UNIVERSIDAD VERACRUZANA

FACULTAD DE ESTADÍSTICA E INFORMÁTICA

Un Modelo de Regresión Poisson Inflado con Ceros para Analizar datos de un Experimento de Fungicidas en Jitomate

Tesis

Que como requisito parcial para obtener el grado de

MAESTRO EN ESTADÍSTICA

Presenta:

María de Lourdes Velasco Vázquez

TUTOR

Dr. Sergio Francisco Juárez Cerrillo

Xalapa, Enríquez, Ver. Enero de 2008

RESUMEN

Los datos de conteo generalmente se modelan con los modelos de regresión Poisson. Sin embargo, cuando los datos presentan una alta frecuencia de ceros, el ajuste es pobre. Para analizar datos de conteo con exceso de ceros, Lambert (1992) propuso el *modelo regresión Poisson inflado con ceros*. En este modelo la variable respuesta se distribuye como una mezcla de una distribución que contiene unos y ceros, con probabilidad p; y una distribución Poisson con parámetro λ, con probabilidad $1 - p$. Ambos parámetros λ y p dependen de covariables a través de un modelo lineal generalizado. En esta tesis analizamos los datos generados por un experimento de pruebas de fungicidas para el control de daños en plantas de jitomate. El experimento se realizó bajo condiciones de invernadero bajo un diseño anidado con tres tratamientos (fungicidas) y un testigo (no se aplicó fungicida). Las variables respuesta son el número de daños que presentaron las plantas de jitomate en foliolos, hojas, racimos y frutos. El objetivo del experimento es determinar al fungicida que más reduce el número de daños en las plantas. Debido a que los datos presentaban exceso de ceros, utilizamos el modelo regresión Poisson inflado con ceros de Lambert para cada variable explicatoria. El análisis nos permitió determinar que los fungicidas efectivamente reducen el número de daños y además pudimos determinar cuál de los fungicidas es el más efectivo. Los modelos se ajustaron con el programa STATA.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

CAPÍTULO 1
INTRODUCCIÓN

Las variables respuesta que son conteos surgen en disciplinas tales como la epidemiología, la sociología, la psicología, la ingeniería, y la agricultura. Cuando la variable respuesta es un conteo los datos observados se deben modelar estadísticamente con distribuciones discretas como por ejemplo la Poisson y la binomial negativa. Sin embargo, no es raro que el número de ceros observados en la variable respuesta exceda a la frecuencia que se espera observar bajo la distribución que se ajusta. En este caso se dice que los datos presentan *exceso de ceros* o están *inflados con ceros*. El exceso de ceros indica que se ha especificado incorrectamente al modelo ya que no se cumplen los supuestos distribucionales de la modelación. Esto no debe ignorarse, no debe ajustarse una distribución que no considera al exceso de ceros pues los valores nominales asociados a las inferencias -niveles de significancia, niveles de confianza- no serán correctos.

El exceso de ceros es común en la práctica, por lo que se han desarrollado modelos estadísticos que describen este fenómeno y, por lo tanto, permiten derivar conclusiones realistas y confiables a partir de las inferencias. Esta tesis está

motivada por un problema en agronomía en el cual las variables respuesta que se observan son conteos que presentan exceso de ceros.

El tema central de este trabajo es la aplicación del modelo de regresión Poisson a un conjunto de datos experimentales que presentan exceso de ceros. En este capítulo presentamos algunos antecedentes de este tema. El resto del capítulo está organizado del siguiente modo. En la Sección 1.1 caracterizamos a los ceros de acuerdo a su origen. En la sección 1.2 hacemos una revisión de los principales modelos que se han propuesto para manejar datos de conteo con exceso de ceros. Por último, en la sección 1.3 describimos la estructura de la tesis.

1.1 Naturaleza del Exceso de Ceros

El exceso de ceros puede ocurrir con datos de variables continuas o discretas. En esta tesis solo trataremos los ceros en datos discretos. Como ya se mencionó, el inflamiento con ceros ocurre cuando hay una gran frecuencia de observaciones iguales a cero de tal manera que ninguna de las distribuciones discretas estándar proporciona un ajuste adecuado. Para modelar el exceso de ceros es fundamental entender la naturaleza del origen de los ceros. De acuerdo a lo anterior, los ceros se clasifican en dos tipos: *ceros estructurales* y *ceros muestrales*. Expliquemos los tipos de ceros con un par de ejemplos. Supongamos que se observa el número de lesiones en plantas. Una planta podría no tener lesiones por que es resistente a la

enfermedad que ocasiona la lesión. Para esta planta siempre observaremos cero lesiones. En este caso tenemos un cero estructural. Por otro lado, supongamos que la planta sí es vulnerable a la enfermedad pero al momento de observarla no está enferma y por lo tanto no presenta lesiones y observamos un cero. Sin embargo se podría observar un dato diferente de cero en una medición posterior. En este caso tenemos un cero muestral. Así, un cero inevitable es un cero estructural, mientras que un cero que ocurre debido al mecanismo de muestreo es un cero muestral. Consideremos otro ejemplo. En datos de abundancia de especies los ceros estructurales pueden ocurrir porque cierta especie está ausente en la zona de observación. Pero si la especie está presente en la zona bajo estudio y el observador simplemente no la detecta durante el muestreo, entonces se tiene un cero muestral.

1.2 Modelos para Conteos Inflados con Ceros

Para desarrollar modelos estadísticos que modelen adecuadamente el exceso de ceros es importante entender los mecanismos que originan a los ceros. En este sentido, se debe determinar si se trata de ceros estructurales o muestrales o, más aun, una combinación de ambos tipos. Es así que se han desarrollado modelos para datos con exceso de ceros que consideran diversos escenarios. Estos modelos describen ceros estructurales (Welsh *et al*. 1996, 2000, Barry and Welsh 2002, Podlich *et al*. 2002,) y ceros muestrales (Mackenzie *et al*. 2002). Si no se entiende

el origen de los ceros y se modelan incorrectamente, se corre el riesgo de estimar deficientemente los parámetros del modelo (Lambert 1992; Mackenzie *et al*. 2002). Así pues, recalcamos la importancia de entender el origen de los ceros en los datos bajo estudio, especialmente en el proceso de selección o desarrollo de un modelo estadístico. A continuación presentamos una breve descripción de los procedimientos de modelación más comunes para cuando los datos presentan ceros estructurales, ceros muestrales y una combinación de ambos.

1.2.1 Modelos Lineales Generalizados

Supongamos que la variable respuesta Y es una variable aleatoria de conteo y X es un vector de covariables. El enfoque más simple consiste en modelar la relación entre Y y X a través un modelo de regresión aditivo $Y = g(X;\theta) + \varepsilon$, donde ε es un componente aleatorio de error y g es una función conocida excepto por el vector de parámetros θ el cual se estima generalmente por mínimos cuadrados ordinarios o máxima verosimilitud. Si g es una función lineal en el vector de parámetros θ, entonces tenemos al modelo lineal general clásico. De otro modo, tenemos un modelo de regresión no lineal y debemos usar la metodología de estos modelos. Debido a la naturaleza discreta de la variable respuesta Y, generalmente el modelo se ajusta después de transformar a Y. Las transformaciones más comunes para variables discretas son $\sqrt{Y}$ y $\log Y$. Sin embargo, estas

transformaciones no son de gran utilidad para tratar el exceso de ceros. La raíz cuadrada no elimina a los ceros y el logaritmo ni tan siquiera está definido para los ceros. Adicionalmente, los procedimientos de inferencia en estos modelos están basados en los supuestos de normalidad de Y así como de homogeneidad de varianza de Y. En el caso de que Y sea un conteo estos supuestos obviamente no son válidos.

Un mejor enfoque de modelación consiste en ajustar un modelo lineal generalizado en el cual la variable respuesta Y se modela con alguna distribución discreta que pertenece a la familia exponencial. Cuando se usa a la distribución Poisson, el modelo lineal generalizado es conocido como el modelo de regresión Poisson. Sin embargo, ante la presencia de exceso de ceros, el ajuste de un modelo de regresión Poisson generalmente será pobre. Por lo que se debe adecuar al modelo para que considere al exceso de ceros. Una exposición amplia sobre la teoría y aplicaciones de los modelos de regresión Poisson sin exceso de ceros y en general de los modelos lineales generalizados, se puede consultar en McCullagh and Nelder (1989).

1.2.2 Modelos en dos Partes

Para analizar datos con exceso de ceros estructurales se han propuesto los modelos de *dos partes*, también conocidos como modelos *condicionales* (Lambert 1992,

Welsh *et al.* 1996, Mullahy 1986, Heilbron 1994). Estos modelos tienen gran aplicación, principalmente en Ecología. El enfoque consiste en primero modelar la presencia/ausencia (conteos diferentes de cero y ceros) con un modelo de regresión logística. Después se condiciona sobre los datos de conteo positivos y se modelan éstos con una distribución discreta con el cero truncado. Si se usa la distribución Poisson, se tiene el modelo Poisson de dos componentes. En estos modelos, todos los ceros se modelan en el componente presencia/ausencia. Un supuesto fundamental de este enfoque de análisis es que los ceros se originan a partir de un mecanismo simple que no afecta a las observaciones diferentes de cero. Una ventaja computacional de este enfoque es que es posible ajustar estos modelos en dos etapas. Primero se ajustan los ceros y los no ceros con el modelo de regresión logística y posteriormente se ajustan los conteos positivos usando la distribución Poisson con el cero truncado. De este modo, la log-verosimilitud del modelo es la suma de la log-verosimilitud de cada componente. Estos modelos son fáciles de ajustar e interpretar.

Formalmente, sea Y_{ij} la observación *j* en la variable respuesta asociada al sujeto *i*. Sean x_{ij} y z_{ij} vectores de covariables, las covariables no necesariamente son diferentes. Supongamos que $Y_{ij} = 0$ con probabilidad $1 - p(x_{ij})$ y $Y_{ij} > 0$ con probabilidad $p(x_{ij})$. Cuando Y_{ij} es mayor que cero, suponemos que tiene una

distribución Poisson con el cero truncado y parámetro $\lambda(z_{ij})$. El modelo tiene la siguiente forma:

$$P(Y_{ij} = 0 \mid x_{ij}, z_{ij}) = 1 - p(x_{ij})$$

$$P(Y_{ij} = y_{ij} \mid x_{ij}, z_{ij}) = \frac{p(x_{ij}) \exp(-\lambda(z_{ij})) \lambda(z_{ij})^{y_{ij}}}{y_{ij}!\{1 - \exp[-\lambda(z_{ij})]\}}, \quad y_{ij} = 1,2,\ldots$$

En este modelo tenemos que para la observación *j* en el sujeto *i*, la probabilidad de observar al menos una ocurrencia en la variable respuesta es $p(x_{ij})$ y $\lambda(z_{ij})$ es el parámetro de la distribución Poisson (con el cero truncado) que describe el número de ocurrencias en la variable respuesta. Para el componente presencia/ausencia, se postula el modelo de regresión logística

$$h_1(p(x_{ij})) = \text{logit}(p(x_{ij})) = \log\left(\frac{p(x_{ij})}{1 - p(x_{ij})}\right) = x_{ij}^{\mathrm{T}}\beta,$$

y para el componente de conteos positivos tenemos

$$h_2(\lambda(z_{ij})) = \log(z_{ij}) = z_{ij}^{\mathrm{T}}\gamma,$$

donde β y γ son parámetros desconocidos y, como se puede ver, h_1 y h_2 son las funciones liga logit y logarítmica, respectivamente. La log-verosimilitud del componente presencia/ausencia es

$$l(\beta) = \sum_{y_{ij}=0} \log\left(\frac{1}{1+\exp(x_{ij}^{\mathrm{T}}\beta)}\right) + \sum_{y_{ij}>0} \log\left(\frac{\exp(x_{ij}^{\mathrm{T}}\beta)}{1+\exp(x_{ij}^{\mathrm{T}}\beta)}\right),$$

y la log-verosimilitud para los conteos positivos es la distribución Poisson con el cero truncado

$$l(\gamma) = \sum_{y_{ij}>0} (y_{ij} z_{ij}^{\mathrm{T}}\gamma - \exp(z_{ij}^{\mathrm{T}}\gamma) - \log(1-\exp(-\exp(z_{ij}))) - \log y_{ij}!)\,.$$

Si las observaciones y_{ij} son independientes, entonces la parametrización en términos de β y γ es ortogonal y entonces la log-verosimilitud completa del modelo es la suma de las log-verosimilitudes de los componentes presencia/ausencia (modelo de regresión logística) y conteos positivos, $l(\beta)$ y $l(\gamma)$. Para más detalles de este modelo véase Dobbie, M. J. & Welsh, A. H. (2001).

1.2.3 Modelos de Mezclas

Los modelos de mezclas de distribuciones son combinaciones lineales convexas de distribuciones de probabilidad. Los ceros de estos modelos pueden ser una mezcla de ceros estructurales y muestrales. De esta clase de modelos, el llamado modelo de regresión *Poisson Inflado con Ceros* (PIC) es el más usado para datos de conteo con exceso de ceros. Este modelo fue propuesto por Lambert (1992). En estos modelos los ceros se dividen en dos grupos, uno tiene los ceros provenientes de la distribución que genera a la variable respuesta, el otro grupo tiene a los ceros "extra". Los ceros del primer grupo se modelan con la distribución Poisson. Un cero en este grupo ocurre con probabilidad $1-p$. Los ceros extra ocurren con probabilidad p. Lambert (1992), Welsh *et al.* (1996), y Böhning *et al.* (1999) presentan aplicaciones del modelo PIC.

El modelo PIC está dada por

$$P(Y=y)=\begin{cases} p+(1-p)\exp(-\lambda) & y=0, \\ (1-p)\dfrac{e^{-\lambda}\lambda^{y}}{y!} & y=1,2,\ldots \end{cases}$$

Otra distribución que se ha propuesto bajo este enfoque es la distribución binomial negativa (Welsh *et al*. 2000). Esta distribución es particularmente adecuada para

cuando además del exceso de ceros también se presenta sobredispersión. La recomendación es que si el exceso de ceros consiste de ceros muestrales, se debe usar un modelo de mezcla de distribuciones (Mackenzie *et al.* 2002). Aunque la interpretación de estos modelos no es tan sencilla como la de los modelos de dos partes.

Hasta donde es de nuestro conocimiento, aún no se dispone de una discusión formal en la literatura acerca de cómo modelar datos con ceros estructurales y de muestreo. Este es, al parecer, un problema abierto en modelación estadística. Es posible que un enfoque de inferencia Bayesiano, donde el modelo incorpore información acerca de los ceros muestrales como información a priori, pudiera dar respuestas de utilidad. Una revisión de otros métodos para modelar conteos inflados con ceros está disponible en Ridout *et al.* (1998).

1.3 Organización de la Tesis

El resto de esta tesis está organizado del siguiente modo. En el Capítulo 2 se presenta con detalle al modelo de regresión PIC propuesto por Lambert (1992). Se presentan los procedimientos de inferencia basados en la verosimilitud así como las propiedades asintóticas de los estimadores. En nuestra presentación seguimos de cerca a Lambert (1992). En el Capítulo 3 se describe el experimento de prueba de fungicidas para el control de daños en plantas de jitomate que motivó esta tesis.

Se presenta un análisis inicial de los datos junto con los resultados del proceso de modelación con el modelo de regresión PIC. Finalmente, discutimos algunos de los problemas con los que nos hemos encontrado y que creemos merecen investigación adicional. El ajuste de los modelos PIC lo realizamos con el programa STATA. Las salidas del análisis están en el Apéndice.

CAPÍTULO 2
MODELO DE REGRESIÓN POISSON INFLADO CON CEROS

Los modelos Poisson y binomial negativa para datos inflados con ceros se han ajustado a datos sin considerar covariables (Jonhson, Kotz and Kemp 1969). Lambert (1992) extiende estos modelos a la forma general de modelos de regresión que permiten la consideración de covariables. Como se mencionó en el capítulo anterior, el modelo de regresión PIC se usa para modelar datos de conteo como una mezcla de ceros y una distribución Poisson. En este modelo se considera que los ceros pueden provenir de dos procesos, el proceso generador de la variable respuesta y otro proceso que genera a los ceros extra. Además las covariables se relacionan con el proceso que genera los ceros extra así como con el proceso Poisson que genera a los conteos incluyendo a una proporción de los ceros observados (Lambert, 1992). En el resto del capítulo presentamos con detalle el modelo de regresión PIC.

2.1 El Modelo de Regresión PIC

En los modelos de regresión para datos inflados con ceros se mezcla una distribución degenerada en el cero con una distribución discreta no degenerada. La estructura de la regresión se construye a través de la media de la distribución no

degenerada y, posiblemente, a través de mezclar las probabilidades. A continuación presentamos el modelo de regresión Poisson inflado con ceros desarrollado en Lambert (1992).

Sea $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ el vector de la variable respuesta. El modelo de regresión PIC supone qua las Y_i son independientes con la siguiente distribución

$$Y_i = 0 \text{ con probabilidad } p_i,$$

$$Y_i \sim \text{Poisson}(\lambda_i) \text{ con probabilidad } 1 - p_i.$$

De tal manera que la función de masa de probabilidad de Y_i es

$$P(Y_i = y) = \begin{cases} p_i + (1 - p_i)\,\mathrm{e}^{-\lambda_i}, & \text{para } y = 0, \\ (1 - p_i)\dfrac{\mathrm{e}^{-\lambda_i}\,\lambda^y}{y!}, & \text{para } y = 1,2,\ldots \end{cases}$$

Denotamos lo anterior por $Y_i \sim \text{PIC}(p_i, \lambda_i)$. La variable respuesta se modela mediante los parámetros del modelo $\lambda = (\lambda_1, \ldots, \lambda_n)^{\mathrm{T}}$ y $\mathbf{p} = (p_1, \ldots, p_n)^{\mathrm{T}}$. Se supone que estos parámetros dependen de las covariables a través de un modelo lineal generalizado. Es decir, se supone que

$$\log(\lambda) = (\log(\lambda_1),...,\log(\lambda_n))^{\mathrm{T}} = \mathbf{B}\beta\,,$$

$$\mathrm{logit}(\mathbf{p}) = (\mathrm{logit}(p_1),\ldots,\mathrm{logit}(p_n))^{\mathrm{T}} = \mathbf{G}\gamma\,,$$

donde β y γ son parámetros desconocidos, y **B** y **G** son matrices con las covariables.

2.2 Ajuste del Modelo de Regresión PIC

Cuando los parámetros γ y β no están relacionados y las covariables que afectan a la distribución Poisson son las mismas covariables que afectan a las probabilidades p_i, la log-verosimilitud para el modelo de regresión PIC es

$$\begin{aligned} \mathrm{L}(\gamma,\beta;Y) = \sum_{i=1}^{n} \log p_i + \sum_{Y_i=1} \log\{\exp(\mathbf{G}_i^{\mathrm{T}}\gamma) + \exp[-\exp(\mathbf{B}_i^{\mathrm{T}}\beta)]\} \\ + \sum_{Y_i>0} [Y_i \mathbf{B}_i^{\mathrm{T}}\beta - \exp(\mathbf{B}_i^{\mathrm{T}}\beta)] - \sum_{Y_i>0} \log(Y_i!), \end{aligned}$$

donde $\mathbf{G}_i$ y $\mathbf{B}_i$ son las filas *i* de **G** y **B**, respectivamente. Notemos que

$$p_i = \frac{\exp(\mathbf{G}_i^{\mathrm{T}}\gamma)}{1+\exp(\mathbf{G}_i^{\mathrm{T}}\gamma)}\,.$$

Así que la log-verosimilitud del modelo de regresión PIC es

$$\begin{aligned} \mathrm{L}(\gamma,\beta;Y) = & \sum_{i=1}^{n} \mathbf{G}_i^{\mathrm{T}}\gamma - \sum_{i=1}^{n} \log(1+\exp(\mathbf{G}_i^{\mathrm{T}}\gamma)) \\ & + \sum_{Y_i=0} \log\{\exp(-\mathbf{G}_i^{\mathrm{T}}\gamma) + \exp[-\exp(\mathbf{B}_i^{\mathrm{T}}\beta)]\} \\ & + \sum_{Y_i>0} [Y_i \mathbf{B}_i^{\mathrm{T}}\beta - \exp(\mathbf{B}_i^{\mathrm{T}}\beta)] - \sum_{Y_i>0} \log(Y_i!). \end{aligned}$$

Puesto que es difícil maximizar numéricamente la log-verosimilitud debido a que es la suma de las funciones exponenciales, los estimadores de máxima verosimilitud se deben obtener con algún procedimiento numérico como por ejemplo el algoritmo de Newton-Raphson o el algoritmo EM. Generalmente el algoritmo de Newton-Raphson es más rápido que el EM pero el algoritmo EM es más simple de implementar.

2.3 Algoritmo EM

Supongamos que sabemos de donde provienen los ceros, es decir, conocemos cuales ceros provienen del mecanismo que genera los ceros extra y cuales provienen de la distribución Poisson. Definimos la variable Z_i tal que $Z_i = 0$ si la observación Y_i proviene de la Poisson y $Z_i = 1$ si la observación Y_i viene del

proceso generador de los ceros. Si observáramos a Y y Z, la log-verosimilitud completa del modelo de regresión PIC sería

$$\mathrm{L_c}(\gamma,\beta;Y) = \sum_{i=1}^{n}(Z_i\mathbf{G}_i\gamma - \log(1+\mathrm{e}^{\mathbf{G}_i\gamma})) + \sum_{Y_i=1}^{n}(1+Z_i)(y_i\mathbf{B}_i\beta - \mathrm{e}^{\mathbf{B}_i\beta}) - \sum_{i=1}^{n}(1-Z_i)\log(Y_i!).$$

El algoritmo EM involucra dos pasos: el paso E (estimación) y el paso M (maximización). La verosimilitud se maximiza en forma iterativa alternando entre estos dos pasos. El paso E consiste en estimar los valores de Z dada las estimaciones actuales de β y γ. El paso M consiste en estimar los valores de β y γ a través de maximizar la log-verosimilitud usando los valores actuales de Z. Para el modelo regresión PIC el algoritmo consiste de tres pasos en cada iteración, un paso de estimación, un paso de maximización para β y un paso de maximización para γ. Los detalles se muestran a continuación.

1. Paso E

En la iteración k se estima Z_i con la regla de Bayes y el uso de las estimaciones más recientes de β y γ. Denotemos a estas estimaciones como $\beta^{(k)}$ y $\gamma^{(k)}$. Entonces

$$
\begin{aligned}
Z_i^{(k)} &= \Pr(\text{proceso que solo genera ceros} \mid y_i, \beta^{(k)}, \gamma^{(k)}) \\
&= \frac{\Pr(y_i \mid \text{proceso que solo genera ceros}) \Pr(\text{proceso que solo genera ceros})}{\Pr(y_i \mid \text{proceso que solo genera ceros}) \Pr(\text{proceso que solo genera ceros}) + \Pr(y_i \mid \text{Poisson}) \Pr(\text{Poisson})} \\
&= \{1 + e^{-\mathbf{G}_i\gamma^{(k)} - \exp(\mathbf{B}_i\beta^{(k)})}\} \quad \text{si } y_i = 0, \\
&= 0, \qquad\qquad\qquad\qquad\qquad \text{si } y_i = 1,2,\ldots
\end{aligned}
$$

2. Paso M para estimar β

Se maximizar $L_c(\beta; y, Z^{(k)})$ para así obtener la estimación $\beta^{(k+1)}$

$$L_c(\beta; y, Z) = \sum_{i=1}^{n} (1 - Z_i)(y_i \mathbf{B}_i \beta - e^{\mathbf{B}_i\beta}).$$

Esto se puede hacer mediante una regresión log-lineal Poisson ponderada con pesos $1 - Z^{(k)}$, véase McCullagh and Nelder (1989).

3. Paso M para estimar γ

Se maximiza $L_c(\gamma; y, Z^{(k)})$ para así obtener la estimación de $\gamma^{(k+1)}$. Puesto que $Z_i^{(k)} = 0$ siempre que $y_i > 0$, tenemos que

$$L_c(\gamma; y, Z) = \sum_{i=1}^{n} (Z_i \mathbf{G}_i \gamma - \log(1 + e^{\mathbf{G}_i\gamma}))$$

se puede reescribir como

$$L_c(\gamma; y, Z) = \sum_{y_i=0} Z_i^{(k)} \mathbf{G}_i\gamma - \sum_{y_i=0} Z_i^{(k)} \log(1+e^{\mathbf{G}_i\gamma}) - \sum_{i=1}^{n} (1-Z_i)\log(1+e^{\mathbf{G}_i\gamma}).$$

Para maximizar la log-verosimilitud anterior como una función de γ, suponemos que n_0 resultados de las n y_i's son cero. Por ejemplo, suponemos que las observaciones $y_{i1},...,y_{in_0}$ son todas cero. Ahora definimos lo siguiente:

$$\begin{aligned}
\mathbf{y}_*^{\mathrm{T}} &= (y_1,\ldots,y_n; y_{i1},\ldots,y_{in_0}),\\
\mathbf{G}_*^{\mathrm{T}} &= (\mathbf{G}_1^{\mathrm{T}},\ldots,\mathbf{G}_n^{\mathrm{T}}; \mathbf{G}_{i1}^{\mathrm{T}},\ldots,\mathbf{G}_{in_0}^{\mathrm{T}}),\\
\mathbf{P}_*^{\mathrm{T}} &= (p_1,\ldots,p_n; p_{i1},\ldots,p_{in_0}).
\end{aligned}$$

Definimos también una matriz diagonal $\mathbf{W}^{(k)}$ con elementos en la diagonal $w^{(k)} = (1-Z_1^{(k)},\ldots,1-Z_n^{(k)}; Z_{i1}^{(k)},...,Z_{in_0}^{(k)})$. Ahora podemos escribir

$$L(\gamma; y, Z^{(k)}) = \sum_{i=1}^{n+n_0} \mathbf{y}_{*i} w_i^{(k)} \mathbf{G}_{*i}\gamma - \sum_{i=1}^{n+n_0} w_i^{(k)} \log(1+e^{\mathbf{G}_{*i}\gamma}).$$

Con la función score $\mathbf{G}_*^{\mathbf{T}}\mathbf{W}^{(k)}(\mathbf{y}_* - \mathbf{P}_*)$ y la matriz negativa definida $\mathbf{G}_*^{\mathbf{T}}\mathbf{W}^{(k)}\mathbf{Q}_*\mathbf{G}_*$, donde $\mathbf{Q}_*$ es una matriz diagonal con $\mathbf{P}_*(1-\mathbf{P}_*)$ en la diagonal. Estas funciones son las mismas que las de la regresión logística ponderada con

respuesta $\mathbf{y}_*$, matriz de diseño $\mathbf{G}_*$, y pesos $\mathbf{W}^{(k)}$. Entonces $\gamma^{(k)}$ se puede encontrar a través de la regresión logística ponderada.

El algoritmo EM inicia con un valores iniciales para β y γ, e itera a través de los pasos E y M hasta que las estimaciones convergen numéricamente. El valor inicial de β puede ser la estimación obtenida de un análisis de regresión Poisson considerando solamente las respuestas positivas.

2.4 Propiedades Asintóticas

En esta sección presentamos la distribución asintótica de los estimadores de máxima verosimilitud $(\hat{\gamma}, \hat{\beta})$. Definamos las siguientes expresiones

$$r_i = \frac{\exp(\mathbf{G}_i^{\mathrm{T}}\gamma + \lambda_i)}{1 + \exp(\mathbf{G}_i^{\mathrm{T}}\gamma + \lambda_i)} \quad \text{si } y_i = 0,$$
$$r_i = 0 \quad \text{si } y_i = 1,2,\ldots,$$

para $i = 1,2,\ldots,n$. Definimos también $\mathbf{r} = (r_1, r_2, \ldots, r_n)^{\mathrm{T}}$, $\mathbf{s} = 1 - \mathbf{r}$, y $\mathbf{q} = 1 - \mathbf{p}$ donde $\mathbf{p} = (p_1, \ldots, p_n)^{\mathrm{T}}$. Sean $\hat{\mathbf{p}}$, $\hat{\mathbf{q}}$, $\hat{\mathbf{r}}$ y $\hat{\mathbf{s}}$ las cantidades análogas con los verdaderos valores de los parámetros sustituidos por los estimadores de máxima

verosimilitud. La matriz de información observada correspondiente a la log-verosimilitud del modelo de regresión PIC es

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_1 & \mathbf{I}_2 \\ \mathbf{I}_3 & \mathbf{I}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{G}^{\mathrm{T}} & 0 \\ 0 & \mathbf{B}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{D}_{\hat{\gamma},\hat{\gamma}} & \mathbf{D}_{\hat{\gamma},\hat{\beta}} \\ \mathbf{D}_{\hat{\gamma},\hat{\beta}} & \mathbf{D}_{\hat{\beta},\hat{\beta}} \end{bmatrix} \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{B} \end{bmatrix},$$

donde $\mathbf{D}_{\hat{\gamma},\hat{\gamma}}$ es una matriz diagonal con elementos $\hat{\mathbf{q}} - \hat{\mathbf{s}}$, $\mathbf{D}_{\hat{\gamma},\hat{\beta}}$ es una matriz diagonal con elementos $-\hat{\boldsymbol{\lambda}}\hat{\mathbf{s}}$, y $\mathbf{D}_{\hat{\beta},\hat{\beta}}$ es una matriz diagonal con elementos $\hat{\lambda}(1-\hat{\mathbf{r}})(1-\hat{\lambda}\hat{\mathbf{r}})$. La información esperada es

$$\mathbf{i}_{\gamma,\beta} = \begin{bmatrix} \mathbf{G}^{\mathrm{T}} & 0 \\ 0 & \mathbf{B}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{\gamma,\gamma} & \mathbf{d}_{\gamma,\beta} \\ \mathbf{d}_{\gamma,\beta} & \mathbf{d}_{\beta,\beta} \end{bmatrix} \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{B} \end{bmatrix},$$

donde $\mathbf{d}_{\gamma,\gamma}$ es una matriz diagonal con elementos $\mathbf{p(r-p)}$, $\mathbf{d}_{\gamma,\beta}$ es diagonal con elementos $-\lambda\mathbf{p}(1-r)$ donde $\lambda = (\lambda_1,\ldots,\lambda_n)^{\mathrm{T}}$ y $\mathbf{d}_{\beta,\beta}$ es una matriz diagonal con elementos $\lambda(1-\mathbf{p}) - \lambda^2\mathbf{p}(1-\mathbf{r})$. Si $n^{-1}\mathbf{i}_{\gamma,\beta}$ tiene un límite positivo definido cuando $n \to \infty$, entonces

$$n^{1/2}\begin{bmatrix} \hat{\gamma} - \gamma \\ \hat{\beta} - \beta \end{bmatrix}$$

es asintóticamente $\mathrm{N}(0, n\mathbf{i}_{\gamma,\beta}^{-1})$. La matriz de información observada se puede substituir por la matriz de información esperada en la distribución asintótica. Si $(\hat{\gamma}_0, \hat{\beta}_0)$ maximiza a la log-verosimilitud

$$\mathrm{L_c}(\gamma, \beta; Y) = \sum_{i=1}^{n} (Z_i \mathbf{G}_i \gamma - \log(1 + \mathrm{e}^{\mathbf{G}_i \gamma})) + \sum_{Y_i=1}^{n} (1 + Z_i)(y_i \mathbf{B}_i \beta - \mathrm{e}^{\mathbf{B}_i \beta}) - \sum_{i=1}^{n} (1 - Z_i) \log(Y_i!)$$

bajo una hipótesis nula H_0 de dimensión q_0 y $(\hat{\gamma}, \hat{\beta})$ maximiza la log-verosimilitud anterior bajo una hipótesis alternativa H de dimensión $q > q_0$ y anidada en H_0, entonces $D = 2\{L(\hat{\gamma}, \hat{\beta}) - L(\hat{\gamma}_0, \hat{\beta}_0)\}$ tiene una distribución asintótica ji-cuadrada con $q - q_0$ grados de libertad. Por lo tanto, dos veces la diferencia de la log-verosimilitud del modelo PIC bajo la hipótesis alternativa anidada en la nula, es aproximadamente ji-cuadrada. El estadístico *D* se llama devianza.

CAPÍTULO 3
ANÁLISIS DE UN EXPERIMENTO DE FUNGICIDAS EN JITOMATE

En este capítulo presentamos el experimento en agronomía que llevó al desarrollo de esta tesis. También presentamos los resultados y damos respuesta al problema planteado por los investigadores que realizaron el experimento.

3.1 El Experimento, los Datos y el Problema

Este trabajo está motivado por un experimento que se realizó bajo condiciones de invernadero siguiendo un diseño anidado en dos etapas, véase p. 557 de Montgomery (2002), con cuatro tratamientos incluido el testigo. Los tratamientos están determinados por cuatro fungicidas *cupravit* (Cu), *manzanate* (Mz), *acrobat* (Ac) y un testigo (Ts). Cada tratamiento se replicó en cinco cuadrantes tal y como se ilustra en la Figura 1. Dentro de cada cuadrante se sembraron 32 plantas y en cada una de ellas se registró el número de daños en los foliolos, en las hojas, en los racimos y los frutos. La primera aplicación de los fungicidas fue a los dieciocho días después del trasplante excepto del testigo. Posteriormente se realizaron todas las aplicaciones de los fungicidas en intervalos de diez días en siete ocasiones. En cada ocasión se contó el número de daños en foliolos, hojas, racimos y frutos. La Tabla 2 muestra la estructura de la base datos que se conformó.

1 Cu	5 Ac	9 Cu	13 Mz	17 Ts
2 Ts	6 Mz	10 Ts	14 Ac	18 Cu
3 Ac	7 Cu	11 Ac	15 Ts	19 Mz
4 Mz	8 Ts	12 Mz	16 Cu	20 Ac

Tabla 1. Distribución de los tratamientos en los 20 cuadrantes.

Fungicida	Cuadrante	Planta	Tiempo	Foliolos	Hojas	Racimos	Fruto
Ts	1	1	10	0	0	0	0
Ts	1	1	20	0	0	0	0
Ts	1	1	30	0	0	0	0
Ts	1	1	40	0	0	0	0
Ts	1	1	50	0	0	0	0
Ts	1	1	60	0	0	0	0
Ts	1	1	70	10	4	2	0
Ts	1	2	10	0	0	0	0
Ts	1	2	20	0	0	0	0
Ts	1	2	30	0	0	0	0
Ts	1	2	40	0	0	0	0
Ts	1	2	50	0	0	0	0
Ts	1	2	60	0	0	0	0
Ts	1	2	70	0	0	0	0

Tabla 2.Fragmento de la base de datos. Se identifica al fungicida, al cuadrante, e tiempo en que se hizo la medición y el número de daños en foliolos, hojas, racimos y fruto.

El problema de investigación queda planteado en dos puntos:

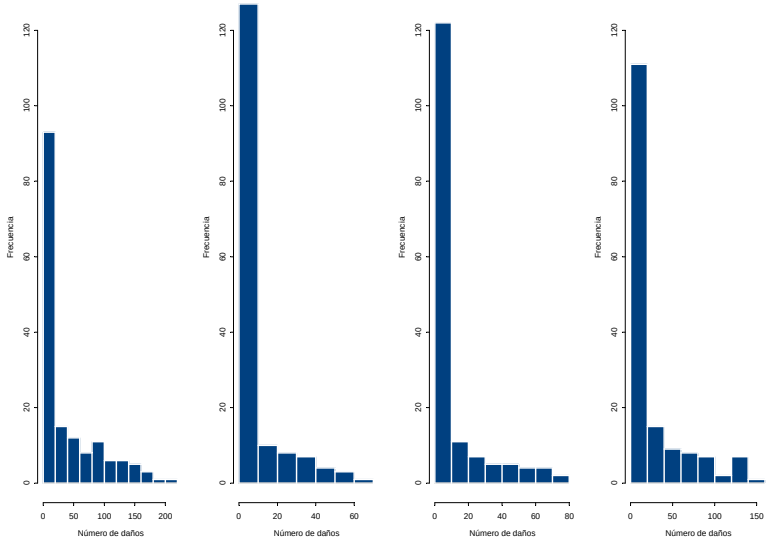
- Determinar si los fungicidas significativamente reducen el número de daños.
- En caso afirmativo de lo anterior, determinar cual fungicida es el más efectivo.

3.2 Análisis

El análisis de los datos generados por el experimento requiere de un modelo con cuatro variables respuesta de conteo: daños en foliolos, hojas, racimos y frutos. Además el modelo debe considerar la correlación entre las variables respuesta. Notemos también que los datos son medidas repetidas, por lo que existe una autocorrelación temporal dentro de las plantas. Los conteos además presentan una alta frecuencia de ceros, véase la Figura 1. El desarrollo de un modelo estadístico multivariado para varias variables respuesta de conteo con exceso de cero todas y que además tome en cuenta la autocorrelación dentro de las medidas repetidas, es un problema abierto. En este trabajo proponemos una primera solución al problema de investigación de determinar cuál es fungicida más efectivo. Para responder a esta cuestión analizaremos cada variable de manera marginal mediante el ajuste de modelos de regresión PIC al número total de daños observados en las plantas durante el período de observación.

3.2.1 Análisis Inicial

En las figuras 1 y 2 se muestran las distribuciones de frecuencias del número total de daños en las cuatro variables respuesta. Destaca inmediatamente la alta frecuencia de plantas que no presentaron daño. Esto nos indica que tenemos exceso de ceros.



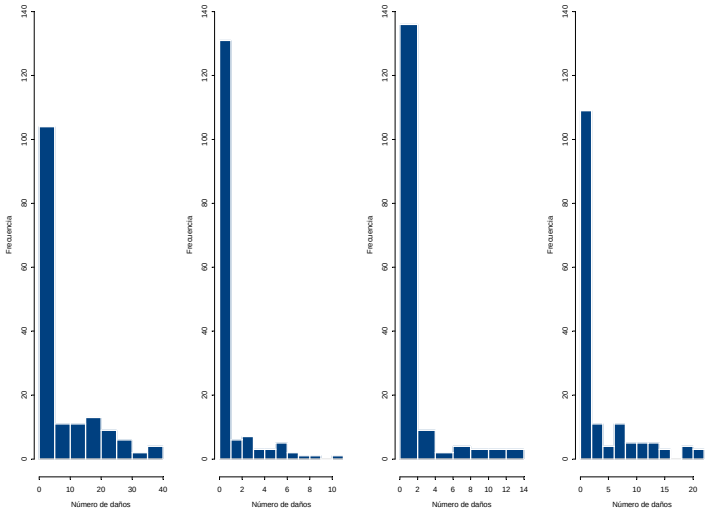
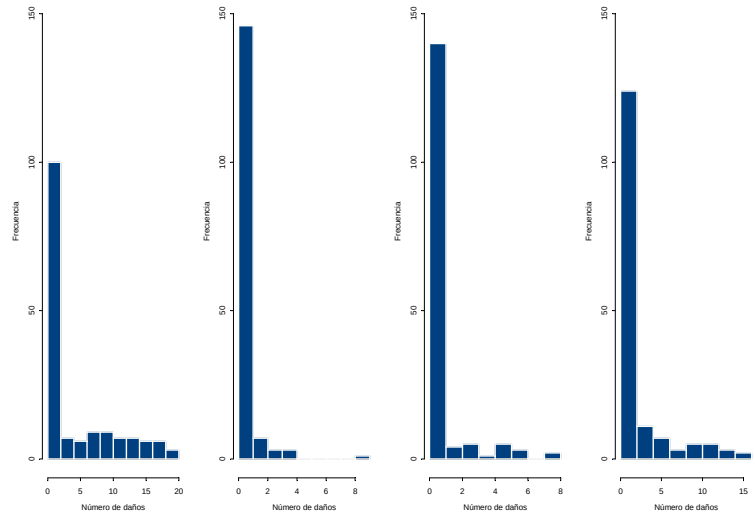


Figura 1. Distribución del número total de daños en Foliolos y Hojas.



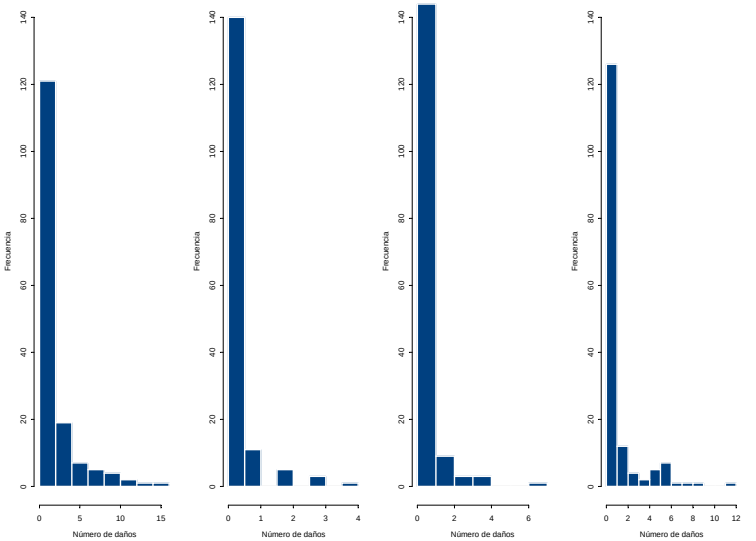


Figura 2. Distribución del número total de daños en Racimos y Fruto.

3.2.2 Análisis Definitivo

Sea Y_{ijk} la variable aleatoria número total de daños en la variable respuesta *i* de la planta *k* bajo el tratamiento *j*; donde $i = 1,\ldots,4$, con 1 daño en foliolos, 2 daño en hojas, 3 daño en racimo y 4 daño en fruto; $k = 1,\ldots,160$, pues ignoramos al cuadrante, lo que nos resulta en $5 \times 32 = 160$ plantas; $j = 1,\ldots,4$, con 1 acrobat (Ac), 2 manzanate (Mz), 3 cupravit (Cu), y 4 el testigo.

El diagrama en la Figura 3 ilustra como conceptualizamos los datos para su modelación estadística.

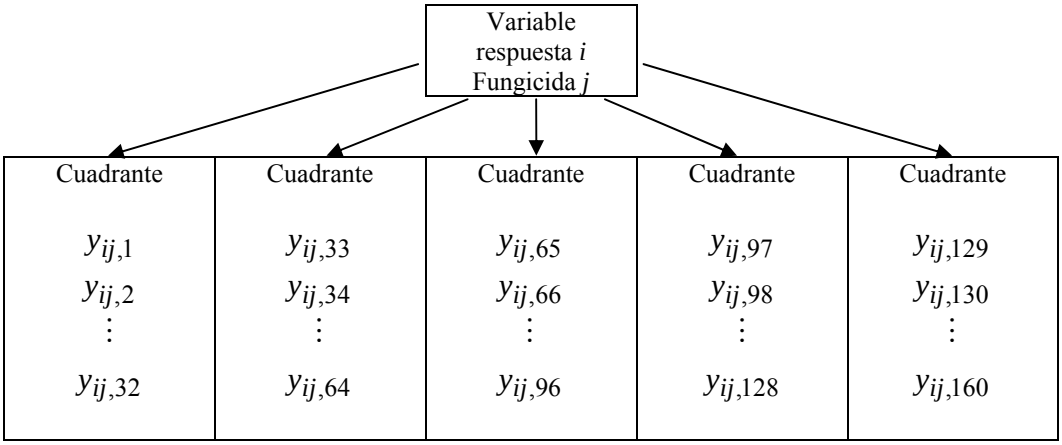


Figura 3. Daños acumulados observados en las 160 plantas en la variable respuesta *i* bajo el fungicida *j*. El cuadrante lo ignoramos en el modelo.

El modelo que ajustamos a los conteos es un modelo de regresión PIC. Es decir, suponemos que $Y_{ijk} \sim \text{PIC}(p_{ij}, \lambda_{ij})$, además suponemos que las observaciones entre plantas son independientes e idénticamente distribuidas ($k = 1,\ldots,160$). Los parámetros del modelo de regresión PIC son

$$\log(\lambda_{ij}) = \beta_{0ij} + \beta_{1ij}\,\text{Ac} + \beta_{2ij}\,\text{Mz} + \beta_{3ij}\,\text{Cu},$$

$$\text{logit}(p_{ij}) = \gamma_{0ij} + \gamma_{1ij}\,\text{Ac} + \gamma_{2ij}\,\text{Mz} + \gamma_{3ij}\,\text{Cu},$$

donde la variable explicatoria del modelo es la variable indicadora del fungicida.

Fungicida	Valores		
Ac (acrobat)	1	0	0
Mz (manzanate)	0	1	0
Cu (cupravit)	0	0	1
Testigo	0	0	0

Tabla 3. Valores de la variable explicatoria del modelo.

Con propósitos de comparación, también ajustamos modelos de regresión Poisson

$$\log(\lambda_{ij}) = \theta_{0ij} + \theta_{1ij}\,\text{Ac} + \theta_{2ij}\,\text{Mz} + \theta_{3ij}\,\text{Cu}.$$

3.3 Resultados

La Tabla 4 muestra los modelos de regresión Poisson estimados sin considerar el exceso de ceros y la Tabla 5 muestra las estimaciones de los parámetros λ_{ij}. Las estimaciones de los coeficientes son todas significativas para las cuatro variables respuesta. Por otro lado, los signos de las estimaciones producen que disminuya el valor estimado de $\log \lambda_{ij}$ y por lo tanto aumenta el valor de λ_{ij}. Esto indica que existe una relación negativa entre el número de daños y el fungicida, es decir, la aplicación de los fungicidas produce una disminución en el número esperado de daños.

Variables:	Foliolo	Hoja	Racimo	Fruto
Constante (β_0)	3.5921 (0.0131)	1.4109 (0.3904)	0.4777 (0.0622)	0.4777 (0.0622)
Ac (β_1)	−1.7280 (0.0337)	−1.5589 (0.0936)	−1.7028 (0.1585)	−2.0265 (0.1824)
Mz (β_2)	−1.3921 (0.0294)	−1.1978 (0.0810)	−0.9782 (0.1191)	−1.6017 (0.1520)
Cu (β_4)	−0.4822 (0.0212)	−0.2361 (0.0587)	0.1541 (0.0848)	−0.5356 (0.1024)

Tabla 4. Estimaciones de los parámetros del modelo de regresión Poisson. En paréntesis están los errores estándar.

Variables:	Foliolo	Hoja	Racimo	Fruto
Testigo	36.3103	4.0996	1.6124	1.6124
Ac	6.45013	0.8624	0.2937	0.2125
Mz	9.0250	1.2375	0.6062	0.3250
Cu	22.4188	3.2375	1.8809	0.9437

Tabla 5. Estimaciones de λ obtenidas del modelo de regresión Poisson.

Las Figura 4 muestra las distribuciones ajustadas con el modelo de regresión Poisson. Por ejemplo, la distribución ajustada para el daño en foliolos cuando se administra el fungicida Ac es

$$P(Y_{11} = y) = \frac{e^{6.45013}(6.45013)^y}{y!}, \quad y = 0,1,2,\ldots,$$

donde la estimación del parámetro se obtiene de $\log \hat{\lambda}_{11} = 3.5921 - 1.728 = 1.8641$ y por lo tanto $\hat{\lambda}_{11} = \exp(1.8641) = 6.45013$. Podemos ver que, en general, el uso de fungicidas reduce el número de daños en foliolos, hojas, racimos y frutos. También podemos ver que el mejor fungicida es el acrobat (Ac) y el menos efectivo es cupravit (Cu). Sin embargo, de la inspección de las Figuras 1 y 2, vemos que el modelo ajustado está subestimando sustancialmente el número observado de ceros.

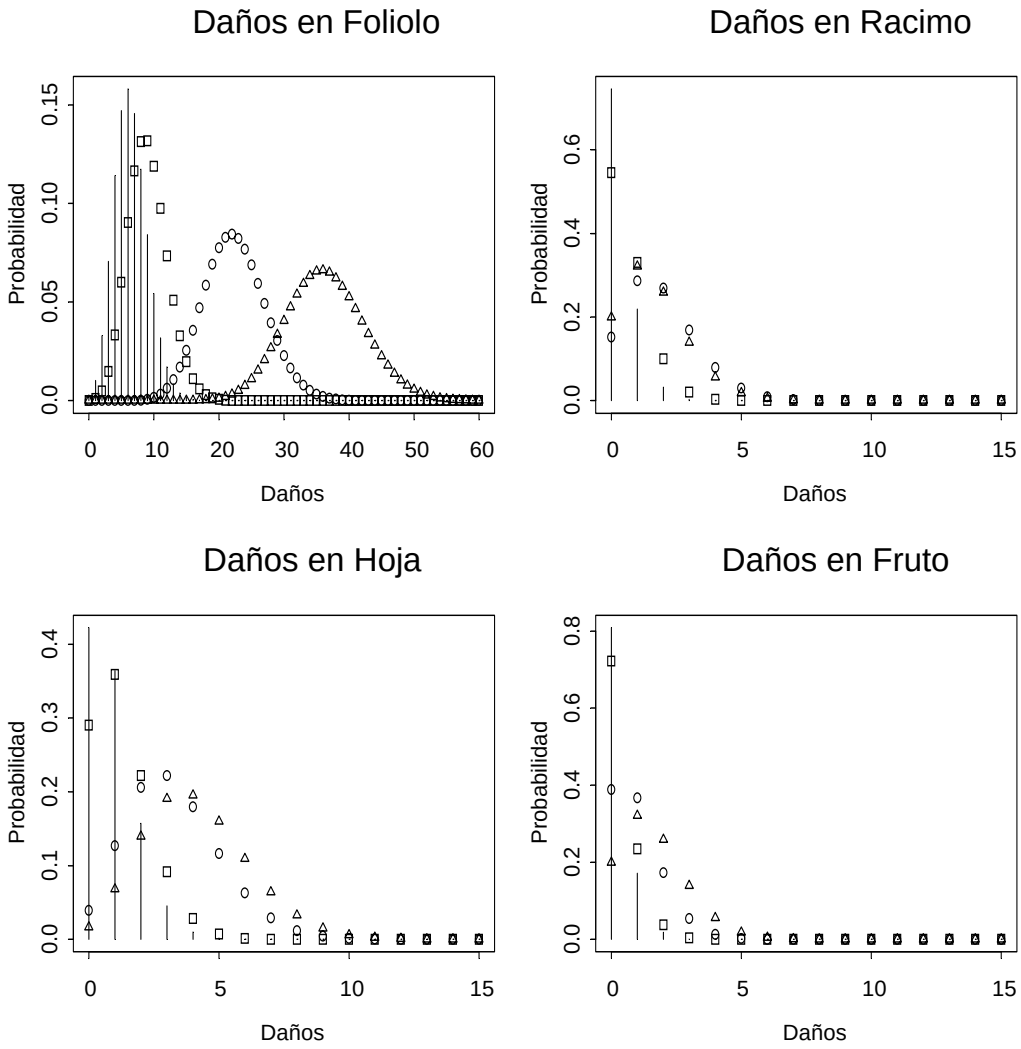


Figura 4. Modelo de regresión Poisson ajustado al número de daños. Las líneas sólidas corresponden a Ac, las líneas con cuadrados corresponden a Mz, las líneas con círculos corresponden a Cu y las líneas con triángulos corresponden al testigo.

Las estimaciones de los parámetros de los modelos de regresión PIC se muestra en la Tabla 6.

Variables	Foliolo	Hoja	Racimo	Fruto
Poisson				
Constante	4.2365 (0.0131)	2.2233 (0.0390)	1.5359 (0.0637)	1.5359 (0.0637)
Ac	−1.0108 (0.0337)	−0.8727 (0.0971)	−0.6004 (0.1766)	−1.3746 (0.2568)
Mz	−0.8733 (0.0294)	−0.7068 (0.0826)	−0.4480 (0.1278)	−0.8116 (0.1775)
Cu	−0.3281 (0.0212)	−0.0679 (0.0588)	0.1973 (0.0863)	−0.1768 (0.1061)
Ceros				
Constante	−0.10008 (0.1583)	0.2257 (0.1591)	0.6318 (0.1673)	0.6318 (0.1673)
Ac	1.1656 (0.2405)	1.01984 (0.2498)	1.4061 (0.3111)	0.8787 (0.3444)
Mz	0.8885 (0.2327)	0.7607 (0.2395)	0.7280 (0.2637)	1.0450 (0.2945)
Cu	0.3007 (0.2243)	0.2847 (0.2280)	0.0654 (0.2373)	0.5074 (0.2511)

Tabla 6. Ajuste del modelo PIC. Los valores entre paréntesis son los errores estándar.

Los parámetros λ_{ij} y p_{ij} de las distribuciones PIC ajustadas se obtienen directamente de la Tabla 6. Estos parámetros se muestran en la Tabla 7.

Poisson	Foliolo	Hoja	Racimo	Fruto
Ac	25.1712	3.8597	2.5484	1.1750
Mz	28.8815	4.5563	2.9680	2.0633
Cu	49.8192	8.6313	5.6587	3.8927
Testigo	69.1654	9.2378	4.6455	4.6456
Ceros	Foliolo	Hoja	Racimo	Fruto
Ac	0.744	0.777	0.885	0.819
Mz	0.687	0.728	0.914	0.842
Cu	0.550	0.625	0.845	0.758
Testigo	0.475	0.556	0.653	0.653

Tabla 7. Estimaciones de λ_{ij} y p_{ij} obtenidas del modelo de regresión PIC.

La Figura 5 muestra las distribuciones PIC ajustadas. Por ejemplo, la distribución ajustada al daño en foliolos cuando se administra el fungicida Ac es

$$P(Y_{11}=y)=\begin{cases}0.744+(1-0.744)\,\mathrm{e}^{-25.1712}, & \text{para } y=0,\\ (1-0.744)\dfrac{\mathrm{e}^{-25.1712}\,25.1712^{y}}{y!}, & \text{para } y=1,2,\ldots,\end{cases}$$

donde la parte Poisson se estima por $\hat{\lambda}_{11}=\exp(4.2365-1.0108)=25.1712$ y la parte del exceso de ceros se estima por

$$\hat{p}_{11} = \frac{\exp(-1.0008 + 1.1656)}{1 + \exp(-1.0008 + 1.1656)} = 0.744.$$

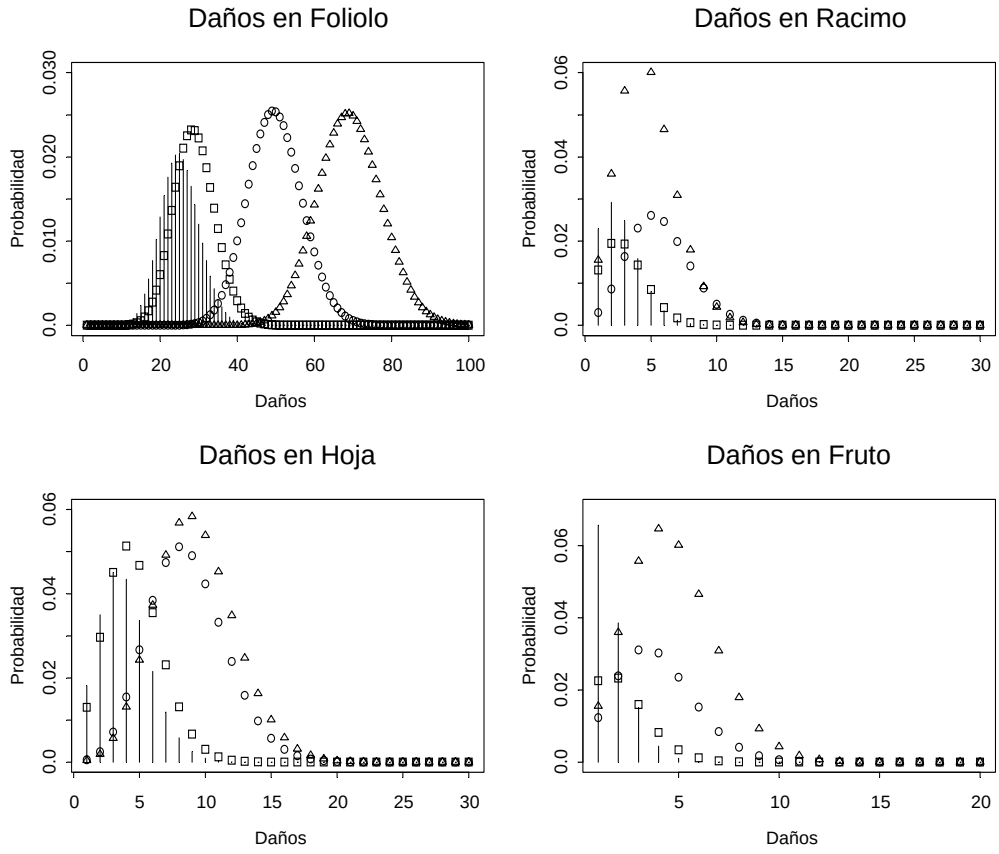


Figura 5. Modelo de regresión PIC (con el cero truncado) ajustado al número de daños. Las líneas sólidas corresponden a Ac, las líneas con cuadrados corresponden a Mz, las líneas con círculos corresponden a Cu y las líneas con triángulos corresponden al testigo.

En los modelos de regresión PIC todos los coeficientes son significativos en todas las variables respuesta, además los signos del modelo Poisson son negativos, igual que en los modelos de regresión Poisson que no toman en cuenta el exceso de

ceros. En la parte de excesos de ceros, también los coeficientes resultaron significativos, pero con signo positivo. Lo que significa que la probabilidad del proceso que genera sólo los ceros aumenta a conforme disminuye la media del proceso Poisson. Podemos establecer que con los fungicidas Ac y Mz observamos más ceros en las variables respuesta ocasionando una disminución en el número de daños en foliolos. Sin embargo, esto no sucede con el fungicida cupravit. En general, los resultados indican que el fungicida más efectivo es Ac.

Para comparar el modelo de regresión Poisson con el modelo de regresión PIC se han utilizado las devianzas dadas en la Sección 2.4 y cuyos valores están en el Apéndice. En todos los casos, resulta ser más adecuado el modelo de regresión PIC. También hacemos una comparación gráfica en las Figuras 6, 7, 8 y 9, las cuales muestran las distribuciones ajustadas a los daños en foliolos, hojas, racimos y frutos, respectivamente.

De la inspección de estas figuras es claro que el modelo de regresión Poisson subestima el número de ceros y que el modelo PIC proporciona un mejor ajuste a los datos.

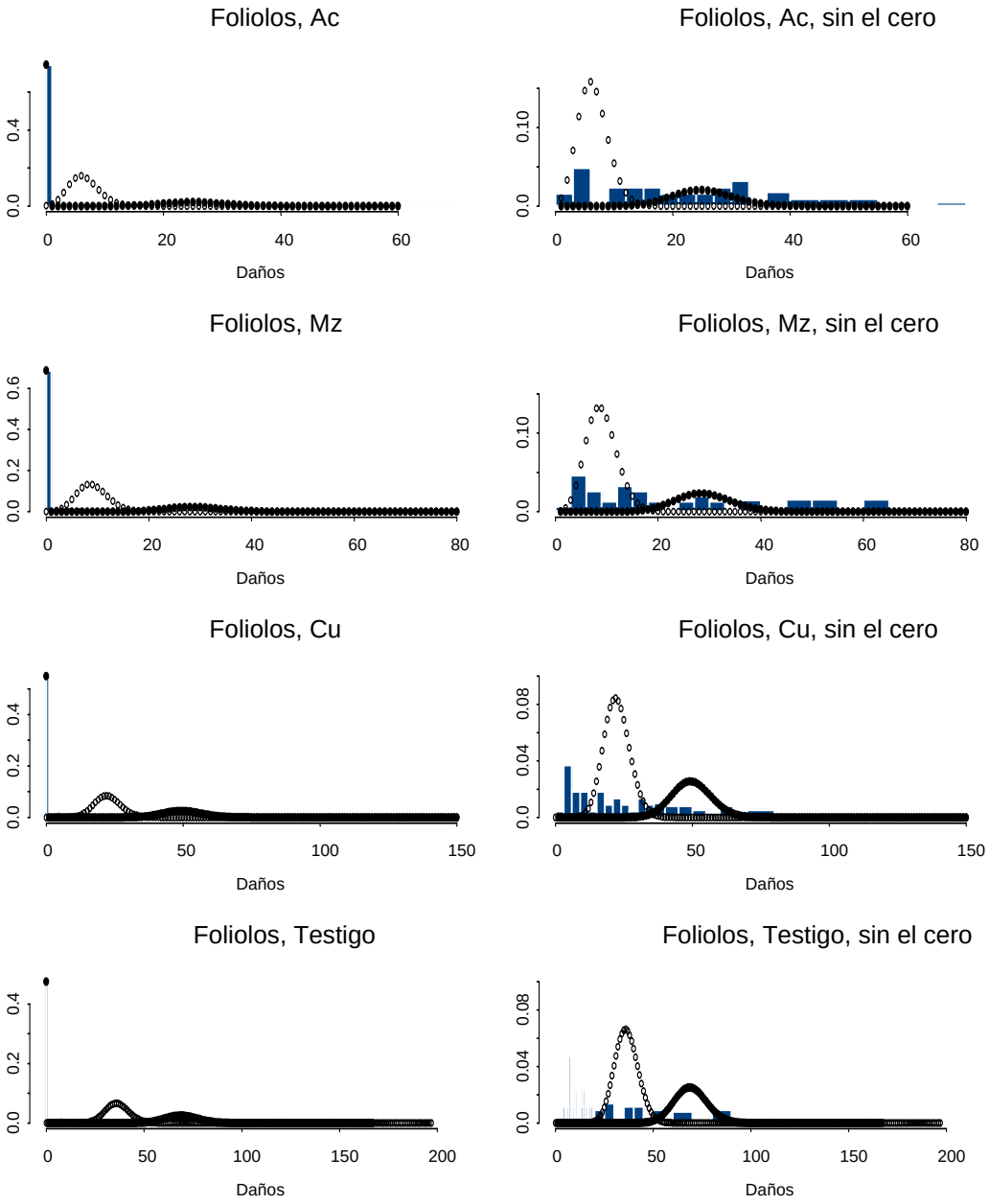


Figura 6. Distribuciones Poisson (círculos) y PIC (círculos rellenos) ajustadas y observada del número de daños en foliolos para cada fungicida.

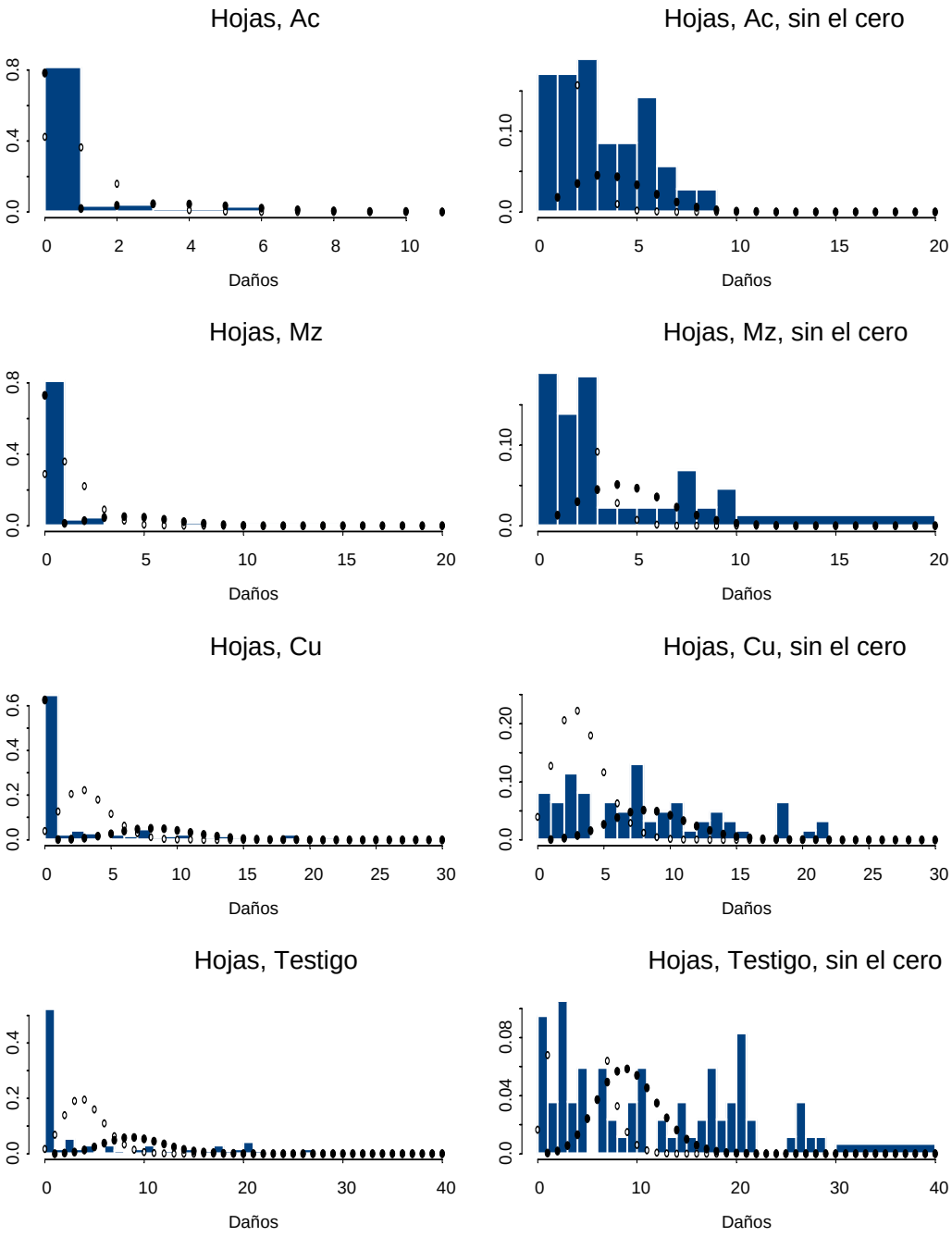


Figura 7. Distribuciones Poisson (círculos) y PIC (círculos rellenos) ajustadas y observada del número de daños en hojas para cada fungicida.

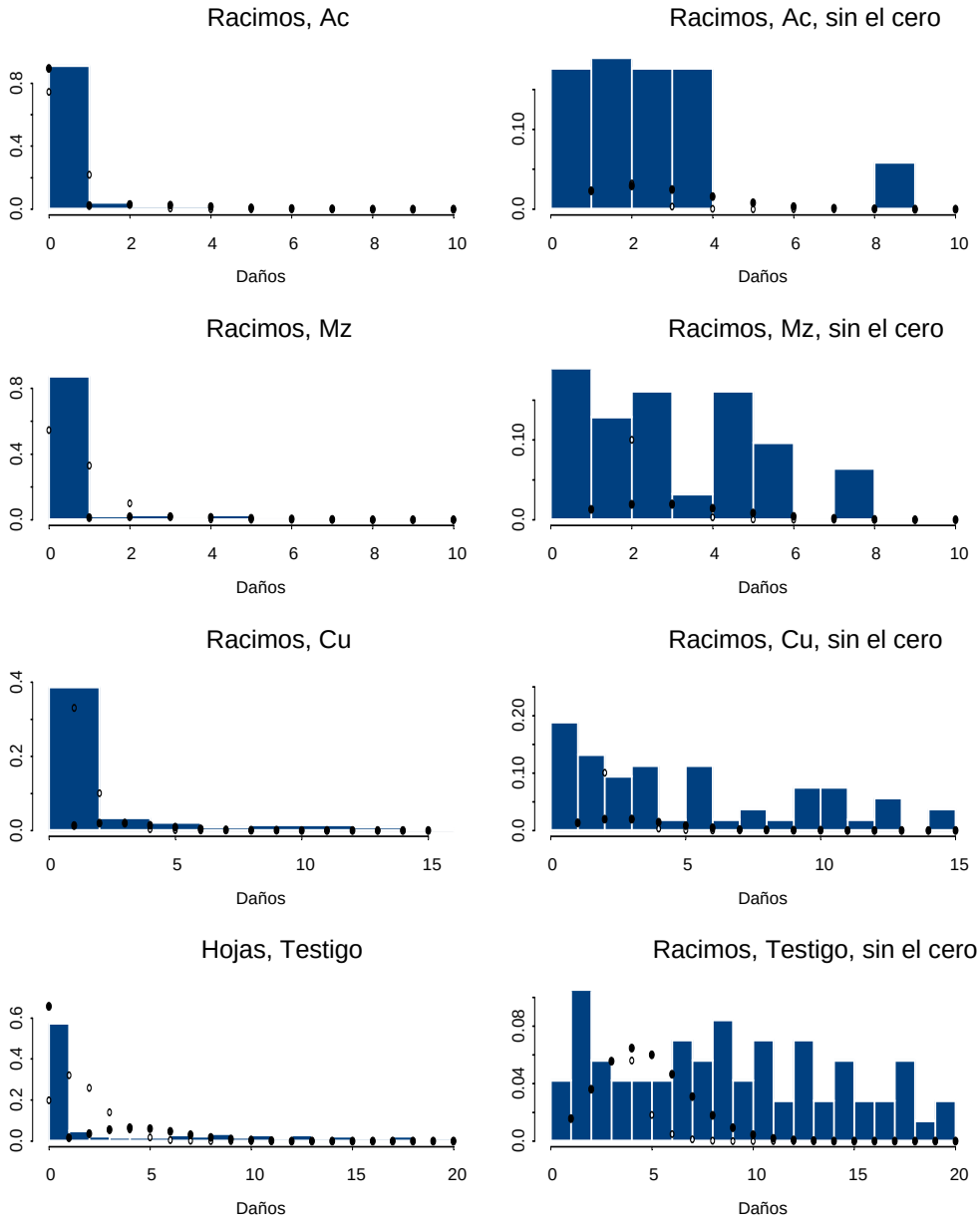


Figura 8. Distribuciones Poisson (círculos) y PIC ajustadas (círculos rellenos) y observada del número de daños en racimo para cada fungicida.

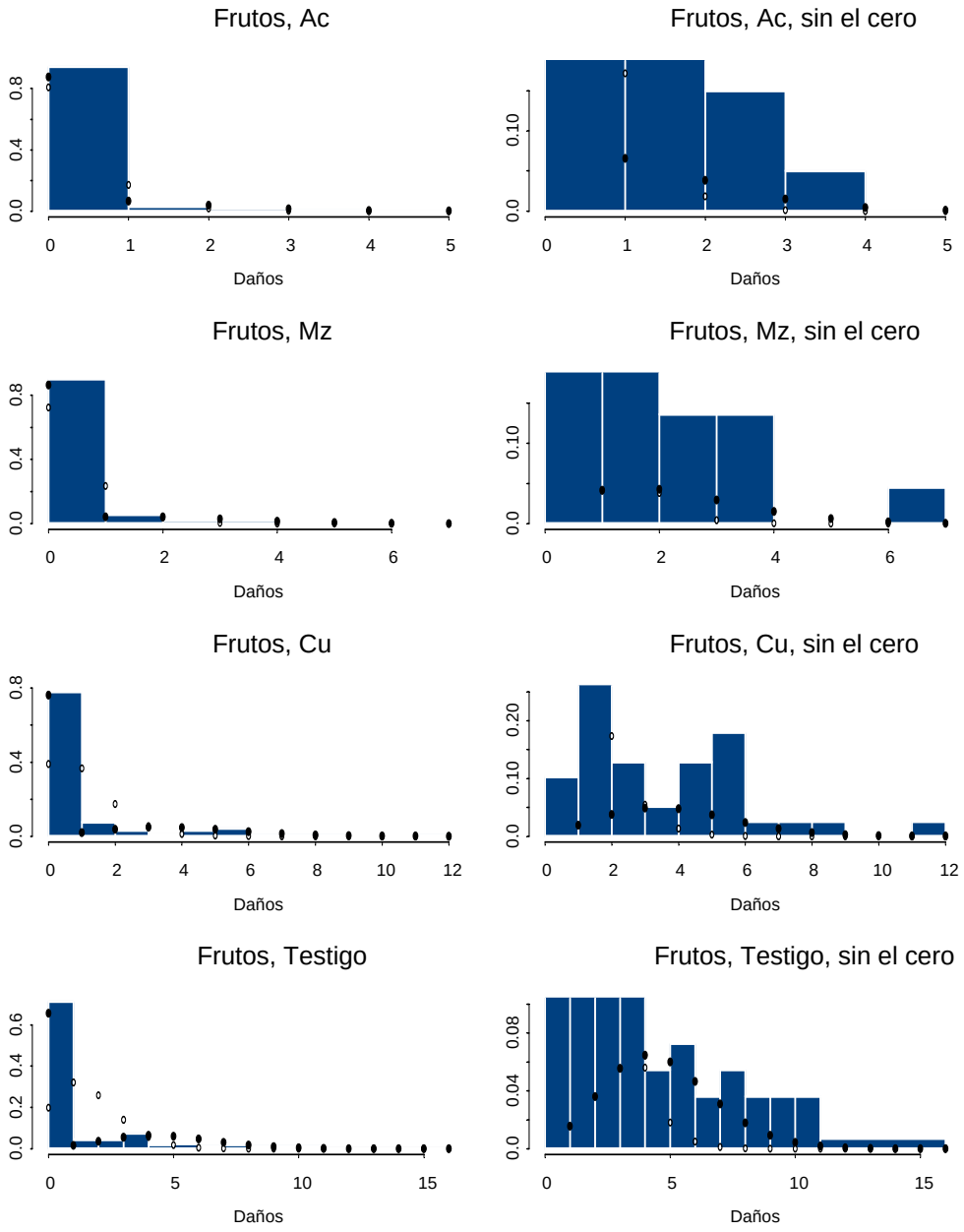


Figura 9. Distribuciones Poisson (círculos) y PIC ajustadas (círculos rellenos) y observada del número de daños en fruto para cada fungicida.

Foliolos									
Intervalo	Empíricas			Poisson			PIC		
	Ac	Mz	Cu	Ac	Mz	Cu	Ac	Mz	Cu
0	0.7440	0.6880	0.5500	0.0015	0.00012	1.83E-10	0.7440	0.6870	0.5500
1 - 10	0.0500	0.0750	0.0880	0.9343	0.7028	0.00278	0.0001	1.48E-05	3.35E-12
11 - 20	0.0630	0.0690	0.0560	0.0640	0.2965	0.35093	0.0451	0.01669	6.20E-07
21 - 30	0.0500	0.0440	0.0440	0.0000	0.0004	0.5969	0.1737	0.18018	0.00077
31 - 40	0.0440	0.0310	0.0500	0.0000	0.0000	0.04914	0.0364	0.11004	0.03974
41 - 50	0.0250	0.0310	0.0380	0.0000	0.0000	0.00027	0.0006	0.00603	0.20595
51 - 60	0.0190	0.0250	0.0190	0.0000	0.0000	1.56E-07	0.0000	3.92E-05	0.17267
61 +	0.0060	0.0380	0.1560	0.0000	0.0000	1.37E-11	0.0000	4.06E-08	0.02564
Hojas									
Intervalo	Empíricas			Poisson			PIC		
	Ac	Mz	Cu	Ac	Mz	Cu	Ac	Mz	Cu
0	0.7810	0.7310	0.6250	0.4221	0.2901	0.0392	0.7746	0.7328	0.6251
1 a 5	0.1560	0.1810	0.1250	0.5775	0.7081	0.8510	0.1751	0.1857	0.0524
6 a 10	0.0560	0.0500	0.1250	0.0002	0.0017	0.1091	0.0426	0.1697	0.2282
11 a 15	0.0060	0.0380	0.0750	0.0000	0.0000	0.0005	0.0004	0.0115	0.0883
16 +	0.0000	0.0000	0.0500	0.0000	0.0000	0.0000	7.01E-07	7.72E-05	0.0057
Racimos									
Intervalo	Empíricas			Poisson			PIC		
	Ac	Mz	Cu	Ac	Mz	Cu	Ac	Mz	Cu
0	0.8940	0.8060	0.6690	0.7455	0.5454	0.1524	0.8940	0.9184	0.8455
1 a 5	0.1000	0.1630	0.1810	0.2544	0.4545	0.8349	0.1007	0.0746	0.077
6 a 10	0.0060	0.0310	0.0880	6.93E-07	4.11E-05	0.0126	0.0052	0.0069	0.0725
11 a 15	0.0000	0.0000	0.0630	2.68E-14	5.84E-11	4.70E-06	8.37E-06	2.29E-05	0.004
16 +	0.0000	0.0000	0.0000	1.11E-22	8.98E-18	2.00E-10	1.59E-09	9.26E-09	0.000
Frutos									
Intervalo	Empíricas			Poisson			PIC		
	Ac	Mz	Cu	Ac	Mz	Cu	Ac	Mz	Cu
0	0.8750	0.8630	0.7630	0.8085	0.7225	0.3891	0.8749	0.8620	0.7629
1 a 5	0.1250	0.1310	0.1690	0.1914	0.2774	0.6103	0.1248	0.1349	0.1890
6 a 10	0.0000	0.0060	0.0630	1.06E-07	1.23E-06	0.0004	0.0002	0.0029	0.0474
11 a 15	0.0000	0.0000	0.0060	8.22E-16	7.94E-14	5.59E-09	9.14E-09	1.74E-06	0.0005
16 +	0.0000	0.0000	0.0000	6.76E-25	5.45E-22	7.79E-15	3.78E-14	1.17E-10	8.46E-07

Tabla 8. Probabilidades de los modelos Poisson y PIC ajustados; y probabilidades empíricas.

En la Tabla 8 se muestra las probabilidades ajustadas para cada modelo. Al comparar las probabilidades empíricas y las ajustadas por el modelo Poisson se pone de manifiesto lo ya mencionado para este caso, el ajuste del modelo es pobre ante la presencia de exceso de ceros. En cambio, el modelo PIC modela adecuadamente el exceso de ceros.

3.4 Conclusiones

Los resultados obtenidos del análisis nos permiten concluir, en primer lugar, que el modelo de regresión Poisson no es la herramienta adecuada para los datos debido al exceso de ceros. Esto se refleja en el ajuste pobre que proporciona a los datos, en particular a los ceros. Por otro lado, el modelo PIC modela adecuadamente al exceso de ceros y proporciona en general un buen ajuste al número de daños en foliolos, hojas, racimos y frutos. De hecho, las devianzas indican que el modelo PIC proporciona un tener mejor desempeño que el modelo Poisson en las cuatro variables respuesta. Los modelos ajustados, incluidos los de regresión Poisson, ponen en evidencia una relación negativa y significativa entre el número de daños y el uso de fungicida. Esto es, el número de daños es menor cuando se usa fungicida que cuando no se usa. Más aún, el modelo de regresión PIC indica que los fungicidas que dan los mejores resultados para el control de daños son acrobat (Ac) y manzanate (Cu) en las cuatro variables respuesta.

3.5 Desarrollos e Investigación Adicional

En este último capítulo discutimos brevemente algunos de los problemas que hemos identificado y que consideramos potenciales para desarrollos futuros. Son líneas de investigación que, hasta donde es de nuestro conocimiento, merecen investigación adicional.

3.5.1 Modelo de Regresión Poisson para Medidas Repetidas

Los datos que hemos analizado son medidas repetidas ya que son conteos que se observaron longitudinalmente en las plantas en diferentes momentos. Si por ejemplo interesara estudiar la evolución temporal del efecto de los fungicidas en las variables respuesta se tendría que ajustar un modelo de regresión para datos de medidas repetidas que tome en cuenta la dependencia entre las observaciones hechas dentro de una misma unidad. Es decir, el modelo debe tomar en cuenta la correlación serial que hay entre los conteos observados *dentro* de las plantas. Los elementos para desarrollar estos modelos están disponibles en la literatura, por ejemplo, los modelos de regresión Poisson para medidas repetidas con una estructura de dependencia autorregresiva se pueden ver en Lindsey (1993).

La extensión del modelo de regresión Poisson para medidas repetidas con exceso de ceros también ya se ha estudiado también. Algunos de estos modelos se pueden ver en Dobbie y Welsh (2001), y Hall y Zhang (2004).

3.5.2 Modelo de Regresión Poisson Multivariado

Ha habido intentos para extender la distribución Poisson al caso bivariado. Pero el problema es difícil, véase por ejemplo Karlis y Meligkotsidou (2005). La construcción de la distribución Poisson multivariada es un problema aún más difícil. La construcción de un modelo de Regresión PIC para datos multivariado para mediciones repetidas es, por lo tanto, todo un reto en la modelación estadística.

REFERENCIAS

Barry, S.C., and Welsh, A. H. (2002). Generalized additive modeling and zero inflated count data. *Ecological Modelling*, 157, 179-188.

Böhning, D., Dietz, E., and Schlattmann, P. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society A*, 162, 195-209.

Dobbie, M. J., and Welsh, A. H. (2001). Modelling correlated zero inflated count data. *Australian and New Zealand Journal of Statistics*, 43, 431-444.

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

Hall, D.B., and Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling*, 4, 161-180.

Heilbron, D. C. (1994). Zero altered and other regression models for count data with added zeros. *Biometrics*, 36, 531-547.

Hinde, J., and Demétrio, C. G. B. (1998) Overdispersion: models and estimation. *Computational Statistics and Data Analysis*, 27, 151-170.

Johnson, N.L., Kotz, S., and Kemp, A.W. (1992). *Distributions in Statistics: Discrete Distributions*. Wiley: New York.

Karlis, D., and Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing*, 15, 255-265.

Lambert, D. (1992). Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics*, 34, 1-14.

Lindsey, (1993). *Models for Repeated Measurements*. Oxford: Oxford.

Mackenzie, D. I., Nichols, J. D., Lachman, G.B., Droege, S., Royle, J.A and Langtimm, C. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83, 2248-2255.

Montgomery, D.C. (2002). *Diseño y Análisis de Experimentos*. Limusa Wiley: México D.F.

Podlich, H.M., Faddy, M.J. & Smyth, G.K. (2002). A general approach to modeling and analysis of species abundance data with extra zeros. *Journal of Agriculture, Biology, and Environmental Statistics*, 7, 324-334.

Romero, R. M., Los Arcos, E., Cano, F. V y Sánchez, P. M. (2001). *Modelos para datos de recuento de corte transversal con exceso de ceros. Aplicado a citas de patentes*. Documento de trabajo # 05. Universidad de la Laguna, España.

McCullagh, P., and Nelder, J.A. (1989). *Generalised Linear Models,* 2nd ed. Chapman and Hall: London.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Economics*, 33, 341-365.

Tara, G. M., Brendan, A. W., Jonathan, R.R., Petra, M. K., Scout, A. F., Samantha J. L., Andrew, J.T and Hugh, P. P. (2005). Zero tolerance ecology: improving ecological inference by modeling the source of zero observations. *Ecology*, 8, 1235-1246.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. and Lindenmayer, D.B. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88, 297-308.

Welsh, A. H., Cunningham, R. B. & Chambers, R. (2000). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88, 297-308.

APÉNDICE
Salidas de STATA

A continuación presentamos las salidas del programa STATA con el ajuste del modelo de regresión Poisson para el número de daños en foliolos, hojas, racimos y frutos.

A.1 Modelo de Regresión Poisson

Ajuste de la variable foliolos.

```
Iteration 0:   log likelihood = -13788.348
Iteration 1:   log likelihood = -13787.433
Iteration 2:   log likelihood = -13787.433

Poisson regression                                Number of obs   =        640
                                                  LR chi2(3)      =    4900.91
                                                  Prob > chi2     =     0.0000
Log likelihood = -13787.433                       Pseudo R2       =     0.1509

------------------------------------------------------------------------------
     foliolo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ac |  -1.728082   .0337803   -51.16   0.000     -1.79429   -1.661874
          mz |  -1.392164   .0294047   -47.34   0.000    -1.449796   -1.334531
          cu |  -.4822644   .0212344   -22.71   0.000    -.5238831   -.4406456
       _cons |   3.592162   .0131193   273.81   0.000     3.566449    3.617875
------------------------------------------------------------------------------
```

Ajuste de la variable hojas.

```
Iteration 0:   log likelihood = -2223.2667
Iteration 1:   log likelihood =  -2223.228
Iteration 2:   log likelihood =  -2223.228

Poisson regression                                Number of obs   =        640
                                                  LR chi2(3)      =     519.51
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2223.228                       Pseudo R2       =     0.1046

------------------------------------------------------------------------------
       hojas |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ac |  -1.558907   .0936524   -16.65   0.000    -1.742462   -1.375352
          mz |  -1.197894   .0810857   -14.77   0.000    -1.356819   -1.038969
          cu |  -.2361855   .0587783    -4.02   0.000     -.351389   -.1209821
       _cons |   1.410987   .0390434    36.14   0.000     1.334463    1.487511
------------------------------------------------------------------------------
```

Ajuste de la variable racimos.

```
Iteration 0:   log likelihood = -1290.7967
Iteration 1:   log likelihood = -1290.6601
Iteration 2:   log likelihood = -1290.6599
Iteration 3:   log likelihood = -1290.6599

Poisson regression                                Number of obs   =        640
                                                  LR chi2(3)      =     282.71
                                                  Prob > chi2     =     0.0000
Log likelihood = -1290.6599                       Pseudo R2       =     0.0987
```

```
------------------------------------------------------------------------------
      racimo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ac |  -1.702812   .1585956   -10.74   0.000    -2.013654    -1.39197
          mz |  -.9782486   .1191018    -8.21   0.000    -1.211684   -.7448133
          cu |   .1541507   .0848424     1.82   0.069    -.0121373    .3204387
       _cons |   .4777858   .0622573     7.67   0.000     .3557637    .5998078
------------------------------------------------------------------------------
```

Ajuste de la variable fruto.

```
Iteration 0:   log likelihood = -966.25433
Iteration 1:   log likelihood = -966.13697
Iteration 2:   log likelihood = -966.13693

Poisson regression                                Number of obs   =        640
                                                  LR chi2(3)      =     261.19
                                                  Prob > chi2     =     0.0000
Log likelihood = -966.13693                       Pseudo R2       =     0.1191

------------------------------------------------------------------------------
       fruto |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ac |  -2.026599   .1824492   -11.11   0.000    -2.384193   -1.669005
          mz |  -1.601716    .152009   -10.54   0.000    -1.899648   -1.303784
          cu |  -.5356796   .1024621    -5.23   0.000    -.7365016   -.3348575
       _cons |   .4777858   .0622573     7.67   0.000     .3557637    .5998078
------------------------------------------------------------------------------
```

A.2 Modelo de Regresión PIC

Ajuste de la variable foliolos.

```
Fitting constant-only model:

Iteration 0:   log likelihood = -958.15579
Iteration 1:   log likelihood = -752.64999
Iteration 2:   log likelihood = -715.25156
Iteration 3:   log likelihood = -638.54777
Iteration 4:   log likelihood = -637.74853
Iteration 5:   log likelihood = -637.74788
Iteration 6:   log likelihood = -637.74788
Fitting full model:

Iteration 0:   log likelihood = -637.74788
Iteration 1:   log likelihood = -608.65154
Iteration 2:   log likelihood = -605.61921
Iteration 3:   log likelihood = -605.56902
Iteration 4:   log likelihood = -605.56889
Iteration 5:   log likelihood = -605.56889

Zero-inflated poisson regression               Number of obs   =        640
                                               Nonzero obs     =        135
                                               Zero obs        =        505

Inflation model = logit                        LR chi2(3)      =      64.36
Log likelihood  = -605.5689                    Prob > chi2     =     0.0000
```

```
------------------------------------------------------------------------------
       fruto |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
fruto        |
          ac |  -1.374697   .2568582    -5.35   0.000     -1.87813    -.871264
          mz |  -.8116209    .177596    -4.57   0.000    -1.159703   -.4635391
          cu |   -.176887   .1061377    -1.67   0.096    -.3849131     .031139
       _cons |   1.535979   .0637085    24.11   0.000     1.411113    1.660845
-------------+----------------------------------------------------------------
inflate      |
          ac |   .8787203   .3444423     2.55   0.011     .2036257    1.553815
          mz |    1.04501    .294503     3.55   0.000     .4677945    1.622225
          cu |   .5074396   .2511082     2.02   0.043     .0152765    .9996027
       _cons |   .6318887   .1673573     3.78   0.000     .3038744    .9599031
------------------------------------------------------------------------------
```

Ajuste de la variable hojas.

```
Fitting constant-only model:

Iteration 0:   log likelihood = -1870.7964
Iteration 1:   log likelihood = -1291.8797
Iteration 2:   log likelihood = -1186.9918
Iteration 3:   log likelihood = -1185.8508
Iteration 4:   log likelihood = -1185.8504
Iteration 5:   log likelihood = -1185.8504
```

```
Fitting full model:

Iteration 0:   log likelihood = -1185.8504
Iteration 1:   log likelihood = -1107.1226
Iteration 2:   log likelihood = -1104.0025
Iteration 3:   log likelihood = -1103.9971
Iteration 4:   log likelihood = -1103.9971

Zero-inflated poisson regression                Number of obs   =        640
                                                Nonzero obs     =        209
                                                Zero obs        =        431

Inflation model = logit                         LR chi2(3)      =     163.71
Log likelihood  = -1103.997                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
       hojas |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
hojas        |
          ac |  -.8727761   .0971019    -8.99   0.000    -1.063092   -.6824599
          mz |  -.7068732   .0826612    -8.55   0.000    -.8688862   -.5448601
          cu |  -.0679314   .0588153    -1.15   0.248    -.1832072    .0473445
       _cons |   2.223384    .039061    56.92   0.000     2.146826    2.299942
-------------+----------------------------------------------------------------
inflate      |
          ac |    1.01984   .2498623     4.08   0.000     .5301191    1.509561
          mz |   .7607279   .2395833     3.18   0.001     .2911533    1.230302
          cu |   .2847585    .228028     1.25   0.212    -.1621682    .7316853
       _cons |   .2257817   .1591364     1.42   0.156    -.0861199    .5376833
------------------------------------------------------------------------------
```

Ajuste de la variable racimos.

```
Fitting constant-only model:

Iteration 0:   log likelihood = -1204.7278
Iteration 1:   log likelihood = -871.47266
Iteration 2:   log likelihood = -788.20871
Iteration 3:   log likelihood = -775.38774
Iteration 4:   log likelihood = -775.27909
Iteration 5:   log likelihood = -775.27904

Fitting full model:

Iteration 0:   log likelihood = -775.27904
Iteration 1:   log likelihood = -752.91511
Iteration 2:   log likelihood = -751.98956
Iteration 3:   log likelihood = -751.98607
Iteration 4:   log likelihood = -751.98607

Zero-inflated poisson regression                Number of obs   =        640
                                                Nonzero obs     =        156
                                                Zero obs        =        484

Inflation model = logit                         LR chi2(3)      =      46.59
Log likelihood  = -751.9861                     Prob > chi2     =     0.0000
```

```
------------------------------------------------------------------------------
      racimo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
racimo       |
          ac |  -.6004694   .1766466    -3.40   0.001    -.9466903   -.2542485
          mz |  -.4480178   .1278423    -3.50   0.000    -.6985841   -.1974514
          cu |    .197349   .0863024     2.29   0.022     .0281995    .3664985
       _cons |   1.535979   .0637085    24.11   0.000     1.411113    1.660845
-------------+----------------------------------------------------------------
inflate      |
          ac |   1.406182    .311117     4.52   0.000     .7964036     2.01596
          mz |   .7280684   .2637855     2.76   0.006     .2110584    1.245078
          cu |   .0654242    .237325     0.28   0.783    -.3997242    .5305727
       _cons |   .6318887   .1673573     3.78   0.000     .3038744    .9599031
------------------------------------------------------------------------------
```

Ajuste de la variable frutos.

```
Fitting constant-only model:

Iteration 0:   log likelihood = -958.15579
Iteration 1:   log likelihood = -752.64999
Iteration 2:   log likelihood = -715.25156
Iteration 3:   log likelihood = -638.54777
Iteration 4:   log likelihood = -637.74853
Iteration 5:   log likelihood = -637.74788
Iteration 6:   log likelihood = -637.74788
```

```
Fitting full model:

Iteration 0:   log likelihood = -637.74788
Iteration 1:   log likelihood = -608.65154
Iteration 2:   log likelihood = -605.61921
Iteration 3:   log likelihood = -605.56902
Iteration 4:   log likelihood = -605.56889
Iteration 5:   log likelihood = -605.56889

Zero-inflated poisson regression                Number of obs   =        640
                                                Nonzero obs     =        135
                                                Zero obs        =        505

Inflation model = logit                         LR chi2(3)      =      64.36
Log likelihood  = -605.5689                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
       fruto |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
fruto        |
          ac |  -1.374697   .2568582    -5.35   0.000     -1.87813    -.871264
          mz |  -.8116209    .177596    -4.57   0.000    -1.159703   -.4635391
          cu |   -.176887   .1061377    -1.67   0.096    -.3849131     .031139
       _cons |   1.535979   .0637085    24.11   0.000     1.411113    1.660845
-------------+----------------------------------------------------------------
inflate      |
          ac |   .8787203   .3444423     2.55   0.011     .2036257    1.553815
          mz |    1.04501    .294503     3.55   0.000     .4677945    1.622225
          cu |   .5074396   .2511082     2.02   0.043     .0152765    .9996027
       _cons |   .6318887   .1673573     3.78   0.000     .3038744    .9599031
------------------------------------------------------------------------------
```