinde y Demetrio, 2005; Cordeiro y Demetrio, 2007; Camargo Mendes, 2007).

El método iterativo Newton-Raphson para la solución de una ecuación $f(x)=0$ basado en la aproximación de Taylor para una función $f(x)$ en el entorno del punto x_0 es:

$$f(x)=f(x_0)+(x-x_0)f'(x_0)=0,$$

obteniéndose $x = x_0 - \frac{f(x_0)}{f'(x_0)}$,

de una forma más general:

$$x^{m+1} = x^m - \frac{f(x^m)}{f'(x^m)},$$

(Larson *et. al*, 1995; Cordeiro y Demetrio, 2007), siendo x^{m+1} el valor del punto x, valorado en la instancia $(m+1)$, x^m es el valor del punto x en la instancia m, $f(x^m)$ la función valorada en x^m y $f'(x^m)$ la derivada primera de la función evaluada en el punto x^m.

La solución del sistema de ecuaciones $\mathbf{U} = \mathbf{U}(\beta) = \frac{\partial l(\beta)}{\partial \beta} = 0$, permite obtener los vectores máximo verosímiles de los parámetros estimados y la matriz de información, constituida por la inversa de la matriz negativa (J) de las derivadas parciales de segundo orden de la función de log-verosimilitud, con elementos $-\frac{\partial^2 l(\beta)}{\partial \beta_r \partial \beta_s}$, evaluada en el punto m, así

$$\beta^{(m+1)} = \beta^{(m)} + \left(J^{(m)}\right)^{-1} U^{(m)},$$

siendo $\beta^{(m)}$ y $\beta^{(m+1)}$ los vectores de parámetros estimados en los puntos m y $(m+1)$ respectivamente, $U^{(m)}$ el vector *score* evaluado en m. Cuando las derivadas parciales de segundo orden son evaluadas de manera directa, el método Newton-Raphson es óptimo. En caso contrario se utiliza el método *score* de Fisher, que es más simple en su cálculo, y coincide con el método de Newton-Raphson en el caso de funciones con enlaces canónicos (Cordeiro y McCullagh, 1991; Cordeiro y Demetrio, 2007). Este método sustituye la matriz de derivadas parciales de segundo orden por la matriz de valores esperados de las derivadas

parciales, así se sustituye la matriz de información observada J por la inversa de la matriz de información esperada de Fisher K

$$\beta^{(m+1)} = \beta^{(m)} + \left(K^{(m)}\right)^{-1} U^{(m)},$$

teniendo como elementos de K, a diferencia de los elementos de J:

$$k_{r,s} = -\mathrm{E}\left[\frac{\partial^2 l(\beta)}{\partial\beta_r \partial\beta_s}\right] = \mathrm{E}\left[\frac{\partial l(\beta)}{\partial\beta_r}\frac{\partial l(\beta)}{\partial\beta_s}\right],$$

que es la matriz de covarianzas de $\mathbf{U}'_r s$.

Si a la ecuación $\beta^{(m+1)} = \beta^{(m)} + \left(K^{(m)}\right)^{-1} U^{(m)}$ se la premultiplica por $K^{(m)}$ se obtiene $K^{(m)}\beta^{(m+1)} = K^{(m)}\beta^{(m)} + U^{(m)}$ y consecuentemente como elemento típico

$$k_{r,s} = \mathrm{E}(U_r U_s) = \frac{1}{\phi^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2 \frac{1}{V^2{}_i}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_{ir}x_{is},$$

trabajando algebraicamente la ecuación y renombrando como $w_i = \frac{1}{V^2{}_i}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$ a la ponderación o peso se reduce el elemento típico a $k_{r,s} = \phi^{-1}\sum_{i=1}^{n} w_i x_{ir} x_{is}$. La matriz de información de Fisher para β adquiere la forma $\mathbf{K} = \phi^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}$, siendo $\mathbf{W} = diag\{w_1,...,w_n\}$ una matriz diagonal de ponderaciones que contiene información sobre la distribución y la función de enlace utilizada, podrá también incluir un término de pesos a priori (Cordeiro y Demetrio, 2007). En el caso de funciones de enlace canónicas se tiene que $w_i = V_i$, ya que $V_i = V(\mu_i) = \frac{\partial\mu_i}{\partial\eta_i}$. Obsérvese que la información es inversamente proporcional al parámetro de dispersión (Ruiz Maya Pérez, 1999; Montero Mercadé, 2004; Cordeiro y Demetrio, 2007).

Reescribiendo el vector de *score* $\mathbf{U} = \mathbf{U}(\beta)$ como $\mathbf{U} = \phi^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{G}(\mathbf{y} - \boldsymbol{\mu})$, con $\mathbf{G} = diag\left\{\frac{\partial \eta_1}{\partial \mu_1},...,\frac{\partial \eta_n}{\partial \mu_n}\right\} = diag\{g'(\mu_1),...,g'(\mu_n)\}$. De este modo la matriz diagonal $\mathbf{G}$, queda formada a partir de las derivadas de primer orden de la función de enlace.

Al sustituir las matrices $\mathbf{K}$ y $\mathbf{U}$ en $K^{(m)}\beta^{(m+1)} = K^{(m)}\beta^{(m)} + U^{(m)}$ se obtiene

$$\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{X}\beta^{(m+1)} = \mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{X}\beta^{(m)} + \mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{G}^{(m)}\left(\mathbf{y} - \boldsymbol{\mu}^{(m)}\right),$$

o de manera equivalente

$$\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{X}\beta^{(m+1)} = \mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\left[\boldsymbol{\eta}^{(m)} + \mathbf{G}^{(m)}\left(\mathbf{y} - \boldsymbol{\mu}^{(m)}\right)\right],$$

si se define como variable dependiente ajustada, siguiendo a Hinde y Demetrio (2007) a $\mathbf{z} = \boldsymbol{\eta} + \mathbf{G}(\mathbf{y} - \boldsymbol{\mu})$, entonces podemos reescribir el sistema como:

$$\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{X}\beta^{(m+1)} = \mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{z}^{(m)},$$

$$\beta^{(m+1)} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{z}^{(m)}.$$

Así esta ecuación es válida para cualquier modelo lineal generalizado y muestra que la solución de las ecuaciones ML equivale a calcular de manera iterativa una regresión lineal ponderada de una variable dependiente ajustada $\mathbf{z}$ sobre la matriz $\mathbf{X}$ utilizando una función de ponderación $\mathbf{W}$ que se modifica en el proceso iterativo; así las funciones de varianzas y de enlace entran en el proceso a través de la matriz de ponderación $\mathbf{W}$ y de la variable ajustada $\mathbf{z}$. Es importante destacar que la ecuación iterativa no depende del parámetro de dispersión ϕ (Nelder y Wedderburn, 1972; Cordeiro y Demetrio, 2007).

Un método usual para iniciar el proceso iterativo es la especificación de la estimación inicial y cambiarla de manera sucesiva hasta que la convergencia sea obtenida y en consecuencia $\beta^{(m+1)}$ se aproxime a $\hat{\beta}$ siempre que m aumente.

Podría considerarse a cada observación como una estimación de su valor medio, implicando $\mu^{(1)}{}_i = y_i$ y calcular:

$$\eta^{(1)}{}_i = g\left(\mu^{(1)}{}_i\right) = g(y_i) \text{ y } w^{(1)}{}_i = \frac{1}{V(y_i)\left[g'(y_i)\right]^2}.$$

Al usar $\eta^{(1)}{}_i$ como variable respuesta, $\mathbf{X}$ como la matriz del modelo y $\mathbf{W}^{(1)}$ como la matriz diagonal de ponderaciones, con elementos $w^{(1)}{}_i$, se obtiene el vector buscado $\beta^{(2)} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(1)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(1)}\eta^{(1)}$, continuando con el proceso, para $m = 2,...,k$ se requiere $k-1$ iteraciones necesarias para la convergencia.

Siguiendo a Cordeiro y Demetrio (2007), y generalizando lo anteriormente expuesto se puede resumir el proceso en tres pasos:

a. se obtienen las siguientes estimaciones:

$$\eta_i^{(m)} = \sum_{r=1}^{p} x_{ir}\beta_r^{(m)} \quad \text{y } \mu_i^{(m)} = g^{-1}\left(\eta_i^{(m)}\right).$$

b. se obtiene la variable dependiente ajustada y las ponderaciones:

$$z_i^{(m)} = \eta_i^{(m)} + \left(y_i - \mu_i^{(m)}\right)g'\left(\mu_i^{(m)}\right),$$

$$w_i^{(m)} = \frac{1}{V\left(\mu_i^{(m)}\right)\left[g'\left(\mu_i^{(m)}\right)\right]^2}.$$

c. se calcula:

$$\beta^{(m+1)} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{z}^{(m)}.$$

Se repite el proceso hasta obtener la convergencia.

Entre los criterios existentes para verificar la convergencia, uno de ellos podría ser que la suma del cuadrado de la diferencia entre los estimados sea menor a un valor infinitesimal dado ξ, en términos matemáticos:

$$\sum_{r=1}^{p}\left(\frac{\beta_r^{(m+1)}-\beta_r^{(m)}}{\beta_r^{(m)}}\right)<\xi$$.

Cuando la función $g(\cdot)$ no está definida para algunos valores de y_i el proceso no puede ser iniciado, por ejemplo si la función de enlace viene dada por $\eta=g(\mu)=\ln(\mu)$ y son observados valores nulos.

La desventaja del método Newton-Raphson es que al utilizar la matriz de derivadas parciales de segundo orden, no siempre converge para determinados valores iniciales (Cordeiro y Demetrio, 2007; Camargo Mendes, 2007; Hilbe, 2008).

A modo de ejemplo, se presenta la estimación para la Binomial negativa. En términos generales, si la variable aleatoria y_i representa conteos, con media θ_i y se especifica el modelo con distribución Poisson, tal que $y_i \sim Poi(\theta_i)$, siendo las θ_i variables aleatorias con distribución gamma: $\Gamma(k,\lambda_i)$, entonces se deriva la distribución Binomial negativa para y_i con:

$$f(y_i;\mu_i,k)=\frac{\Gamma(k+y_i)}{\Gamma(k)y_i!}\frac{{\mu_i}^{y_i}k^k}{(\mu_i+k)^{k+y_i}} \Big/ y_i=0,1,...$$,

y la esperanza del modelo $\mathrm{E}(y_i)=\frac{k}{\lambda_i}=\mu_i$, con varianza de la siguiente forma:

$$\begin{aligned} Var(Y) &= \mathrm{E}_{\theta_i}\left[Var(y_i|\theta_i)\right]+Var_{\theta_i}\left(\mathrm{E}[y_i|\theta_i]\right) \\ &= \mathrm{E}(\theta_i)+Var(\theta_i)=\frac{k}{\lambda_i}+\frac{k}{{\lambda_i}^2} \\ &= \mu_i+\frac{{\mu_i}^2}{k} \end{aligned}$$.

La función de verosimilitud estándar del modelo de regresión Binomial negativo es:

$$l(\beta,Y,X)=\prod_{i=1}^{n}\Pr(y_i|x_i)$$

$$=\prod_{i=1}^{n}\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma(y_i+1)\Gamma\left(\alpha^{-1}\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}\Big/\ y=0,1,2,...$$

,

y $\alpha\geq 0$, $\mathrm{E}(Y|X)=\exp(x_i\beta)=\mu_i$.

(Cameron y Trivedi, 1998; Vives Brosa, 2002; Hilbe, 2008).

Aplicando logaritmos, se obtiene la función de log-verosimilitud, que permitirá estimar el modelo:

$$\ln l(\beta,Y,X)=\sum_{i=1}^{n}\left\{\sum_{j=0}^{y_i-1}\left(\ln\left(j+\alpha^{-1}\mu_i\right)\right)-\ln y_i!-\left(y_i+\alpha^{-1}\mu_i\right)\ln(1+\alpha)+y_i\ln\alpha\right\},$$

$$\ln l(\mu,k;y)=\sum_{i=1}^{n}\{y_i\ln\mu_i-k\ln k-(k+y_i)\ln(k+\mu_i)+d\ln(y_i,k)-\ln y_i!\}$$

.

Operando con la distribución propuesta, se obtienen las siguientes ecuaciones máximo verosímiles de *score* estimadas:

$$\frac{\partial l}{\partial\beta_r}=\sum_{i=1}^{n}\left\{\frac{y_i}{\mu_i}-\frac{k+y_i}{k+\mu_i}\right\}\frac{\partial\mu_i}{\partial\beta_r}$$

$$=\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\mu_i\left(1+\frac{\mu_i}{k}\right)}\frac{1}{g'(\mu_i)}x_{ir}$$

,

$$\frac{\partial l}{\partial k}=\sum_{i=1}^{n}\left\{d\ln(y_i,k)-\ln(\mu_i+k)-\frac{k+y_i}{k+\mu_i}+\ln k+1\right\}.$$

Para valores fijos de k, la estimación de β se realiza usando el ajuste usual con función de varianza $V(\mu)=\mu+\frac{\mu^2}{k}$.

Para β fijo, y μ dado, usando el método iterativo de Newton-Raphson

$$k^{(m+1)} = k^{(m)} - \left(\frac{\partial l}{\partial k} \Big/ \frac{\partial^2 l}{\partial k^2} \right)\Bigg|_{k^{(m)}},$$

hasta la convergencia. La derivada segunda con respecto a k está dada por:

$$\frac{\partial^2 l}{\partial k^2} = \sum_{i=1}^{n} \left\{ d\ln(y_i, k) - 2\left(\frac{1}{\mu_i + k} \right) + \frac{k + y_i}{(k + \mu_i)^2} + \frac{1}{k} \right\},$$

$$\frac{\partial^2 l}{\partial \beta_r \partial k} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{(k + \mu_i)} \frac{1}{g'(\mu_i)} x_{ir},$$

y como la $\mathrm{E}\left(\frac{\partial^2 l}{\partial \beta_r \partial k} \right) = 0$ implica que k y β no están correlacionadas, al menos asintoticamente (Hinde y Demetrio, 2005; Hilbe, 2008).

El Diagnóstico y la Selección de Modelos

Los métodos de diagnóstico son generales en el sentido que detectan ausencia de equidispersión y no únicamente presencia de superdispersión (Vives Brosa, 2002).

Entre las pruebas para detectar superdispersión, se pueden destacar tres grandes grupos de pruebas específicas (Vives Brosa, 2002; Llorens, 2005):

1. Pruebas para modelos anidados; estas están basadas en la comparación de la varianza del modelo de Poisson con una función de varianza generalizada, quedando anidada a la primera.
2. Pruebas para modelos no anidados; evaluar alguna restricción particular en base a la metodología de máxima verosimilitud, comparando el modelo restringido respecto al ampliado, estos pueden estar anidados o no.

3. Pruebas basadas en la regresión; similar a un análisis de residuos en modelo lineal general con errores distribuidos normalmente pueden revelar heterosedasticidad, los residuales de Poisson pueden indicar una violación del supuesto de equidispersión.

Este último caso, las pruebas basadas en la regresión, no se van a realizar en este trabajo, por lo cual se las menciona pero no se las presenta.

1. Pruebas para modelos anidados

Entre las pruebas para modelos anidados se pueden destacar: la prueba Razón de Verosimilitud (LR), la Prueba de Wald y la Prueba de Multiplicadores de Lagrange (LM). Estas pruebas son asintoticamente equivalentes bajo hipótesis nula verdadera. A medida que el tamaño de n aumenta, la distribución muestral de las tres pruebas converge a la misma distribución χ^2, con grados de libertad igual al número de restricciones evaluadas, sin embargo no se puede afirmar lo mismo para muestras de tamaños chicos (Long, 1997).

1.1. Razón de Verosimilitud (LR)

Esta prueba se basa en diferencia entre el valor de la función log-verosímil de la estimación restringida $\hat{\ell}_r$ -como puede ser el modelo Poisson, por ejemplo- y el valor de la función log-verosímil evaluada en las estimaciones de la máxima verosimilitud no restringido $\hat{\ell}_{nr}$ -como puede ser el modelo Binomial Negativa-. Si *k* es la cantidad de restricciones –en el caso del planteo entre MRP y MRBN, *k*=1- entonces bajo H_0 correcta:

$$LR = -2\left(\hat{\ell}_r - \hat{\ell}_{nr}\right) \approx \chi^2_{(k)},$$

siendo (*k*) los grados de libertad de la distribución χ^2. (Long, 1997)

1.2. Prueba de Wald

A diferencia de la prueba de LR, es suficiente la estimación de un solo modelo, siendo el punto de partida la distribución asintótica del estimador máximo verosímil del modelo no restringido. Se trabaja con un sistema de ecuaciones lineales, y en su forma más genérica puede ser utilizada para evaluar restricciones no lineales, para una ampliación al respecto ver Long, 1997.

$$R\theta = q \qquad \text{siendo} \begin{cases} \text{R la matriz de restricciones} \\ \text{q el vector de resultados} \\ \hat{\theta} \text{ coeficientes de regresión estimados} \end{cases}.$$

Al especificar R y q, quedan definidas una gran variedad de restricciones lineales. En un modelo con dos variables explicativas, por ejemplo, se podría establecer la restricción $\beta_1 = \beta_2 = 0$ y el sistema de ecuaciones lineales quedaría:

$$R\theta = q \Rightarrow \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

esta hipótesis, puede ser evaluada mediante el estadístico de Wald:

$$W = \left[R\hat{\theta} - q\right]' \left[R\hat{Var}\left(\hat{\theta}\right)R'\right]^{-1} \left[R\hat{\theta} - q\right],$$

$$W \approx \chi_r^2,$$

el estadístico de Wald sigue una distribución χ^2 con grados de libertad igual al número de restricciones. Las componentes de este estadístico son:

• $\left[R\hat{\theta} - q\right]$al principio y final de la estructura, mide la distancia entre el valor estimado y el hipotético.

• $\left[R\hat{V}ar\left(\hat{\theta}\right)R'\right]^{-1}$ refleja la variabilidad en el estimador, alternativamente la curvatura de la función de verosimilitud.

1.3. Prueba Multiplicadores de Lagrange (LM)

Esta prueba es también conocida como la prueba de puntuaciones o *score test*, estima solamente el modelo restringido y evalúa la pendiente de la función log-verosímil en la restricción, (Long, 1997; Winkelmann, 2000) si la hipótesis nula es verdadera, entonces la pendiente evaluada a través de $\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_r}$ debe aproximarse a 0. Por lo cual, la base de la prueba será:

$$LM = \left(\frac{\partial \ell}{\partial \theta}\right)'_{\theta_r} \left[\mathrm{I}\left(\hat{\theta}_r\right)\right]^{-1} \left(\frac{\partial \ell}{\partial \theta}\right)_{\theta_r},$$

este estadístico sigue una distribución χ^2 con grados de libertad igual al número de restricciones impuestas en la hipótesis nula.

Si como estimador consistente de la varianza, se obtiene a $\sum_{i=1}^{n} \frac{1}{2\hat{\mu}_i^2}$, se puede reexpresar al estadístico LM como:

$$LM = \sqrt{\left[\sum_{i=1}^{n} \frac{1}{2\hat{\mu}_i^2}\right]} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \hat{\mu}_i\right)^2 - y_i .$$

Bajo la hipótesis nula, la prueba LM sigue una distribución asintoticamente normal, dado que es la raíz cuadrada de una distribución χ^2 con un grado de libertad. La prueba de superdispersión consta en una prueba unilateral con valor crítico en z_α.

2. Pruebas para modelos no anidados

2.1. Prueba de Vuong

Una extensión de la prueba de razón de verosimilitud, para modelos no anidados es la prueba de Vuong (Vuong, 1989). El objetivo de la misma consiste en la selección del modelo más cercano a la distribución condicional verdadera. Si consideramos que dos modelos condicionales $\begin{cases} F_\alpha = \{f(y \mid x;\alpha), \alpha \in A\} \\ G_\beta = \{g(y \mid x;\beta), \beta \in A\} \end{cases}$, no son modelos anidados tenemos $F_\alpha \cap G_\beta = \emptyset$. La hipótesis que se plantea es la equivalencia entre los modelos:

$$H_0 = E_0\left[l_f(\hat{\alpha}) - l_g(\hat{\beta})\right] = 0.$$

Bajo la hipótesis nula, el estadístico de prueba LR_{NA} converge a una distribución normal.

$$LR_{NA} = \frac{\frac{1}{\sqrt{n}}\left[\left(l_f(\hat{\alpha})\right) - \left(l_g(\hat{\beta})\right)\right]}{\omega},$$

siendo

$$\omega = \frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i \mid x_i, \hat{\alpha}) - l_g(y_i \mid x_i, \hat{\beta})\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i \mid x_i, \hat{\alpha}) - l_g(y_i \mid x_i, \hat{\beta})\right]\right]^2.$$

La selección se hará en base a un valor crítico c para un nivel de significación dado. Si el estadístico $LR_{NA} > c$, entonces el rechazo de hipótesis de equivalencia de modelos implicará que se escoja el modelo f respecto al g; si por el contrario el estadístico $LR_{NA} < c$, el rechazo de hipótesis nula, implicará en este caso que se escoja el modelo g respecto al f. Por último si $|LR_{NA}| < c$, no se rechaza la hipótesis de nulidad y por consiguiente no se puede discriminar entre ambos modelos (Winkelmann, 2000).

En este caso particular, el test de Vuong será utilizado para comparar los modelos con exceso de ceros respecto a los modelos corrientes, como el modelo ZIP (f) y el Poisson (g), o bien entre el ZINB (f) y el Binomial negativo (g).

La *Deviance*

La función de *deviance* se utiliza para medir la calidad de un modelo lineal generalizado. El ajuste de un modelo a un conjunto de datos, consiste en sustituir los valores observados de y por valores estimados con un modelo que contenga la menor cantidad de parámetros posibles (McCullagh y Nelder, 1989). La intención es tener una pequeña discrepancia entre lo observado y lo esperado y tener cierta confianza que lo que se modeló representa satisfactoriamente lo que se observó. Se puede tener un conjunto grande de variables explicativas que mejoren el modelo, pero puede complicar la interpretación del mismo. Así mismo se puede tener un pequeño número de variables que expliquen y facilite la interpretación pero no justifiquen de manera coherente el comportamiento de lo observado. Por esto se necesita encontrar un modelo satisfactorio en términos de explicación, con la menor cantidad de parámetros posibles.

Para un conjunto de n observaciones puede ajustarse un modelo con n parámetros, denominado *saturado* o *completo* porque tiene un parámetro para cada observación, atribuyendo toda la variación al componente sistemático del modelo. Por otro lado, un modelo absolutamente simple de ser ajustado es aquel que contiene un único parámetro para el total de las observaciones y es conocido como modelo *nulo*, el cual atribuye toda la variabilidad al componente aleatorio del modelo.

Estos casos representan los extremos, generalmente se busca un modelo intermedio, el cual posea el menor número de parámetros posibles y que explique

mejor la respuesta observada. Entonces podemos decir que el modelo propuesto es el que tenga p parámetros linealmente independientes, denominado modelo *común* o *corriente*.

La *deviance* es la medida propuesta por Nelder y Wedderburn (1972) como medida de discrepancia entre los máximos del los logaritmos de la función de verosimilitud entre los modelos saturado y corriente, respectivamente $\hat{l}_n$ y $\hat{l}_p$. Su forma matemática está dada por:

$$D = 2\left(\hat{l}_n - \hat{l}_p\right),$$

Esta medida que depende del número de variables explicativas del modelo, como se explicó anteriormente, a mayor número de variables explicativas, menor es el valor que adquiere la *deviance* y más complejidad en la interpretación del modelo.

El valor de la deviance puede proporcionar una medida de la calidad del ajuste de un modelo lineal generalizado, siempre que se lo compare con un valor de distribución de probabilidad conocida. Si bien la distribución de la *deviance* es desconocida, en el caso de la distribución de Poisson, puede demostrarse que asume una distribución asintoticamente Ji cuadrado con $(n-p)$ grados de libertad (Nelder y Wedderburn, 1972).

Criterios de Información de Akaike y de Bayes - AIC y BIC -

Los criterios de AIC y BIC son ampliamente utilizados en la comparación de modelos, tanto anidados como no anidados. Funcionalmente son expresados mediante:

$$AIC = 2\left\{\left[l\left(\hat{\theta}_2\right) - l\left(\hat{\theta}_1\right)\right] - \left(p_2 - p_1\right)\right\},$$

$$BIC = 2\left[l\left(\hat{\theta}_2\right) - l\left(\hat{\theta}_1\right)\right] - \log n\left(p_2 - p_1\right),$$

siendo $l\left(\hat{\theta}_1\right)$ y $l\left(\hat{\theta}_2\right)$ los logaritmos de las funciones de verosimilitud de dos modelos M_1 y M_2 así como p_1 y p_2 son los grados de libertad respectivos de cada modelo y n el numero de observaciones.

Al comparar dos modelos con el mismo conjunto de datos, el mejor modelo será aquel que posea menor valor de criterio de información considerado. Se utilizaran ambos criterios en este trabajo para comparar los modelos con exceso de ceros.

"Cuando las leyes de la matemáticas se refieren a la realidad, no son ciertas; cuando son ciertas, no se refieren a la realidad". (Albert Einstein)

ESTUDIO DE CASO

La ecología de poblaciones se dedica a estudiar las poblaciones formadas por los organismos de una misma especie desde el punto de vista del tamaño (número de individuos), la estructura (sexo y edad) y la dinámica (variación en el tiempo), en otras palabras, estudia las causas de la distribución y abundancia de los agregados de organismos (Soberón y Dirzo, 1989).

Una de las características ecológicas más importantes es la distribución espacial. Iawo (1979) propone un resumen sobre éstas, sus posibles causas y los modelos matemático-estadísticos asociados. Esto se resume en la Tabla 1.

Tabla 1: Distribuciones espaciales, sus posibles causas y los modelos matemático - estadísticos asociados a las mismas.

Distribución Espacial	**Posible Explicación del Mecanismo Básico**	**Modelo Correspondiente**
Aleatorias o al Azar	*Los individuos se reparten independientemente y aleatoriamente.	Serie de Poisson
Agregadas o de Contagio	*Colonias distribuidas aleatoriamente (tamaño medio de colonia fijo). *Atracción mutua: La presencia de un individuo en una unidad cuadrado incrementa la probabilidad de encontrar otros individuos en la misma unidad. *Reparticiones de individuos con una definida tendencia a la agregación (en el sentido de la constante probablemente debida, en muchos casos, a la naturaleza heterogénea de las unidades de muestreo. *Colonias repartidas contagiosamente (tamaño de colonia fijo); muchos otros procesos pueden incluirse.	*Poisson Binomial *Neyman tipo A *Thomas *Binomial negativa (Quenonille) Polya-Aeppli. *Binomial negativa (Polya-Eggenberger con un k común) Serie binomial negativa con un k común. No hay modelos matemáticos (p.e, reparticiones de colonias que siguen la serie binomial negativa con un k común).
Regulares o Uniformes	*El máximo número de individuos posible en una unidad está limitado por competición o alguna otra razón. *Repulsión mutua; pero no limitación en la capacidad de una unidad: la presencia de un individuo en una unidad decrece la probabilidad de que otros individuos se encuentren en la mayor unidad.	Binomial positiva con un común N (N: capacidad de una unidad) Serie binomial positiva Distribuciones completamente uniformes.

Si bien estas estrategias han sido la forma estándar de uso, no se ha profundizado sobre la manera más idónea de describir las poblaciones que presentan superdispersión.

La ecología de poblaciones se vuelve un campo de aplicación válido para este tipo de modelos. En este trabajo se focalizará en las poblaciones animales –invertebrados, específicamente artrópodos-, en la que los individuos viven en el espacio, en hábitats discontinuos, que pueden ser divididos en unidades discretas como hojas, flores, suelo, frutos o plantas enteras tomadas como unidades de hábitat, siendo habitual que se analicen mediante el uso de las frecuencias obtenidas por los conteos del número de individuos por unidad de hábitat (Cadahia, 1977).

Las poblaciones de invertebrados, específicamente, los lepidópteros tortrícidos (como la carpocapsa), presentan generalmente un patrón de distribución agregada (Capuccino y Price, 1995). Este patrón genera en el muestreo estacional una gran cantidad de ceros, ya que dependen directamente tanto de las condiciones climáticas como de la disponibilidad del recurso.

La carpocapsa (Lepidoptera: Tortricidae) es una de las especies cosmopolitas más agresivas y difíciles de controlar en cultivos de pomáceas (Fernández, 2007). Figura 1.

Figura 1: *Cydia Pomonella* (L.) adulto.

Es un insecto de ciclo de vida multivoltino con diapausa facultativa y en la región del Alto Valle de Río Negro, Argentina, cumple con tres generaciones completas por año. La tercera generación se ve interrumpida al final de una temporada y se completa a la siguiente, luego de pasar por un período de diapausa (Cichón, 1999). Figura 2.

El primer vuelo de adultos se extiende desde fines de septiembre hasta mediados de diciembre. El segundo y tercer vuelo se superponen en el tiempo ocurriendo desde mediados de diciembre a mediados de febrero y desde inicio de febrero a fines de marzo respectivamente (Cichón, 1999). A cada vuelo de adultos le corresponde una generación de larvas. Figura 3.

Figura 2: Ciclo completo anual. Tres generaciones y su desarrollo: Huevo, Larva, Pupa y Adulto.

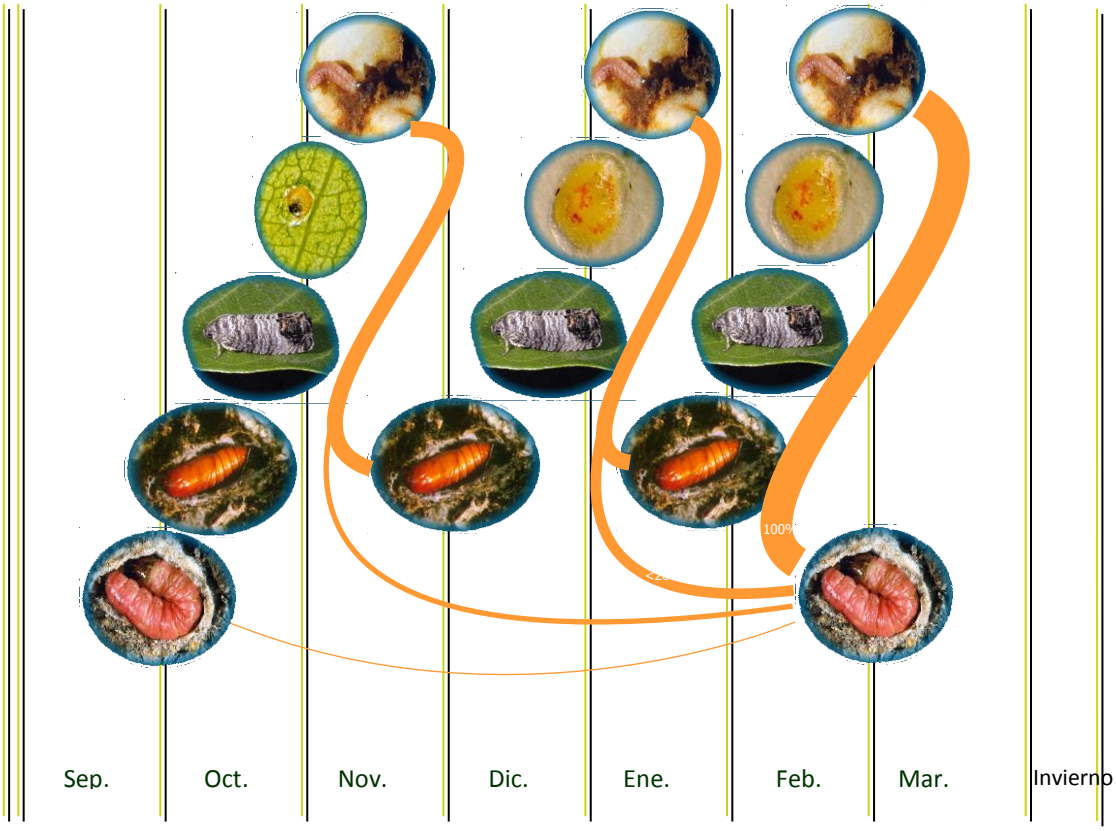
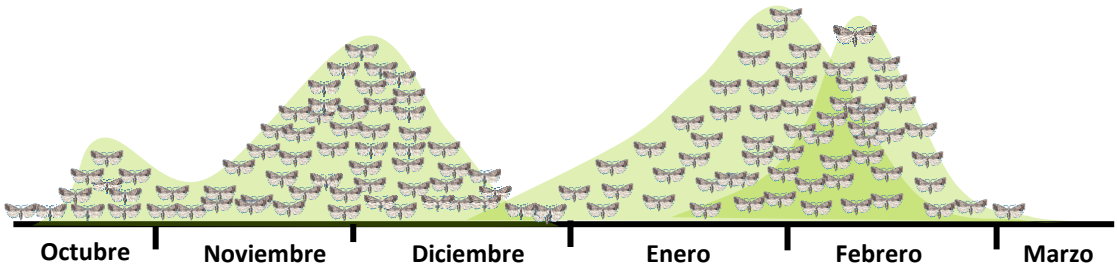


Figura 3: Esquema de los vuelos de adultos a lo largo de una temporada.



El codlemone, feromona producida por la hembra, se utiliza tanto para su control como para su monitoreo. En este último caso se definen umbrales de captura en trampas, para decidir acciones de control y de esta manera evitar las pérdidas económicas que genera la plaga.

Sin embargo, siguiendo a Fernández (2007) la utilización de la codlemone para el monitoreo de la carpocapsa se ve restringida en algunos casos por su falta de precisión para detectar adecuadamente diferentes niveles poblacionales de la plaga, ya que son muchos los factores que influencian el nivel de capturas. Entre ellos están los que dependen de la trampa en sí como el diseño, el tipo de emisor, el mantenimiento, la ubicación y la densidad de trampeo; los que dependen de la población del insecto como la densidad de la población, la relación de sexos, la edad de los individuos y por último los dependientes de las condiciones meteorológicas como la temperatura, el viento y la lluvia (Riedl et al. 1986, según Fernández).

Comparar trampas con diferentes cebos de atracción, utilizadas para el monitoreo de la plaga es una manera de volver eficiente los recursos disponibles. Entre estas nuevas opciones actualmente se ha incorporado la combinación de codlemone con éster de pera.

Metodología y Toma de Datos

Los datos utilizados para el análisis, corresponden al conjunto de datos muestreados por el grupo de sanidad de la Estación Experimental del INTA Alto Valle (Figura 4), en las temporadas 2003-2004 y 2004-2005.Las trampas estaban colocadas en montes frutales de perales (n=6283) y manzanos (n=10950) en un bloque de 200 ha contiguas. Los manzanos y perales estaban intercalados. El bloque estuvo tratado con feromonas de confusión sexual durante toda la temporada.

Dentro de cada temporada se considero la emergencia de las tres generaciones anuales, correspondientes al ciclo biológico del mismo. (Fernández, 2007).

Figura 4: Cuadro de plantación de Manzanas, Estación Experimental INTA Alto Valle.

Se utilizaron cuatro tipos de cebos (Figura 5): trampa 1 (L2), es una trampa tipo delta que contiene un cebo con 1 mg de codlemone (compuesto químico denominado (E,E)-8,10-dodecadien-1-ol (EEOH) (Roelofs, et al. 1971); trampa 2 (Megalure), es una trampa tipo delta con un cebo con 10 mg de codlemone; trampa 3 (DA2313), trampa tipo delta con un cebo con 3 mg de (2E, 4Z) decadienonato de etilo (de nombre común: éster de pera), y la trampa 4 (CM-DA o

Combo), trampa tipo delta con un cebo con 3 mg de éster de pera más 3 mg de codlemone (Tabla 2).

Figura 5: Cebos utilizados en las trampas.

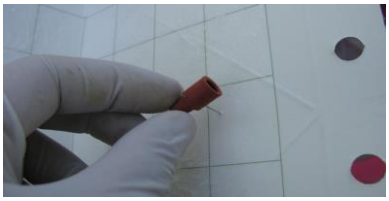


Tabla 2: Trampas delta y cebos utilizados para la captura de adultos de (Fernández, 2007)**.**

Trampa	Cebo	Contenido
1: L2	Codlemone	1 mg
2: Mega	Codlemone	10 mg
3: DA2313	Ester de pera	3 mg
4: Combo ó CM-DA	Codlemone + Ester de pera	3 mg + 3 mg

En cada parcela se colocó un grupo de cada trampa dispuestas en forma aleatoria a una distancia mínima de 25 m unas de otras, para evitar interferencias en las capturas (Trematerra *et al*. 2004; Knight& Light, 2005; Fernández *et al*. 2010). Figura 6.

Figura 6: Trampas para la captura de la carpocapsa.

Programas Estadísticos

Los ajustes de los modelos propuestos, así como la estadística descriptiva, se realizaron con el programa STATA SE versión 11.0.

Se ejecutaron a través de las rutinas GLM, para los modelos corrientes de Poisson y Binomial Negativa, así como los gráficos de los residuos; y a través de las rutinas ZIP y ZINB para el ajuste de las distribuciones con exceso de ceros, así como la comparación de los modelos mediante el test de Vuong y los criterios de AIC y BIC así como los gráficos de los modelos ajustados.

En las tablas que se presentan en el capítulo siguiente se organizo la información de los modelos propuestos de la siguiente manera: la parte del modelo que se está estimando, las variables y sus respectivas categorías, la estimación de las mismas; el error muestral, la significancia de los efectos considerados, a través del valor p, dado por el estadístico de Wald. En la columna final se manifiesta el valor del *Odds Ratio* y su intervalo de confianza del 95%, para poder expresar en términos de chances.

Para los modelos con exceso de ceros, en la base de las tablas se resumen los criterios de información para la selección de modelos (AIC y BIC). El test de Vuong con el valor z y la probabilidad correspondiente. Para los modelos Binomial Negativo la estimación del parámetro de dispersión α y su respectivo intervalo de confianza (introducido y explicado en el capítulo V).

El cambio es lo que nos da un propósito (Heráclito).

ANÁLISIS DE DATOS

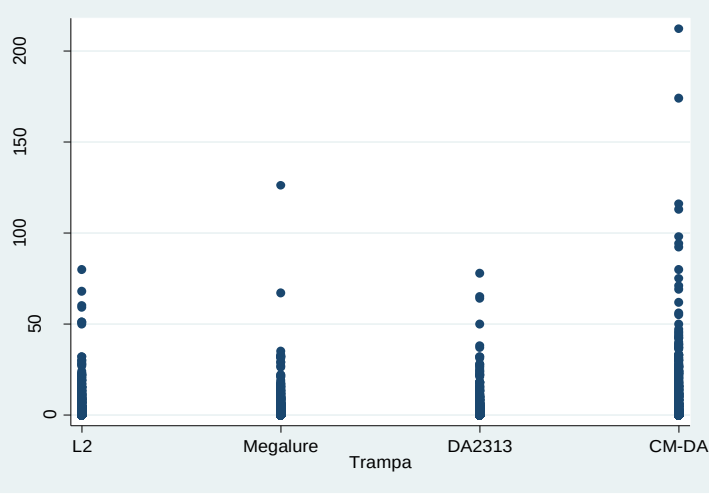
A partir de una revisión descriptiva, la cual es presentada a continuación, se intenta motivar el uso de modelos probabilísticos que sean adecuados para respuestas de conteos con exceso de ceros, es decir aquellos que contemplen la superdispersión por exceso de valores nulos.

En la Tabla 3 se observan los valores obtenidos para los estadísticos resumen para el conjunto de datos anteriormente presentado. Se muestra que en todas las trampas, el valor mediano es cero, indicando que el 50% de las observaciones como mínimo son nulas. La captura promedio es de tres insectos, con la salvedad de la trampa 4 que captura en promedio 8 insectos, pareciendo ser la más efectiva dada la combinación de componentes. La Figura 7 pone de manifiesto, a través del coeficiente de variación que el promedio resulta no representativo dada la alta variabilidad Sd (desvío estándar > al promedio), este fenómeno es provocado por el exceso de valores nulos, siendo el posible causante de la superdispersión.

Tabla 3: Descripción del conteo de *C. pomonella (L.)*por trampas a través de las siguientes medidas de resumen: promedio, mediana, desvío estándar, percentil 75, error estándar, coeficiente de variación, percentil 25 y valor mínimo.

Medidas	Trampa 1	Trampa 2	Trampa 3	Trampa 4
Promedio	3,7617	3,0717	3,0400	8,1500
Mediana (p50)	0	0	0	0
Desvío Estándar (Sd)	8,4295	7,8909	7,5208	1,8212
Percentil 75 (p75)	4	3	3	10
Error Estándar [se(mean)]	0,3441	0,3221	0,3070	0,7435
Coeficiente de Variación (Cv)	2,2409	2,5689	2,4740	0,2234
Percentil 25 (p25)	0	0	0	0
Valor mínimo(Min)	0	0	0	0

Figura 7: Conteo de capturas de *C. pomonella (L.)*dentro de cada trampa.



En la Tabla 4 se manifiesta que el promedio de captura por especie es mayor la captura en las zonas con manzana, que en las zonas con pera. Así mismo en los perales, hay un 50% de trampas sin capturas, mientras que en los manzanos el 50% de las trampas capturan al menos un insecto. La variabilidad es superior en la especie pera que en la manzana, representado a través de el desvío estándar y en el coeficiente de variación.

Tabla 4: Descripción del conteo de *C. pomonella (L.)*por especie, a través del promedio, de la mediana (p50), del desvío estándar (Sd), del percentil 75 (p75), del error estándar (se(mean)), del coeficiente de variación (Cv), del percentil 25 (p25) y del valor mínimo (Min) .

Medidas	Manzana	Pera
Promedio	5,7642	3,2475
p50	1	0
Sd	1,4312	7,8475
p75	5	3
se(mean)	0,4132	0,2265
Cv	0,2483	2,4165
p25	0	0
Min	0	0

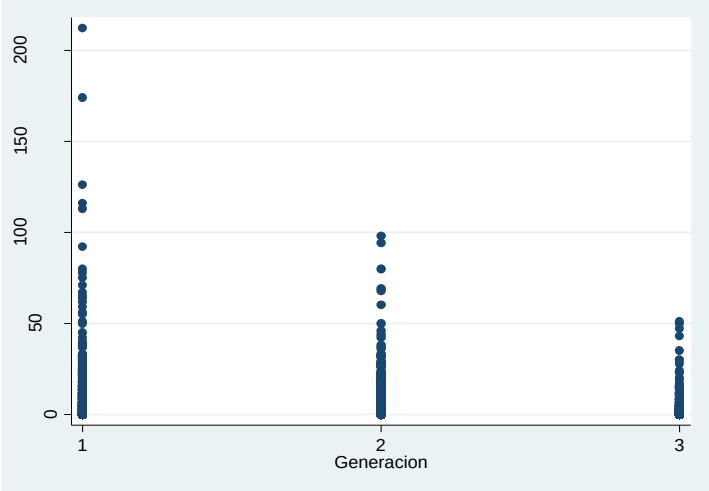
Figura 5: Conteo de capturas de *C. pomonella (L.)*dentro de cada especie.

Respecto a las generaciones, en la Tabla 5, se observa como estas van mermando en abundancia a medida que se completa el ciclo anual, lo cual se ve representado en el promedio de capturas por generación. El percentil 50 nos indica que en la primer generación se encuentra un insecto al menos en el 50% de las observaciones, no así en la segunda y tercer generación que no hay capturas; reforzando que la probabilidad de no encontrar insectos en las trampas a medida que termina el ciclo es muy elevada, no solo por el ciclo biológico de la carpocapsa, sino por la extra variación que existe. Obsérvese que el desvío supera ampliamente el promedio de capturas en las dos últimas generaciones, generando coeficientes de variación muy aumentados.

Tabla 5: Estadística descriptiva del conteo de *C. pomonella (L.)*por generación. Medidas que resumen la información: promedio, mediana (p50), desvío estándar (Sd), percentil 75 (p75), error estándar (se(mean)), coeficiente de variación (Cv), percentil 25 (p25) y valor mínimo (Min).

Medidas	G1	G2	G3
Promedio	7,2813	4,6163	1,6200
p50	1	0	0
Sd	1,6411	9,7298	4,9494
p75	8	5	1
se(mean)	0,5802	0,3440	0,1750
Cv	0,2254	2,108	3,055
p25	0	0	0
Min	0	0	0

Figura 6: Conteo de capturas de *C. pomonella (L.)* dentro de cada generación.

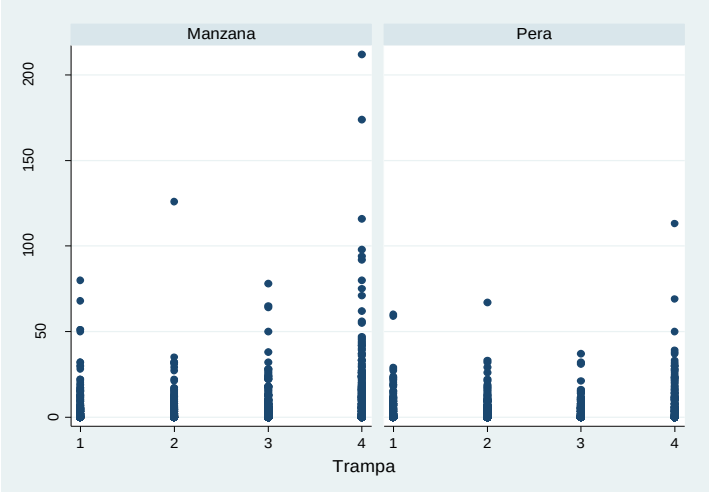


A nivel conjunto, la Tabla 6 muestra en las especies y las trampas, que el nivel medio de capturas es más elevado en la manzana que en la peras con alta variabilidad en ambas especies y destacándose la trampa 4 por sobre las demás en ambas especies con un promedio en manzanas de casi 11 capturas y 5,5 capturas en peras. El 25% de las observaciones son nulas en ambas especies, y en peras llega hasta el 50% de las observaciones, mientras que en manzana difieren según la trampa. El coeficiente de variabilidad es similar y elevado para todas las trampas y especie.

Tabla 6: Estadística descriptiva del conteo de *C. pomonella (L.)*para el conjunto de especie y trampa. Medidas que resumen la información: promedio, mediana (p50), desvío estándar (Sd), percentil 75 (p75), error estándar (se(mean)), coeficiente de variación (Cv), percentil 25 (p25) y valor mínimo (Min).

Medidas	**Manzana**				**Peras**			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Promedio	4,2667	3,7533	4,2000	10,8367	3,2567	2,3900	1,8800	5,4633
p50	0,5	1	0	2	0	0	0	0
Sd	9,4591	9,2385	9,4847	22,8489	7,2364	6,2016	4,5447	11,3105
Se (media)	0,5461	0,5334	0,5476	1,3192	0,4178	0,3580	0,2624	0,6530
p75	4	4	5	13	3	2	2	6
Cv	2,2170	2,4614	2,2583	2,1085	2,2220	2,5948	2,4174	2,0703
p25	0	0	0	0	0	0	0	0
Min	0	0	0	0	0	0	0	0

Figura 7: Conteo de capturas de *C. pomonella (L.)* dentro de cada trampa por cada especie.



En la Tabla 7, se puede observar de manera conjunta que a medida que avanza el ciclo generacional, disminuye el promedio de captura. El 25% de las observaciones son nulas en todas las generaciones, trampas y especies, así como el 50% de las observaciones en peras, mientras que en manzanas son nulas a partir de la tercera generación.

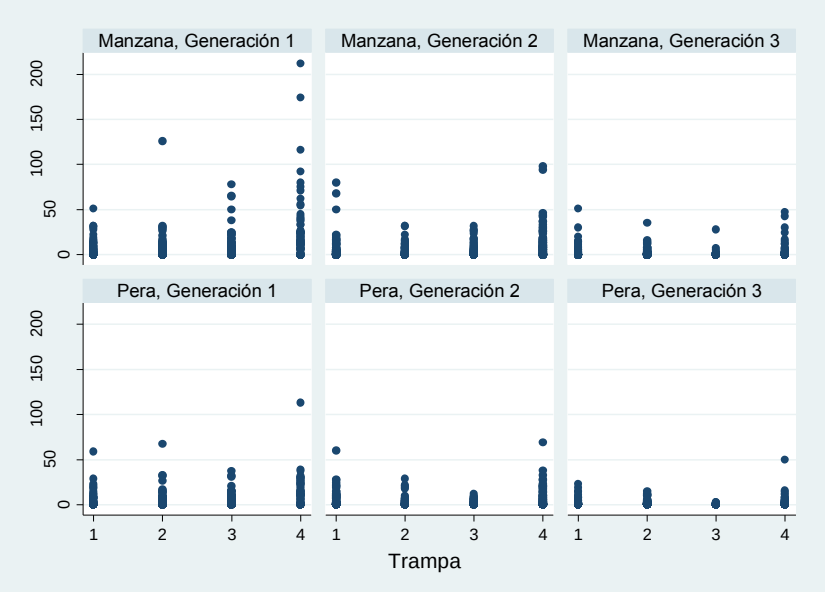
Tabla 7: Estadística descriptiva del conteo de *C. pomonella* (L), para cada generación, dentro de cada especie y por cada trampa. Medidas que resumen la información: promedio, mediana (p50), desvío estándar (Sd), percentil 75 (p75), error estándar (se(mean)), coeficiente de variación (Cv), percentil 25 (p25) y valor mínimo (Min).

Medidas	**Generación 1**							
	Manzana				**Peras**			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Promedio	4,99	5,66	7,89	19,62	3,87	3,65	3,91	8,66
p50	1,5	2	4	11	0	0	0	0
Sd	8,5923	13,9075	13,7576	33,1753	8,1398	8,9108	6,9123	14,9188
se(mean)	0,8592	1,3908	1,3758	3,3175	0,8140	0,8911	0,6912	1,4919
p75	6	5,5	8	24	4	3,5	5	13
Cv	1,7219	2,4572	1,7437	1,6909	2,1033	2,4413	1,7678	1,7227
p25	0	0	0	0	0	0	0	0
Min	0	0	0	0	0	0	0	0
	Generación 2							
	Manzana				**Peras**			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Promedio	5,23	3,98	4,04	10,07	4,02	2,33	1,55	5,71
p50	1,5	2	1	5	0	0	0	0
Sd	12,0536	6,0469	6,8827	16,5348	8,5102	5,1090	2,6643	10,4062
se(mean)	1,2054	0,6047	0,6883	1,6535	0,8510	0,5109	0,2664	1,0406
p75	4	6	5	13	4,5	3	2,5	8
Cv	2,3047	1,5193	1,7036	1,6420	2,1170	2,1927	1,7189	1,8225
p25	0	0	0	0	0	0	0	0
Min	0	0	0	0	0	0	0	0

	Generación 3							
	Manzana				Peras			
	T1	T2	T3	T4	T1	T2	T3	T4
Promedio	2,58	1,62	0,67	2,82	1,88	1,19	0,18	2,02
p50	0	0	0	0	0	0	0	0
Sd	6,8390	4,4125	2,9475	7,7020	4,0733	2,7550	0,5574	5,7577
se(mean)	0,6839	0,4413	0,2948	0,7702	0,4073	0,2755	0,0557	0,5758
p75	2	1	0	2	2	1	0	2
Cv	2,6508	2,7238	4,3993	2,7300	2,1666	2,3151	3,0967	2,8503
p25	0	0	0	0	0	0	0	0
Min	0	0	0	0	0	0	0	0

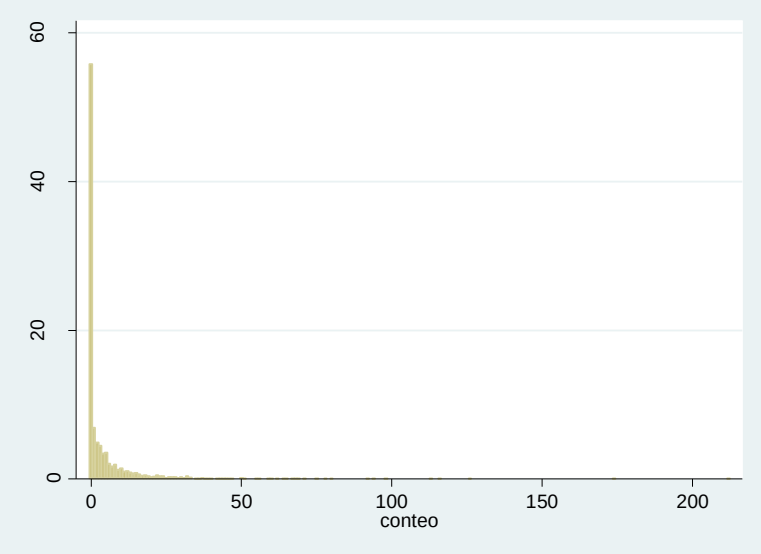
En la Figura 11 queda referenciado que el coeficiente de variación se mantiene relativamente uniforme y elevado en todas las generaciones, oscilando en puntos porcentuales entre el 150 y el 400. Así como la variabilidad relativa excede al promedio en todas las combinaciones.

Figura 8: Conteo de capturas de *C. pomonella (L.)* dentro de cada generación, para cada especie y por cada trampa.



La Figura 12 resume la alta proporción de ceros presente en el conjunto de datos, este fenómeno genera superdispersión ya que cuanto mayor es la probabilidad de ceros, mayor es la varianza de la variable. A su vez sesga la información y genera estimaciones erróneas. Es por este motivo que se debe tener en cuenta las diferentes funciones de probabilidad al proponer un modelo formal del comportamiento de la carpocapsa.

Figura 9: Conteo de capturas de *C. pomonella (L.)* dentro de cada trampa.



Ajuste de Modelos

Respecto a lo presentado en los capítulos precedentes, en términos del trinomio de un MLG la variable aleatoria $\mathbf{Y}$, conteo de capturas de *C. pomonella (L.)*en trampas $\mathbf{Y} \approx P(\boldsymbol{\mu};\varphi)$sigue una distribución Poisson, cuyo predictor lineal se propone como sigue:

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i}$$

con $\beta_i$como el conteo en cada trampa, siendo β_0 la ordenada al origen, el promedio general para el modelo propuesto. Así, para *C. pomonella (L.)*el β_1 indicaría el efecto de la Trampa i en la que cayó el individuo (i=1,...4); β_2, el efecto de la Especie donde se ubicaba la trampa (i=1,2) y el $\beta_3$el efecto de la Generación, el estado biológico en el que se encontraba el individuo en la temporada (i=1,2,3); los sucesivos indicarían los efectos de las interacciones dobles β_4, β_5, β_6 y triple β_7. La función de enlace a utilizar será, naturalmente,

la canónica (logaritmo). Los componentes sistemáticos se modificarán de manera jerárquica en la medida que se propongan los distintos modelos minimales (de efectos principales) o maximales (con interacción). Los componentes aleatorios también se irán verificando, comenzando con el modelo Poisson, siguiendo con el modelo de Binomial Negativo, y modelos de exceso de ceros como el ZIP y el ZINB.

En relación a los modelos ZIP y ZINB, la función de enlace entre la media y el predictor lineal se descompone en dos segmentos, uno de ellos, asociado a la parte del modelo con exceso de ceros y el otro a la parte del modelo sin exceso de ceros (Cap. V). Por ejemplo para el modelo ZIP, la función de enlace para la parte con exceso de ceros, está dada por la función logit: $\text{logit}(\boldsymbol{\pi}) = \ln\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \mathbf{G}\gamma$, siendo $\mathbf{G}$ la matriz de variables explicativas $x_i \,/\, i = 1,2,3$ asociada a la generación y γ el vector de los parámetros a ser estimados de variables asociadas a la parte con exceso de ceros del modelo. La función de enlace para la parte sin exceso de ceros, es la misma que la utilizada por el modelo de Poissoncorriente $\ln(\boldsymbol{\mu}) = \mathbf{B}\beta$, siendo $\mathbf{B}$ la matriz de las variables explicativas especie y trampas, y β el vector de parámetros a ser estimado correspondiente a la parte sin exceso de ceros del modelo.

Modelos de Poisson

La Tabla 8 muestra que todos los efectos son significativos. La especie referente (manzana) captura el doble que la pera. La trampa 4 tiene 2 veces más de capturar individuos con respecto a la trampa 1, siendo este el referente, el testigo comercial.

Tabla 8: Modelo de Poisson con efectos principales.

Modelo Poisson						
Parte	Variable	Categorías	Estimación	Error Std.	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*referencia*	-	-	
		Peras	-0,5737	0,011	0,000	0,5633(0,542-0,586)
	Trampas	1	*referencia*	-	-	
		2	-0,2026	0,026	0,000	0,8165(0,768-0,868)
		3	-0,213	0,025	0,000	0,8081(0,759-0,859)
		4	0,773	0,055	0,000	2,1665(2,061-2,277)
	Generación	1	*referencia*	-	-	
		2	-0,4557	0,013	0,000	0,6339(0,608-0,660)
		3	-1,5028	0,006	0,000	0,2224(0,209-0,236)
	Constante		2,051			
Criterios selección Modelo				AIC=29195,74	BIC=29236,22	gl.7

Al agregar la interacción al modelo de Poisson (Tabla 9), la estimación según el AIC mejora levemente y comienzan a ser no significativos algunos efectos principales como el de la trampa 3. Nótese que el efecto de especie, según el OR, ya no indica el doble de captura en manzana que en pera. Así como la trampa 4 captura casi un 40% más.

Tabla 9: Modelo de Poisson con efecto de interacción.

Modelo Poisson con Interacción						
Parte	Variable	Categorías	Estimación	Error Std.	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,27	0,032	0,000	0,763(0,702-0,829)
	Trampas	1	*Referencia*	-	-	
		2	-0,128	0,035	0,002	0,879(0,811-0,953)
		3	-0,015	0,039	0,691	0,984(0,910-1,063)
		4	0,932	0,083	0,000	2,539(2,380-2,709)
	Generación	1	*Referencia*	-	-	
		2	-0,456	0,053	0,000	0,634(0,608-0,660)
		3	-1,502	0,038	0,000	0,222(0,209-0,236)
	Especie*Trampa	1	*Referencia*	-	-	
		2	-0,181	0,034	0,005	0,834(0,735-0,945)
		3	-0,533	0,013	0,000	0,586(0,515-0,667)
		4	-0,415	0,006	0,000	0,660(0,596-0,731)
	Constante		2,051			
Criterios selección Modelo				AIC=29110,22	BIC=29168,06	gl.11

Modelo Binomial Negativo

La Tabla 10, muestra el resumen del modelo BN, se observa un mejor ajuste respecto a los de Poisson presentados anteriormente (Tablas 8 y 9) basado

en los criterios AIC y BIC. Los efectos son todos significativos pero hay que distinguir que la trampa 4, captura una vez y media más que el resto. La constante de dispersión, dado $\alpha > 0$ nos indica que este modelo está capturando la dispersión generada en el conjunto de datos, es decir la extra-variación respecto al promedio.

Tabla 10: Modelo Binomial Negativo con efectos principales.

Modelo Binomial Negativa						
Parte	Variable	Categorías	Estimación	Error Std.	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,537	0,052	0,000	0,584(0,489-0,697)
	Trampas	1	*Referencia*	-	-	
		2	-0,3239	0,092	0,012	0,723(0,562-0,930)
		3	-0,6056	0,071	0,000	0,545(0,421-0,706)
		4	0,5192	0,213	0,000	1,680(1,310-2,156)
	Generación	1	*Referencia*	-	-	
		2	-0,4646	0,068	0,000	0,628(0,507-0,777)
		3	-1,5152	0,025	0,000	0,219(0,175-0,274)
	Constante		2,24			
α			4,46	0,18		(4,121-4,828)
Criterios selección Modelo				AIC=11442,05	BIC=11514,812	gl.8

Al agregar el efecto de interacción (Tabla 11) mejoran aún más los criterios de información. Pero solo quedan los efectos de trampa 4, y generación como altamente significativos.

Tabla 11: Modelo Binomial Negativo con efecto de interacción.

Modelo Binomial Negativa con Interacción						
Parte	Variable	Categorías	Estimación	Error Std.	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*referencia*	-	-	
		Peras	-0,288	0,134	0,108	0,749(0,527-1,065)
	Trampas	1	*referencia*	-	-	
		2	-0,25	0,139	0,162	0,778(0,547-1,105)
		3	-0,336	0,129	0,062	0,714(0,501-1,017)
		4	0,675	0,349	0,000	1,964(1,385-2,783)
	Generación	1	*referencia*	-	-	
		2	-0,468	0,068	0,000	0,625(0,505-0,774)
		3	-1,517	0,025	0,000	0,219(0,175-0,274)
	Especie*Trampa	1	*referencia*	-	-	
	Especie*Trampa	2	-0,136	0,222	0,593	0,872(0,529-1,438)
	Especie*Trampa	3	-0,559	0,145	0,029	0,571(0,346-0,943)
	Especie*Trampa	4	-0,303	0,185	0,228	0,738(0,451-1,208)
	Constante		2,112			
α			4,444	0,179		(4,106-4,811)
Criterios selección Modelo				AIC=9981,24	BIC=9991,85	gl.11

Modelos Particionados por Generación

En las Tablas 12, 13 y 14se observa que el efecto especie es siempre significativo, no así en el caso de las trampas, ni en las interacciones. Dónde la trampa 4 va perdiendo su significatividad a medida que avanzan las generaciones, concretamente en la generación 3.

Tabla 12: Modelo de Poisson en la generación 1.

Poisson Generación 1						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,254	0,0525	0,000	0,7755(0,679-0,885)
	Trampas	1	*Referencia*	-	-	
		2	0,126	0,069	0,040	1,134(1,005-1,279)
		3	0,458	0,0904	0,000	1,581(1,413-1,768)
		4	1,369	0,1971	0,000	3,931(3,563-4,337)
	especie*trampas	1	*Referencia*	-	-	
		2	-0,184	0,0793	0,053	0,831(0,689-1,002)
		3	-0,448	0,0586	0,000	0,638(0,534-0,765)
		4	-0,564	0,045	0,000	0,569(0,487-0,665)
	Constante		1,607		0,000	

Tabla 13: Modelo de Poisson en la generación 2.

Poisson Generación 2						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,263	0,0509	0,000	0,768(0,674-0,875)
	Trampas	1	*Referencia*	-	-	
		2	-0,273	0,0506	0,000	0,760(0,668-0,866)
		3	-0,258	0,0511	0,000	0,772(0,678-0,879)
		4	0,655	0,1037	0,000	1,925(1,732-2,139)
	especie*trampas	1	*Referencia*	-	-	
		2	-0,272	0,0806	0,010	0,761(0,618-0,937)
		3	-0,695	0,0576	0,000	0,499(0,398-0,625)
		4	-0,304	0,0623	0,000	0,737(0,625-0,870)
	Constante		1,654			

Tabla 14: Modelo de Poisson en la generación 3.

Poisson Generación 3						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*referencia*	-	-	
		Peras	-0,316	0,069	0,010	0,728(0,604-0,879)
	Trampas	1	*referencia*	-	-	
		2	-0,465	0,0629	0,000	0,627(0,516-0,764)
		3	-1,348	0,0356	0,000	0,259(0,198-0,339)
		4	0,889	0,094	0,302	1,093(0,923-1,294)
	especie*trampas	1	*referencia*	-	-	
		2	0,008	0,1554	0,958	1,008(0,745-1,364)
		3	-0,998	0,104	0,000	0,368(0,212-0,641)
		4	-0,017	0,1307	0,898	0,983(0,757-1,276)
	Constante		0,948			

Como puede observarse en las Tablas 15, 16 y 17, cuando es por generación el modelo Binomial Negativo no puede captar la extra-variación producida por la parte fija del modelo, ya que la mayoría de los efectos no resultan significativos.

Tabla 15: Modelo Binomial Negativa en la generación 1.

Binomial Negativa Generación 1						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,254	0,224	0,378	0,775(0,440-1,365)
	Trampas	1	*Referencia*	-	-	
		2	0,126	0,325	0,661	1,134(0,646-1,990)
		3	0,458	0,452	0,109	1,581(0,903-2,769)
		4	1,369	1,119	0,000	3,931(2,250-6,869)
	especie*trampas	1	*Referencia*	-	-	
		2	-0,184	0,339	0,651	0,831(0,374-1,848)
		3	-0,448	0,259	0,271	0,639(0,288-1,418)
		4	-0,564	0,23	0,163	0,569(0,258-1,256)
	Constante		1,607			

Tabla 16: Modelo Binomial Negativa en la generación 2.

Binomial Negativa Generación 2						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,263	0,22	0,359	0,768(0,438-1,348)
	Trampas	1	*Referencia*	-	-	
		2	-0,273	0,218	0,341	0,760(0,433-1,334)
		3	-0,258	0,221	0,368	0,772(0,440-1,354)
		4	0,655	0,547	0,021	1,925(1,103-3,359)
	especie*trampas	1	*Referencia*	-	-	
		2	-0,272	0,311	0,505	0,761(0,342-1,695)
		3	-0,695	0,205	0,091	0,499(0,223-1,17)
		4	-0,304	0,297	0,451	0,737(0,334-1,626)
	Constante		1,654			

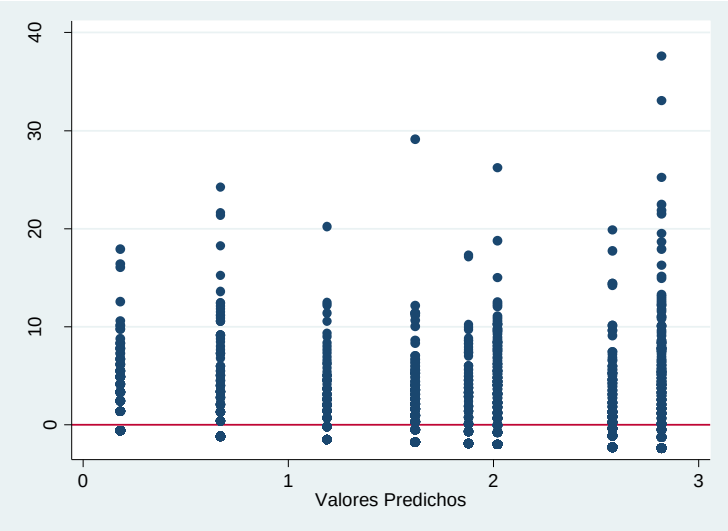
Tabla 17: Modelo Binomial Negativa en la generación 3.

Binomial Negativa Generación 3						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,316	0,261	0,378	0,728(0,360-1,472)
	Trampas	1	*Referencia*	-	-	
		2	-0,465	0,226	0,196	0,627(0,310-1,271)
		3	-1,348	0,096	0,000	0,259(0,125-0,538)
		4	0,089	0,389	0,803	1,093(0,544-2,197)
	especie*trampas	1	*Referencia*	-	-	
		2	0,008	0,516	0,987	1,008(0,369-2,754)
		3	-0,998	0,208	0,077	0,368(0,122-1,115)
		4	-0,017	0,498	0,973	0,983(0,364-2,654)
	Constante		0,947			

Residuales de los Modelos Corrientes

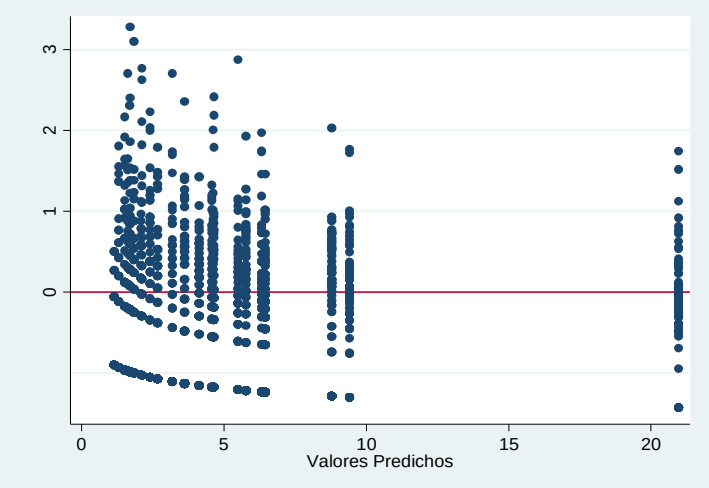
En la Figura 13 los residuales, es decir la diferencia entre lo observado y lo esperado, nos muestran la amplia variabilidad y la gran concentración entorno al cero.

Figura 10: Residuos del modelo de Poisson.



En la Figura 14 se puede observar que a diferencia de los residuos de Poisson (Figura 13) los valores se concentran mucho mas, dado el rango de variación y muestran que comienzan a ser más equidispersos aunque continúan muy concentrados en torno al cero.

Figura 11: Residuos del modelo Binomial Negativo.



A través del análisis de los residuales de los distintos modelos (Figuras 15 y 16) se puede destacar que a medida que avanzan las generaciones más concentración de ceros se observa, esto nos indica que sería conveniente introducir la generación como la variable que presenta un efecto de superdispersión o de inflación (*inflated*) en el modelo.

Figura 12: Residuos del modelo Poisson particionado por generación.

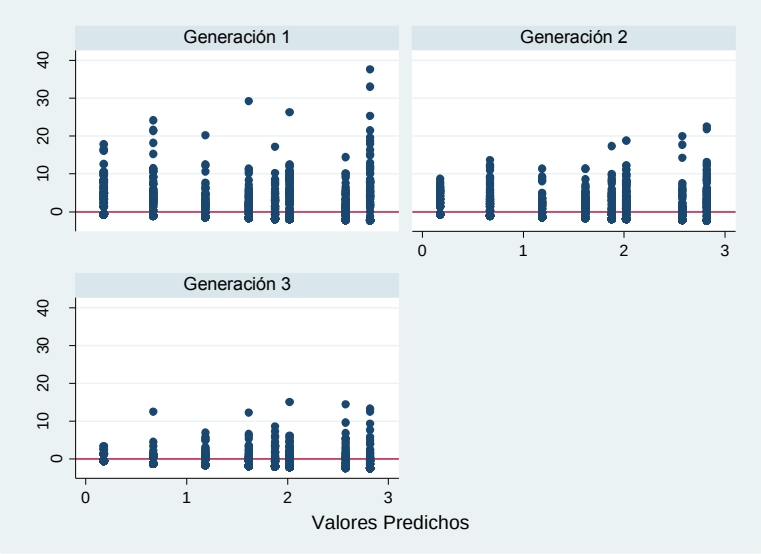
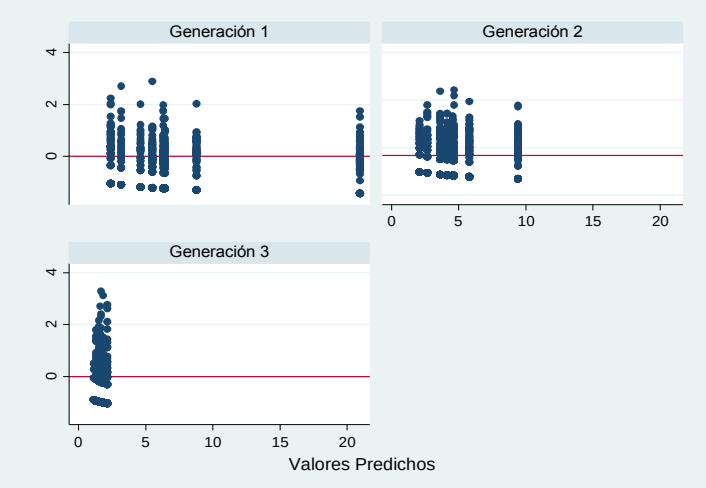


Figura 13: Residuos del modelo Binomial Negativo particionado por generación.



Modelos con Superdispersión (*Inflated*) de Ceros

En la Tabla 18, el test de Vuong indica que el modelo ZIP es más adecuado que el modelo de Poisson corriente confirmando lo que indican los criterios AIC y BIC, mientras que en el modelo de Poisson corriente (Tabla 9) el AIC es de 29195,74, en este se ve que el valor es de 19108,15. Los efectos principales son significativos.

Tabla 18: Modelo ZIP con efectos principales.

Modelo ZIP						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*referencia*	-	-	
		Peras	-0,297	0,0149	0,000	0,743(0,714-0,772)
	Trampas	1	*referencia*	-	-	
		2	-0,179	0,0263	0,000	0,836(0,786-0,889)
		3	-0,065	0,0295	0,037	0,936(0,880-0,996)
		4	0,690	0,0508	0,000	1,994(1,897-2,096)
Inflate	Generación	1	*referencia*	-	-	
		2	0,150	0,1001	0,134	1,162(0,955-1,414)
		3	0,950	0,1046	0,000	2,587(2,107-3,174)
	Constante		-0,122	0,0708	0,085	0,885(0,771-1,017)
Criterios de Selección de Modelos				AIC=19108,15	BIC=19154,42	gl.8
Test de Vuong				z=19,28	p>z=0,0000	

Al incluir el efecto de interacción (Tabla 19) la especie pierde la significatividad. Si bien el test de Vuong continúa siendo significativo, son muchos los grados de libertad que se pierden, para lo poco que se gana en explicación.

Tabla 19: Modelo ZIP con efecto de interacción.

Modelo ZIP con Interacción						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,637	0,0399	0,134	0,938(0,863-1,019)
	Trampas	1	*Referencia*	-	-	
		2	-0,167	0,0346	0,000	0,845(0,780-0,916)
		3	0,075	0,0428	0,059	1,077(0,997-1,165)
		4	0,837	0,0762	0,000	2,309(2,165-2,464)
	especie*trampas	1	*Referencia*	-	-	
		2	-0,002	0,0639	0,970	0,997(0,879-1,131)
		3	-0,378	0,0456	0,000	0,685(0,601-0,780)
		4	-0,383	0,0356	0,000	0,682(0,615-0,755)
Inflate	Generación	1	*Referencia*	-	-	
		2	0,150	0,1001	0,134	1,162(0,955-1,414)
		3	0,950	0,1046	0,000	2,587(2,107-3,174)
	Constante		-0,122	0,0708	0,085	0,885(0,771-1,017)
Criterios de Selección de Modelos				AIC=19028,88	BIC=19092,5	gl.11
Test de Vuong				z=19,26	p>z=0,0000	

El modelo de ZINB con efectos principales presentado en la Tabla 20, mejora respecto al Binomial Negativa corriente así como lo indica el test de Vuong reafirmando los criterios AIC y BIC, comparado con la Tabla 10, se ve que la diferencia es de 1561,958. Los efectos principales son significativos a nivel especie. La constante de dispersión del modelo Binomial Negativo, continúa siendo válido y muestra un mejor ajuste que en las Tablas 10 y 11, ya que aquí $\alpha = 2{,}216$ mientras que en el modelo presentado en la Tabla 10 $\alpha = 4{,}46$ y el de la Tabla 11, modelo que incluye la interacción, $\alpha = 4{,}44$.

Tabla 20: Modelo ZINB con efectos principales.

Modelo ZINB						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*referencia*	-	-	
		Peras	-0,467	0,0532	0,000	0,627(0,530-0,740)
	Trampas	1	*referencia*	-	-	
		2	-0,221	0,0929	0,056	0,801(0,638-1,006)
		3	-0,198	0,0976	0,095	0,819(0,649-1,035)
		4	0,745	0,2418	0,000	2,108(1,684-2,639)
Inflate	Generación	1	*referencia*	-	-	
		2	0,264	0,1868	0,157	1,303(0,903-1,878)
		3	1,411	0,1986	0,000	4,103(2,779-6,056)
	Constante		-1,152	0,2330	0,000	0,316(0,200-0,499)
□			2,216	0,244		(1,785-2,751)
Criterios de Selección de Modelos				AIC=9880,092	BIC=9932,141	gl.9
Test de Vuong				Z=5,44	p>z=0,0000	

El modelo ZINB con efecto de interacción de la tabla precedente (Tabla 21), no muestra una mejora respecto al de efecto principales (Tabla 20) lo que se denota a través de la similitud de los valores del test de Vuong y de los criterios de información, sin embargo si lo hace respecto a los modelos de Poisson de las Tablas 18 y 19.

Tabla 21: Modelo ZINB con efecto de interacción.

Modelo ZINB con Interacción						
Parte	Variable	Categorías	Estimación	Error	valor p	OR (IC 95%)
Fija	Especie	Manzanas	*Referencia*	-	-	
		Peras	-0,211	0,131	0,195	0,809(0,588-1,114)
	Trampas	1	*Referencia*	-	-	
		2	-0,155	0,135	0,326	0,856(0,628-1,166)
		3	0,045	0,169	0,782	1,045(0,762-1,434)
		4	0,937	0,399	0,000	2,553(1,878-3,469)
	especie*trampas	1	*Referencia*	-	-	
		2	-0,122	0,205	0,599	0,885(0,562-1,394)
		3	-0,521	0,140	0,027	0,594(0,374-0,943)
		4	-0,394	0,154	0,084	0,674(0,431-1,054)
Inflate	Generación	1	*Referencia*	-	-	
		2	0,259	0,183	0,156	1,296(0,905-1,857)
		3	1,395	0,193	0,000	4,035(2,761-5,896)
	Constante		-1,126	0,225	0,000	0,324(0,208-0,504))
α			2,176	0,237		(1,757-2,694)
Criterios de Selección de Modelos				AIC=9879,856	BIC=9949,255	gl.12
Test de Vuong				z=5,43	p>z=0,0000	

Por los motivos expuestos en este capítulo, entre los modelos presentados, se seleccionará el ZINB con efectos principales, resumido en la Tabla 20.

DISCUSIÓN

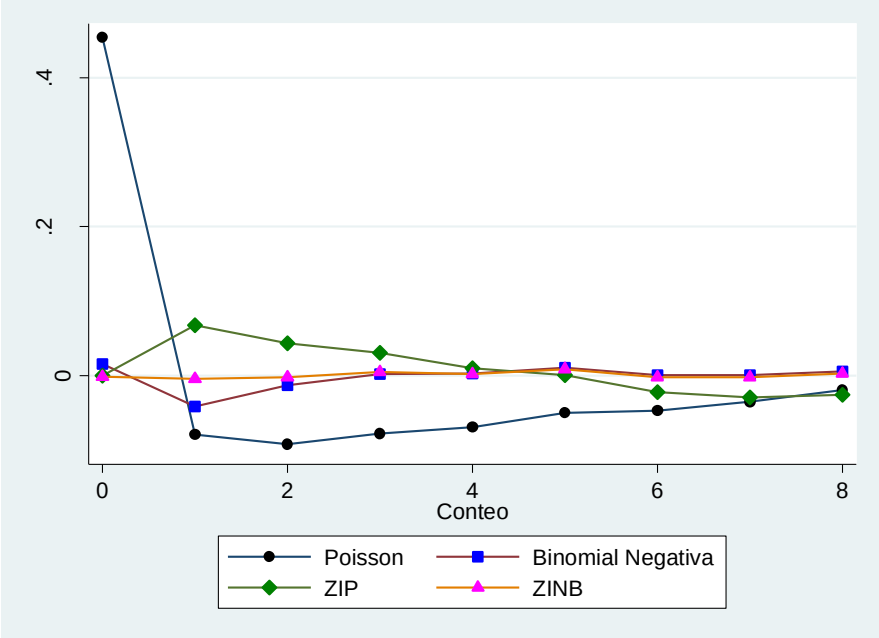
Se pudo obtener un modelo de Poisson, con superdispersión que explica de manera adecuada la relación media-varianza, y a su vez se ha conseguido modelar la respuesta del conteo de *C. pomonella (L.)* por unidad de trampa, evaluada en la región del Alto Valle, y estimar los cambios producidos, así como sus generalizaciones en presencia de extra-variación. (Capítulo I)

Al respecto cabe destacar que el exceso de ceros no es el único fenómeno que está generando la superdispersión en el modelo, por lo observado en el capítulo anterior, pareciera coexistir heterogeneidad no observada, ya que la variabilidad individual no es explicada a través de los modelos propuestos, este comportamiento ha sido observado en estudios de similares características (McCullagh y Nelder, 1989; Díaz, 1999; Hinde y Demetrio, 2005). Más aun, en el proceso de selección de modelos, al proponer solamente un modelo de efectos fijos, el modelo que mejor ajusta es el Binomial Negativo, confirmando que ante situaciones de superdispersión, la distribución Binomial Negativa corrige la probabilidad asociada a valores bajos de conteos que, habitualmente presentan un ajuste deficiente a través del modelo de regresión Poisson (Long, 1997; Veronesi, 2001), con lo cual se evidencia la limitación del modelo de Poisson corriente cunado existe superdispersipon ocasionada por la presencia de exceso de ceros. Puede observarse en la Figura 16, como ante presencia de superdispersión omitida (modelos Poisson y Binomial Negativa) los errores estándar son incorrectos y seriamente subestimados, generando malos resultados sobre la significatividad individual (Hinde y Demetrio, 2005).

Cuando se proponen modelos de ceros modificados, se tiene en cuenta la alternativa de modificar tanto la media como la varianza, y no solamente el incremento de varianza como en el modelo de BN (Long, 1997; Cameron y Trivedi, 1998), por esto al ajustar un modelo ZIP se deben utilizar dos funciones de enlace, una para la parte de exceso de ceros y otra para los conteos corrientes (Ridout *et al*., 1998). De manera análoga ocurre con el ZINB, si bien la media es modificada

respecto a la probabilidad de los conteos sin tener en cuenta el exceso de ceros, la varianza se modifica teniendo presente que la dispersión es mayor que el modelo corriente (Long, 1997; Cameron y Trivedi, 1998; Winkelmann, 2000). Esto queda expuesto en la Figura 17.

Figura 14: Diferencia entre los valores observados y esperados de los cuatro modelos ajustados para el conteo de capturas de *C. pomonella* (L.).



En la Figura 17 se observan los residuales de los modelos Poisson corriente, ZIP, Binomial Negativa corriente y el ZINB. Puede destacarse cómo el modelo de Poisson corriente subestima los valores ceros y sobreestima una vez pasado ese umbral; en el modelo Binomial Negativo de manera análoga se observa el mismo patrón pero más parsimonioso y logra llegar a un ajuste de forma más veloz a partir de los valores pequeños; los modelos que contemplan la superdispersión (*zero inflated*) son mejores desde el punto de vista residual, pero si bien el ZIP mejora el ajuste, aún continúa subestimando los valores bajos y sobrestimando los valores altos. Mientras que el modelo ZINB muestra un ajuste

casi teórico, ya que los valores observados y esperados se mantienen similares y la diferencia entre ellos es nula.

Puede observarse a partir de lo presentado en la parte descriptiva y en el proceso de modelación realizado en el capítulo VIII que:

Con respecto a las trampas, las capturas de *C. pomonella (L.)*son mayores en la Trampa 4 (siendo esta el combo entre codlemone y éster de pera) mostrando una mayor eficiencia. Al menos el 50% de las observaciones fueron nulas, proporcionando la aplicación de modelos que contemplen la superdispersión.

En referencia a las especies, las capturas en manzanas fueron más elevadas que en pera, confirmando la mayor susceptibilidad de esta especie. En ambas especies el 25% de las observaciones fueron ceros al menos, en peras llegaba esta proporción hasta el 50%.

En cuanto a las generaciones, puede destacarse que en la generación 1 es el momento de ocurrencia de mayor captura, mientras que en la generación 2 existe mayor variabilidad en las capturas. La generación 3 es la que va mermando, dado el comportamiento biológico de la plaga (Capítulo VII).

Cuando las estimaciones se realizan por generación, se destacó (Figuras 13 y 14) que a medida que avanzan las generaciones, más concentración de ceros había, indicando que sería conveniente introducir la generación como la variable que presenta un efecto de superdispersión o de inflación (*inflated*) en el modelo.

En todos los modelos tanto los corrientes como los *inflated*, pudo observarse que la Trampa 4 es la de mayor captura respecto al testigo comercial, el referente en el modelo, la Trampa 2 (L2) quedando constatado a través del OR. La chance de captura va variando según el modelo propuesto, pero oscila en el valor 2, es decir que la Trampa 4 tiene la chance de capturar al menos dos veces más que el resto.

En términos generales, los modelos con interacción no mejoran de manera eficaz respecto a los modelos de efectos principales, puede destacarse que en el primer modelo propuesto (Tabla 8) si bien todos los efectos son significativos, el AIC y el BIC con su alto valor, indican que aun hay una parte de variabilidad que no puede ser capturada por el modelo. Por esto los modelos siguientes responden de modo superior, al mejorar los criterios de información y al reforzar que el comportamiento de la varianza no es *pari passu* el de la media, sino que es ampliamente mayor. Más aún, en situaciones de superdispersión, la distribución Binomial Negativa corrige la probabilidad asociada a valores bajos de conteos que, habitualmente presentan un ajuste deficiente a través del modelo de regresión Poisson.

El modelo Binomial Negativo con superdispersión (Tabla 20) fue seleccionado como el apropiado para describir el comportamiento de capturas de *C. pomonella* (L.), debido a que mejora la estimación pero a su vez se pierde menos grados de libertad.

CONCLUSIONES

En el análisis de poblaciones que presentan un patrón de distribución agregado es idóneo analizar las posibles variables a través de modelos lineales generalizados, que contemplen superdispersión a fin de mejorar la calidad de modelación del comportamiento de los individuos de dicha población. La importancia radica en la implementación de estrategias de manejo y de control más adecuadas para la reducción de costos económicos y naturales.

En el caso de *Cydia pomonella (L.)* los modelos propuestos que contemplan el exceso de ceros mejoran las estimaciones y ofrecen la alternativa de resolver el problema mostrando ser más convenientes que los utilizados de manera corriente, al permitir captar la variabilidad generada por el patrón de comportamiento en si mismo y a su vez, por el exceso de ceros que resulta del tipo de muestreo utilizado.

Si bien se logra una mejora substancial en la explicación y comprensión del fenómeno, aún queda por profundizar mediante otros modelos, como son los de verosimilitud restringida, o modelos lineales generalizados mixtos, entre otros estudios que excedían este trabajo.

BIBLIOGRAFÍA

Agresti, A. 1996. An Introduction to Categorical Data Analysis. A Wiley-Interscience Publication; Ed. John Wiley y Sons, INC., Florida.

Alamatsaz, M; Abbasi, S. 2008. Ordering Comparison of Negative Binomial Random Variables with their Mixtures. Statistical and ProbailityLetters, vol. 78 p. 2234-2239.

Arboccó, G. 2005. La Fecundidad y su Relación con Variables Socioeconómicas, Demográficas y Educativas Aplicando el Modelo de Regresión Poisson. Universidad Nacional de San Marcos, cap.III. Perú.

Arnau, J; Balluerka, N. 2004. Análisis de datos longitudinales y de curvas de crecimiento. Enfoque clásico y propuestas actuales". Psicothema, vol. 16, (1), p.156-162. Barcelona.

Breslow, N. 1990. Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-Likelihood Models. Journal of the American StatisticalAssociation, vol. 85, (410), p. 565-571.

Cadahia, D. 1977. Repartición especial de las poblaciones en Entomología aplicada. Bol. Servicio de Plagas, vol. 3 p. 219-233, Madrid.

Caffo, B; An, M-W; Rohde, C. 2007. Flexible Random Intercept Models for Binary Outcomes Using Mixtures of Normals. ComputationalStatistics& Data Analysis, vol. 51 p. 5220–5235.

Camargo Mendes, C. 2007. Modelos Para Dados de Contagem com Aplicações. Dissertaçao; Universidade Estadual de Campinas; SP.

Cameron, A.C; Trivedi, P. 1985. RegressionBased Test for Overdispersion. TechnicalReport Nº 9, p. 1-40. Stanford University, California.

Cameron, A. C. 2009. Advances in Count Data Regression. 28th Annual Workshop in Applied Statistics Southern California Chapter of The American Statistical Association Held at University of California, Los Angeles.

Capriglioni, C. 2005. Estadística: Tomo I y Tomo II. 3C Editores, Buenos Aires, Argentina.

Capuccino, N; Priece, P. 1995. Population Dinamics. New Approches and Synthesis. Ed. AcademicPress, California.

Chiang, A. 1987. Métodos Fundamentales de Economía Matemática. Universidad de Connecticut; 3ª edición española. Ed. Mc.Graw Hill, México.

Cichón, L; Fernández, D; Raffo, D. 1999. Carpocapsa, La Plaga Clave. Manzanos y Perales del Valle". idia XXI, p. 96-99.

Cordeiro, G. 1983. Improved Likelihood Ratio Statistics for Generalized Linear Models. Journal of the Royal Statistical Society, Series B. vol. 45 (3), p. 404-413.

Cordeiro, G; McCullagh, P. 1991. Bias Correction in Generalized Linear Models. J.R. Statistical Society Series B, vol. 53 (3), p. 629-643.

Cordeiro, G; Ferrari, S; Paula, G. 1993. Improved Score Tests for Generalized Linear Models. J.R. Stastistics Society Series B; vol. 55 (3), p. 661-674.

Cordeiro, G; Demetrio, C. 2007. Modelos Lineales Generalizados. Departamento de Estatística e Informática, Recife. PE.

Cox, C. 1984. Generalized Linear Models – The Missing Link. Applied Statistics; vol. 33 (1) p. 18-24.

Cox, D. R. 1983. Some Remarks on Overdispersion. Biometrika, vol. 70 (1), p. 269-274.

Chou, Y. 1977. Análisis Estadístico. 2 edición en Español. Ed. Interamericana, México.

Czado, C; Erhardt, V; Min, A; Wagner, F. 2007. Zero-inflated Generalized Poisson Models with Regression Effects on the Mean, Dispersion and Zero-Inflation Level Applied to Patent outsourcing Rates. Statistical Modelling, vol. 7 (2), p. 125-153.

Dean, C.B. 1992. Testing for Overdispersion in Poisson and Binomial Regression Models. Journal of the American Stastical Association, vol. 87 (418), p.451-457.

Díaz, P; Demetrio, C. 1999. Introducción a los Modelos Lineales Generalizados. Su Aplicación en las Ciencias Biológicas, con Ejemplos en GLIM. Screen Ed.

Dobson, A. 2002. An introduction to Generalized Linear Models. 2nd. ed. Chapman & Hall/CRC texts in statistical science series.

Dyke G, Patterson, H.D. 1952. Analysis Of Factorial Arrangements When The Data Are Proportions. Biometrics, Vol.8, (1), p. 1-12.

Famoye, F; Singh, K. 2006. Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. Journal of Data Science, (4), p.117-130.

Fernández, D. 2007. Evaluación de la Emergencia y las Capturas de Adultos de *Cydiapomonella* (L.) (Lepidoptera: Tortrícidae) en Trampas con Cebos que Contienen Feromonas y/o Cairomonas. Universidad de Lleida, Departamento de Producción Vegetal y Ciencia Forestal.

Fernández, D; Cichón, L; Ribes-Dasi, M. y Avilla, J. 2010.: "Comparison of lures loaded with codlemone and pear ester for capturing codling moth, [Cydiapomonella (L.)], in apple and pear orchards under mating disruption". J. Ins. Sc. p.10:139.

Flowerdew, R; Aitkin, M. 1982. A Method of Fitting the Gravity Model Base on the Poisson Distribution. Journal of Regional Science, vol. 22 (2), p. 191-202.

Frome, E. L. 1983. The Analysis of Rates Using Poisson Regression Models. Biometrics, vol. 39, (3) p. 665-674.

Fujii, T. 2008. On Weak Convergences of the Likelihood Ratio Process in Multi-phase Regression Models. StatisticsandProbailityLetters, vol. 78 p. 2066-2074.

Fumes, G. 2009. Uso de Modelos Inflacionados de Ceros na Análise de Questionários de Frequência Alimentar. Universidad EstadualPaulista, Brasil.

Gareth, M, J. 2002. Generalized Linear Models with Functional Predictors. Journal of the Royal Statistical Society. Series B. Vol. 3, p. 411-432.

Gesell Gamboa, F; Laudien M. 2006. Modelo Logit y OddsRattio: Inferencia bajo Enfoques Clásico y Bayesiano. Universidad del Bío-Bío, Chile; En VII Jornadas Internacionales de Estadística.

Gilberto, P. 2004. Modelos de Regressão com apoio computacional. Instituto de Matemática e Estatística. Universidade de São Paulo, Brasil.

Gittins, C; Menni, M. F; Busso, C. 2009. Modelado de laActividad Metabólica de lasYemas Axilares de Macollas Progenitoras a través de Multinomial Lineal Generalizada. XIV ReuniónCientífica del GAB; Trelew Argentina.

Greene, W. 1994. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regressions Models. New York University.

Gourieroux, C; Monfort, A; Trognon, A. 1984. Pseudo Maximum Likelihood Methods: Theroy. Econometrica, vol. 52, (3), p. 681-700.

Gourieroux, C; Monfort, A; Trognon, A. 1984. Pseudo Maximum Likelihood Methods: Applications to Poisson. Econometrica, vol. 52, (3), p. 701-720.

Gujarati, D. 2004. Econometría. 4ta Edición, Mc. Graw-Hill Interamericana. México.

Györfi, L; Kohler, M; Krzyzak, A; Walk, H. 2002. A Distribution-Free Theroy of Nonparametric Regression. Ed. Springer-Verlag Series in Statistics.

Hall, D. 2000. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. Biometrics, vol. 56 (4), p. 1030-1039.

Hawkin, S. 1988. A Brief History of Time: From the Big Bang to Black. Ed. Grijalbo.

Hilbe, J. M. 2008. Negative Binomial Regression. Cambridge University Press. 1º Ed 2007. U.K.

Hinde, J; Demetrio, C. 2005. Overdispersion: Models and Estimation. MSOR Department, Laver Building, University of Exeter, UK.

Jiang, J. 2007. Linear and Generalized Linear Mixed Models and their Applications. Ed. Springer.

Kokonendji, C; Demetrio, C; Zoochi, S. 2006. On Hinde-Demetrio regression models for overdispersed count data. Statistical Methodology, vol. 4 p. 277-291.

Lambert, D; Roeder, K. 1995. Overdispersion Diagnostics for Generalized Linear Models. Journal of the American Statistical Association. Vol. 90 (432), p. 1225-1236.

Larson, R; Hostetler, R; Edwards, B. 1995. Cálculo y Geometría Analítica, volumen 1 y 2. Ed. McGraw Hill, quinta edición.

Lee, Y; Nelder, J.A. 2000. Two Ways of Modelling Overdispersion in Non-Normal Data. Journal of the Royal Statistical Society. Series C (Applied Statistics) Vol. 49 (4), p. 591-598.

Lee, Y; Nelder J.A y Pawitan, Y. 2006. Generalized Linear Models with Random Effects. Unified Analysis via H-likelihood. Ed. ChapmanyHall/CRC, New York.

Lee, A; Wang, K y otros. 2005. Modelling bivariate count series with escess zeros. Mathematical Biosciences 196 p.226-237. Ed. Elsevier Inc.

Lindsey, J. 1997. Applying Generalized Linear Models. Ed. Springer-Verlag; New York.

Little, R y otros. 2006. SAS® for Mixed Models, Second Edition. SAS Institute Inc. NC, USA.

Long, J. S. 1997. Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences Series. SAGE Publications, Inc.

Llorens, N. 2005. Evaluación en el Modelado de Respuestas de Recuento. Facultad de Psicología, Universidad de las Islas Baleares; Palma de Mallorca.

Llorens, N; Perelló del Río, M; Palmer, P. 2004. Las Estrategias de Afrontamiento: Factores de Protección en el Consumo de Alcohol, Tabaco y Cannabis. Adicciones; vol. 16 (4) p. 261-266.

Llorens, N; Perelló del Río, M; Palmer, P. 2005. Activity Levels and Drug Use in a Sample of Spanish Adolescents. AddictiveBehaviors, vol. 30 (8) p. 1597-1602.

Llorens, N; Perelló del Río, M; Palmer, P. 2005. Características de Personalidad en Adolescentes como Predictores de la Conducta de Consumo de Sustancias Psicoactivas. Trastornos Adictivos, vol. 7 (2) p. 90-96.

Martin, T; y otros. 2005. Zero Tolerance Ecology: Improving ecological Inference by Modelling the source of zero observations. Ecology Letters, Ed. Blackwell Publishing Ltd/CNRS, 8:1235-1246.

Mc.Ardle, B; Anderson, M. 2004. Variance Heterogeneity, Transformations, and Models of Species Abundance: a Cautionary Tale. NRC Research Press, vol. 61 p. 1294-1302, Canadá.

McCullagh, P. 1983. Quasi-Likelihood Functions. The Annals of Statistics, vol. 11, (1), p. 59-67.

McCullagh, P; Nelder, J.A. 1989. Generalizad Linear Models. 2nd Ed. Chapman and Hall, Chicago.

Moffatt, P. 1997. Exploiting a Matrix Identity in the Computation of the Efficient Score test for Overdispersion in the Poisson Regression Model. Statistics and Probability Letters, vol. 32 p. 75-79.

Molenberghs, G; Verbeke, G. 2005. Models for Discrete Longitudinal Data. 1º Ed. Springer.

Montaldo, H. 2002. Comparación de Tres Métodos de Análisis de Variables Binarias. Acta Universitaria, vol. 12 (1). México.

Montero-Mercadé, L. 2004. Models Lineals Generalitzats. Facultat de Matemàtiques i Estadística. Versió 2.2, Cataluña.

Morel, J; Neerchal, N. 2008. Rattio Estimation via Poisson Regression and Generalizad Estimating Equation. Statistics and ProbabilityLetters, vol. 78 p. 2188-2193.

Navarro, A; Utzet, F y otros. 2001. La distribución binomial negativa frente a la de Poisson en el análisis de fenómenos recurrentes. Gacetilla Sanitaria, 15 (5) p. 447-452, Barcelona.

Nayak, T; Bose, S y Kundu, S. 2008. On Inconsistency of Estimators of Parameters of Non-Homogenous Poisson Process Models for Software Reliability. Statistical and Probability Letters. vol. 78 p. 2217-2221.

Naylor, T; Balintfy, J; Burdick, D; Chu, K. 1966. Computer Simulation Techniques. Wiley & Son.

Nelder, J. A; Wedderburn, W. M. 1972. Generalized Linear Models. Journal Royal Statsitical Applied; vol. 135 p. 370-385 [part 3].

Nyquist, H. 1991. Restricted Estimation of Generalized Linear Models. Journal of the Royal Statistical Society. Series C, vol. 40, (1); p. 133-141.

Orbe, J. 2001. Lifetime Data Análisis Using a Semiparametric Generalized Linear Model. QÜESTIIÓ, vol.25 (2) p.337-363, Bilbao.

Owens, M; Tan, F; Berger, M. 2001. Local Influence to Detect Influential Data Structures for Generalizad Linear Mixed Models. Biometrics, vol. 57 (4), p. 1166-1172.

Palmer, A y otros. 2005. Modelado del Número de Días de Consumo de Cannabis. Psicothema Journal, vol. 17 (4) p. 559-574, Palma de Mallorca.

Pardoe, I; Durham, C. 2003. Model Choice Applied to Consumer Preferences. American Statistical Association; Joint Statistical Meetings.

Perry, R; Moens, M. 2006. Plant Nematology. Ed. CABI, UK.

Pinto, E.D; Ponce de Leon A.C. 2006. Modelagem Conjunta da Média e Dispersão de Nelder e Lee Como Alternativa aos Mètodos de Taguchi. Pesquisa Operacional, vol. 26 (2).

Potts, J; Elith, J. 2006. Comparing Species Abundance Models. Science Direct, EcologicalModelling; vol.199 p. 153-163. Ed. Elservier, Australia.

Pregibon, D. 1980. Goodness of Link Tests for Generalized Linear Models. Applied Statistics, vol. 29 (1) p. 15-24.

Pron, J. 2007. Análisis del Desempeño Universitario Utilizando Modelos para Variables Enteras. Tesis de maestría en Economía; Universidad de la Plata, director Porto Alberto.

Rabe-Hesketh, S; Skrondal, A; Pickles, A. 2004. GLLAMM Manual (Generalized Linear Latentand Mixed Models). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160, California.

Rabe-Hesketh, S; Skrondal, A; Pickles, A. 2005. Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects. Journal of Econometrics, vol. 128, p. 301–323.

Ramírez González, A. 1999. Ecología Aplicada. Diseño y Análisis Estadístico, Colección de Estudios de Ecología, Universidad de Bogotá, Ed. Jorge Tadeo Lozano, Colombia.

Raventós, J; Segarra, J. G; Acevedo, M.F. 2005. Modelos de Metapoblaciones y de la Dinámica Espacio-Temporal de Comunidades. Publicaciones de la Universidad de Alicante, España.

Ridout, M; Hinde, J; Demetrio, C. 1991. A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. Biometrics 57, p. 219-223.

Ridout, M; Demetrio, C; Hinde, J. 1998. Models for Count Data with Many Zeros. International Biometric Conference; Cape Town.

Robertson, J.L; Russell, R.M; Preisler, H.K; Savin, N.E. 2007. Bioassays with Arthropods. 2nd Ed. CRC Press Taylor & Francis Group; Boca Raton.

Ruiz Cárdenas, R; Assunçao, R; Demetrio, G. 2009. Spatio-Temporal Modelling of Coffee Berry Borer Infestation Patterns Accounting for Inflation of Zeroes and Missing Values. Sci. Agric. vol. 66 (1) p. 100-109, Piracicaba, Brazil.

Ruiz Maya Pérez, L; Pliego, J. 1999. Fundamentos de Inferencia Estadística. Editorial AC.

Sackman, P; Rabinovich, M; Corley, J.C. 2001. Successful Removal of German Yellowjackets (Hymenoptera: Vespidae) by Toxic Baiting. Journal of Economic Entomology, vol. 94 (4) p.811-816, Bariloche, Argentina.

Sayyad, G. 1973. Bayesian and Classical Analysis of Poisson Regression. Journal of the Royal Statistical Society. Series B, vol. 35 (3); p. 445-451.

Singh, K; Famoye, F. 1993. Analysis of Rates Using a Generalized Poisson Regression Model. Biometrical Journal, vol. 35 (8); p. 917-923.

Slymen, D; Ayala, G; Arredondo, E; Elder, J. 2006. A Demonstration of Modeling Count Data with an Application to Physical Activity. Epidemiologic Perspectives and Innovations, vol. 3 (3) p. 1-9.

Southwood, T.R.E. 1978. Ecological Methods with particular reference to the study of Insect Population. 2nd Ed. ChapmanyHall, New York.

Speight, M; Hunter, M; Watt, A. 1999. Ecology of Insects. Concept and Aplications. Ed. Blackwell Science, UK.

Statistic Data Analysis Special Edition (STATA® SE) versión 11.0. StataCorp, Texas, USA. http://www.stata.com. 1999.

Verbeke, G; Molenberghs, G. 2000. Linear Mixed Model for Longitudinal Data. Ed. Springer.

Veronesi, A. 2001. Modelo Binomial con Superdispersión. Causas, Detección y Modelado. Tesis de Maestría. Universidad Nacional de Córdoba, Argentina.

Vives Brosa, J. 2002. El Diagnóstico de la Sobredispersión en Modelos de Análisis de Recuento. Facultad de Psicología, Universidad Autónoma de Barcelona, España.

Vuong, Q. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. The Econometric Society; Econometrica Journal, vol. 57 (2), p. 307-333.

Weems, K; Smith, P. 2004. On Robustness of Maximum Likelihood Estimates for Poisson-lognormal Models. Statistics and Probability Letters, vol. 66 p. 189-196.

Winkelmann, R; Zimmermann, K. 1995. Recent Developments in Count Data Modelling: Theory and Application. Journal of Economic Surveys, vol. 9 (1), p.1-24.

Winkelmann, R. 2000. Seemingly Unrelated Negative Binomial Regression. Oxford Bulletin of Economics and Statistics, vol. 62 (4), p. 553-560.

Yau, K; Kuk, A. 2002. Robust Estimation in Generalized Linear Mixed Models. Journal of the Royal Statistical Society. Series B, vol. 64 (1), p. 101-117.

Zeger, S. 1988. A Regression Model for Time Series of Counts. Biometrika, vol. 75 (4) p. 621-629.